

DISSERTATION

MODELING ENERGY SYSTEMS USING LARGE DATA SETS

Submitted by

Gerald P. Duggan

Department of Systems Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2024

Doctoral Committee:

Advisor: Peter Young

Co-Advisor: Daniel Zimmerle

Thomas Bradley

Ellison Carter

Copyright by Gerald P. Duggan 2024

All Rights Reserved

ABSTRACT

MODELING ENERGY SYSTEMS USING LARGE DATA SETS

Modeling and simulation are playing an increasingly import role in the sciences, and science is having a broader impact on policy definition at a local, national, and global scale. It is therefore important that simulations which impact policy produce high-quality results. The veracity of these models depend on many factors, including the quality of input data, the verification process for the simulations, and how result data are transformed into conclusions. Input data often comes from multiple sources and it is difficult to create a single, verified data set. This dissertation describes the challenges in creating a research-quality, verified and aggregated data set. It then offers solutions to these challenges, then illustrates the process using three case studies of published modeling and simulation results from different application domains.

ACKNOWLEDGEMENTS

This work was partially supported by the *Collaboratory to Advance Methane Science*, agreement number S877.

DEDICATION

For my family.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
DEDICATION	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ACRONYMS	xi
 Chapter 1	
Introduction	1
1.1 Physical System Simulation	1
1.2 Large Data Sets	2
1.3 A Process for Creating Energy System Simulations based on Large Data Sets	5
1.3.1 Data Management	5
1.3.2 Modeling	8
1.3.3 Analysis	11
1.3.4 Automation	12
1.4 Document Structure	12
 Chapter 2	
Electricity Consumption in Commercial Buildings During Covid-19	14
2.1 Overview	14
2.1.1 Data Management	14
2.1.2 Modeling	15
2.1.3 Related Works	15
2.2 Introduction	16
2.2.1 Energy use during the COVID-19 pandemic	17
2.2.2 Context for the present study	18
2.3 Methods	19
2.3.1 Sample characteristics	19
2.3.2 Procedure for modeling expected electricity usage	23
2.3.3 Timeline of COVID-related events in Fort Collins	25
2.4 Results	27
2.4.1 Utility classification results	27
2.4.2 Detailed property classification results	29
2.5 Discussion	35
2.5.1 Findings and implications	35
2.5.2 Limitations	37
2.5.3 Conclusion	38
2.6 Data Availability Statement	39
 Chapter 3	
Modeling air emissions from complex facilities at detailed temporal and spa- tial resolution: The Methane Emission Estimation Tool (MEET)	40

3.1	Overview	40
3.1.1	Data Management	40
3.1.2	Modeling	40
3.1.3	Analysis	41
3.1.4	Related Works	41
3.2	Introduction	42
3.3	Methods	46
3.4	Results and Discussion	55
3.5	Conclusions	64
Chapter 4	Real-Time Simulation and Modeling	66
4.0.1	Modeling	66
4.0.2	Related Works	66
4.1	Networking for Real-Time Distributed Simulation	67
Chapter 5	Conclusion	70
5.0.1	Future Work	71
Appendix A	Advanced Metering Infrastructure Modeling	95
A.1	A Method to Estimate Residential PV Generation from Net-Metered Load Data and System Install Date	95
A.2	Big Data Analytics for Power Distribution Systems using AMI and Open Source Tools	106
Appendix B	Mechanistic Emissions Modeling	112
B.1	Methane Emissions from the Natural Gas Transmission and Storage Sys- tem in the United States	112
B.2	A Methane Emission Estimation Tool (MEET) for predictions of emis- sions from midstream gas compressor stations with fine scale temporal and spatial resolution. <i>Supplementary Information</i>	123
B.2.1	Compressor Station Equipment	123
B.2.2	Operating Mode Timing	132
B.2.3	Gas Composition	135
B.2.4	Sources of Emissions and Activity Data	135
B.2.5	Leak Emissions	140
B.2.6	Start and Blowdown Episodic Venting	149
B.2.7	Non-Episodic Vented Emissions	152
B.2.8	Compressor Driver Combustion Emissions	153
B.2.9	Simulation Examples	155
Appendix C	Real-Time Distributed Simulation	166
C.1	Precision Timing on Linux Microcomputers	166
C.2	Mitigating Communication Delays in Remotely Connected Hardware-in- the-Loop Experiments	170

Appendix D	Data Validation Taxonomy	181
------------	------------------------------------	-----

LIST OF TABLES

2.1	Number of parcels associated with each one of the groupings from the ENERGYSTAR Portfolio Manager.	22
2.2	Timeline of relevant COVID-related events in Fort Collins, Colorado	26
2.3	Monthly deviation from expected consumption. Note: Shaded values are deviations that are statistically significant and bold shows reduction relative to the expected value.	29
2.4	Deviation on energy consumption broken down by different commercial sectors. Note: Shaded values are deviations that are statistically significant and bold shows reduction relative to the expected value.	30
3.1	Emitter states by compressor mode	51
3.2	Key operating characteristics of test stations	56
B.1	Valve states for compressor operating modes	130
B.2	Emitter states by compressor mode	131
B.3	Compressor transition matrix	133
B.4	Gas starter default distributions	150
B.5	Compressor blowdown default distributions	151
B.6	Key characteristics of stations in test data set	155
B.7	All sources modeled in simulations	156
B.8	Timing of 1-minute Extracts	157
B.9	Fugitive Emitters In Sample Iterations with Emissions ≥ 10 kg/h THC	159
B.10	Tracer Simulation Results	162

LIST OF FIGURES

1.1	System Engineering Flow for Energy Simulations	6
1.2	A Taxonomy for Model Types	9
2.1	Regression Of Energy Daily Consumption vs. Average Daily Temperature	25
2.2	Expected vs. Actual Daily Energy Consumption	27
2.3	Daily consumption deviation from expected.	28
2.4	Daily consumption deviation from expected for the Healthcare Outpatient classification.	31
2.5	Daily consumption deviation from expected for the Poudre School District (K-12 schools) Property Classification.	32
2.6	Daily consumption deviation from expected for the Food Service Property Classification.	32
2.7	Daily consumption deviation from expected for the Lodging Property Classification.	33
2.8	Daily consumption deviation from expected for the Public Assembly Property Classification.	33
2.9	Daily consumption deviation from expected for the Religious Property Classification.	34
2.10	Daily consumption deviation from expected for the Office Property Classification.	34
2.11	The ratio of actual to predicted energy use in 2021 grouped by property classification.	35
3.1	Simulation process	47
3.2	Compressor unit operating states	51
3.3	Simulation examples for compressor 12 on station 6. Left panel displays the time series. Light gray lines indicate the 20 individual Monte Carlo iterations and the dark line indicates the mean over all iterations. Right panel displays the same data as a histogram of values.	57
3.4	Results of tracer simulation. Data are grouped by station, with tracer estimates in declining order by median estimate. Results from all stations combined are shown as a per-station estimate labeled ‘Mean’. Gray blocks represent the confidence interval (CI) for station emissions based upon 365 days of simulated emissions. Tracer estimates for each day are shown as median and mean points and bars with 95% confidence intervals. Confidence interval bars extending outside the gray region represent the probability that a field measurement campaign measuring each of these facilities on one day would estimate emissions outside the confidence interval of annual emissions for these facilities.	58
3.5	The ethane-methane ratio of emissions from the simulated compressor stations. In this simulation, all stations handled natural gas with the same C2/C1 ratio (0.204); variation in the C2/C1 ratio is entirely due to changing activity (vented and combusted emissions) and leak emissions at the station.	60

3.6	Fraction of the time that a continuous monitor (CM) would detect fugitive emissions. The horizontal axis shows the detection thresholds for the CM solution, ranging from 0% (a perfect system that can distinguish <i>any</i> fugitives within any amount of non-fugitive emissions) to 100% (all emissions must be fugitive emissions for a detection to take place). A detection threshold of 50% indicates fugitive emissions equal non-fugitive emissions. The vertical axis shows the the fraction of the time the continuous monitor would successfully indicate that fugitive emissions were occurring. The color scale indicates the time-weighted utilization of compressors on the station, from 0 – no compressors operating – to 1 – all compressor operating.	63
4.1	Network Timing	67
4.2	Reference Signal Synchronization	68
4.3	Ongoing Timing	69
B.1	Compressor station equipment included in the model	124
B.2	Aerial view of a compressor station	124
B.3	Simplified single compressor schematic	125
B.4	Compressor schematic for a multi-stage compressor	126
B.5	Continuous monitoring simulation for Station 6	163
B.6	Continuous monitoring simulation for Station 7	164
B.7	Detection patterns for Stations 6 & 7	165

LIST OF ACRONYMS

FCU Fort Collins Utilities
CSU Colorado State University
EPA U.S. Environmental Protection Agency
PSD Poudre School District
NREL National Renewable Energy Laboratory
CSV Comma Separated Variable
GHG greenhouse gas
GHGRP Greenhouse Gas Reporting Program
GHGI Greenhouse Gas Inventory
LDAR leak detection and repair
T&S transmission and storage
G&P gathering and processing
O&G oil and gas
AGRU acid gas removal unit
OEL open-ended line
PRV pressure reducing valve
PC pneumatic controller
SCADA supervisory control and data acquisition
BU bottom-up
TD top-down
MC Monte Carlo
MEET Methane Emissions Estimation Tool
DES discrete event simulator
AMI Advanced Metering Infrastructure
PV Photovoltaic
PII Personally Identifiable Information
QA Quality Assurance
IoT Internet of Things
I-IoT Industrial Internet of Things
SCADA Supervisory Control and Data Acquisition
NAT Network Address Translation
DJ Denver-Julesberg Basin
EU European Union
NaN Not a Number
UTC Coordinated Universal Time

Chapter 1

Introduction

In 2023, the U.S. Environmental Protection Agency (EPA) established a Waste Emission Charge on oil and gas providers of \$900 per metric ton of methane in 2024, increasing to \$1200 for 2025, and \$1500 for 2026 and beyond [111]. Similarly, the European Union (EU) is driving toward a differentiated natural gas market to reduce greenhouse gas emissions across global natural gas supply chains [106]. Differentiated natural gas is gas that is purported by operators as having undergone independent third-party certification and that the gas has been produced under specified best practices around methane mitigation.

In both these cases, policy positions are being established that will cause operators money if they emit excess amounts of methane in their natural gas production. Also in both these cases, scientific studies are called upon to define the metrics for excess emissions, methods to measure these emissions accurately, and reconcile different measurement approaches.

A critical mechanism to conduct the necessary science is to develop models of the production facilities, and perform physical simulations to ensure measurement systems are fair and accurate.

1.1 Physical System Simulation

A *physical system* is a set of interconnected components that work together to perform a specific function [9, 14, 105]. Such systems exist in many domains including physics, aerospace, biology, and energy. Fritzson [14] further clarifies that physical systems must follow fundamental laws of nature such as Newton's laws of motion, thermodynamics, electromagnetism, etc. A *physical system model* [9] is an information representation of a physical system, used to understand the behavior of a system. These models typically specify inputs to the system and corresponding outputs. A *physical system simulation* is the application of computation models to the study and prediction of physical events on the behavior of an engineered system [9, 14]. Benefits of system simulation include:

- The physical system may not be available, for example, if it is under development and simulation is used model its behaviors for system-of-system development [141, 130].
- Another reason the physical system may not be available is that it is too far from the researcher, either in terms of distance or time. Examples include simulating the creation of a star system [7] or simulating plate tectonics [3].
- The physical system may be in production, in which case performing experiments on a running system may be too disruptive, expensive, or dangerous.
- The time scale of the system may be too slow or too fast for experimental data collection. It takes millions of years to observe small changes in the development of the universe, while these changes can be quickly observed in a simulation [14]. At the other end of the scale, a nuclear reaction happens far too quickly to be observed (at least without expensive equipment), and simulations provide insight into this physical process [5].
- Simulations allow the systems being modeled to be subjected to many inputs, including events or combinations of events that do not happen in normal operation of the system. These "what-if" scenarios can be used to find fault conditions [10] or behavior under other rare events.

Energy system simulation, the topic of this dissertation, is a subset of physical simulation, and can cover such topics as electrical distribution systems [48, 66, 51], impact of renewable energy and distributed generation sources [44], and methane emissions modeling [33, 100].

1.2 Large Data Sets

There has been an explosion of data over the past decades [107, 85]. One cause of this explosion of data is due to the Internet of Things (IoT) ([104]) – that is, sensors deployed with the ability to stream data to a central facility. At the time of this writing, estimates of the number of IoT sensors range between 10B-12B ([101], [40]) and almost doubling to 30B by 2027 ([101]). In terms of

data generated, [101] estimates that there will be 7.31 ZB (10^{15} bytes) of data generated by IoT devices by 2025.

A subset of deployed IoT devices are used for industrial communication, termed Industrial Internet of Things (IIoT) [47, 49]. This subject includes traditional disciplines of industrial automation and control systems, and Supervisory Control and Data Acquisition (SCADA). The TATA Power Delhi Distribution Limited, a power distribution company providing electricity to the North and Northwest part of Delhi, uses IIoT utility operations, including automated integration testing of remote sites, remote data access for diagnostic activities, relay tripping and fault analysis, and transformer asset health maintenance [99]. Bhattacharjee and Nandi [52] describe the use of IIoT for renewable energy, including building energy management, electric vehicle integration, and smart microgrid management.

While not technically IIoT, Advanced Metering Infrastructure (AMI) provides an enormous amount of consumption data about their customers (including commercial, residential, and industrial users) that utilities can use to remotely measure energy use, connect and disconnect service, detect tampering and monitor power quality [60]. Benefits to the utility and customer of AMI include reduced costs in metrics and billing by reducing the need for utility personnel to go on site to read meters, greater visibility of power consumption by the customer to help manage usage time and costs, improved asset utilization and reduced peak demand resulting in lower utility captured expenses and consumer savings, and faster outage restoration due to more precise dispatching of repair crews to the location they are needed [60]. Utilities can use AMI data for modeling purposes, specifically for load forecasting [15] and capacity planning [16, 92, 103].

SCADA has been implemented in the oil and gas sector for decades. Bramantyo [91] provides case studies where increased sensor coverage due to IIoT, combined with machine learning, can lead to reduced costs due to real-time performance monitoring and early detection of equipment failures. Specifically, [91] gives examples of enhanced monitoring of Pressure Relief Valves saving up to 20 percent in operational expenses.

Deloitte [108] describes the use of I-IoT for sensor-based tank monitoring, acoustic operations monitoring, and the creation of digital twins for change analysis during project management and maintenance operations.

Philips ([57]) outlines four critical design factors for a core digital solutions architecture in oil and gas, which are applicable to other I-IoT solutions:

- Establish a foundation of trust for all communications and data sharing. This includes prototyping solutions, building in data validation as it is collected, and cross-checking individual sensor data with other related sensors for data consistency.
- Maintain device identity, where physical machines, individual components, and sensors are tracked digitally both individually and as parts of subsystems.
- Handling time, as incoming data from sensors have different notions and representations of time, including none at all. The collection process must take these differing data streams and merge them into a coherent, unified data product, and this starts with defining a common definition of time and timestamps.
- Monitoring a chain of custody for all data flowing the system, which is critical for both troubleshooting and regulatory activities.

Another contributor to the explosion of data is related to regulators and government agencies, which makes it useful as input to system models [26, 29]. Domains that are developing Data Quality standards for regulator data include finance[11], healthcare [102], and the Food and Drug Administration [93]. In the energy domain, sources such as the Federal Energy Regulatory Commission [126, 127], and the Department of Energy [138, 150, 121, 119, 122] provide extensive data products. In fact, in 2018, the Foundations for Evidence-Based Policy Making act was passed, and one of its major sections, the Open Government Data act, required public government data to be published as machine-readable data. At the state level, many states publicly publish oil and gas and other energy related data [118, 147].

An important source of methane emissions data is the EPA’s Greenhouse Gas Reporting Program [17]. All active oil and gas facilities emitting 25,000 metric tons of greenhouse gas (GHG) over a period of one year are required to report the emissions expressed as their CO₂ equivalent [151]. In 2023, regulations have been strengthened so operators can be fined if they exceed emissions intensity levels set by Congress [111], and this has increased the economic impact of high-quality data.

1.3 A Process for Creating Energy System Simulations based on Large Data Sets

The combination of physical system modeling techniques and the availability of high quality data enables the development of very powerful simulation based analysis tools. The validity of results from these tools requires attention to the process by which they were developed, with several steps necessary for Quality Assurance (QA) of the data and simulation processes used. This process is seldom documented due to space considerations in publications.

Figure 1.1 shows a process for creating high quality simulations. It is divided into three high-level steps: *Data Management*, *Modeling*, and *Analysis*. Data management is the process by which data from many sources is collected, validated, and merged into a coherent data set. Simulations are performed during the Modeling phase, and, very importantly, results of the models are verified against reference data to ensure correctness of the models. Simulations are then run at scale, and during the Analysis phase, the results are examined and conclusions are reported.

1.3.1 Data Management

Data management is the process of collecting data, validating it, and merging it with other sources to create a coherent data set. Data can come from sensors via real-time monitoring, field surveys where staff are sent out to sites to collect measurements on an ad-hoc basis, or public data sources as described above. An important type of data for simulations is configuration data,

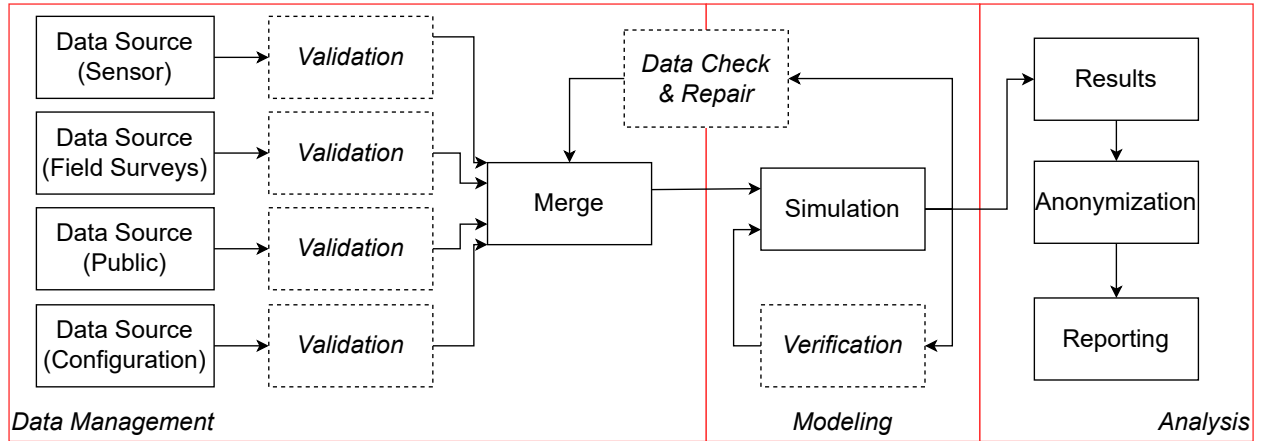


Figure 1.1: System Engineering Flow for Energy Simulations

where equipment types, sizes, and connections are defined. This configuration data can be used to construct a "digital twin" of the site, and can be used to model the site's behavior [67, 58, 98].

Once the data has been collected, it must be validated. Traditional data validation methods focus on rather trivial aspects of the data, such as whether fields are in the correct format or represent proper codes (for example, zip code fields contain values from a list of valid US zip codes) [117, 115]. In fact, these types of data validation processes are so commonplace that it is not unusual for organizations to define their own validation standards [20]. Validation of data used in simulations must be much more rigorous than generic validation. Each type of data has its own unique validation challenges:

- *Sensor Data* can be generated in huge amounts, providing usages challenges in just accessing the proper data. In addition, this data must be checked to determine whether readings fall within expected ranges. Reasons for the readings to be out of range include sensor malfunctions, communications errors – or the underlying physical process being out of control. Further, sensors are often placed in remote locations, and communication difficulties may lead to lost or delayed data values. Data validation must detect this and the overall simulation process must define strategies to deal with missing data.
- *Field Data* is captured by a team going to a site and taking measurements. These teams operate in varied weather conditions, and often visit multiple sites in a single day. A field

campaign can last several weeks or months, so personnel can get fatigued, which can lead to errors in data collection. In order to maintain the integrity of field data, records must be kept on who collected the data, what equipment was used, and where & when the data was collected. These records should be kept with the actual data to keep a history as an integrated data set is created.

Measurements are collected by instruments, and the collection team must take special that the equipment is operating correctly – that it is recently calibrated, set up for field conditions, and that initial sample readings are in expected ranges. If the equipment is to be run autonomously in remote sites, additional care must be taken that it continues to run unattended. Power must be addressed, telemetry must be established if necessary, and the system must be checked periodically to ensure continued correct operation.

- *Public Data Sources.* The organizations supplying this data are primarily responsible for its quality. At a minimum, a data dictionary is a critical resource for customers to understand how to use the data effectively. A description of the method of collecting and managing the data is useful to increase the confidence in data quality.
- *Configuration Data* is difficult to validate. Often the system or sites being modeled are large and complex, so digital records are essential. Moreover, as changes are made to the site, these records must be kept up to date. One way to ensure these records are accurate is to perform manual inventories of a site. If the site is large, a complete inventory is not feasible, instead subsets of the site can be "audited" to ensure consistency with the digital data. Perhaps a better technique to ensure the quality of configuration data is to create a digital twin, and run it against operational data. The results of these runs can be checked against operational outputs, and discrepancies used to debug the configuration information.

Merging data sets refers to the process of matching records from one data source with one or more other data sources. Sometimes this is an easy process, such as matching temperature records from a public data source with power consumption records. Even such a simple example can

introduce uncertainty in the model results. For example, if the temperature data source is in the same metropolitan area, then it is likely that there is relatively little error in temperature readings. If, however, the data source is much further away, or in a different climatological zone, then the likelihood of error increases.

A more significant issue with merging data sets is using keys to merge one data set with another, particularly if the data sets come from other organizations. Occasionally the keys are supplied in a different format – perhaps all keys represent a numeric value and are prefixed with zeros from one organization, and another organization removes preceding zeros. Or one organization uses dashes in the keys, and another organization does not – or puts the dashes in a different location in the key.

Finally, a consideration for data management is confidentiality of the data. A research organization often has access from another organization that has proprietary or Personally Identifiable Information (PII). In these cases, such information must be removed or anonymized before the final report is made public.

1.3.2 Modeling

Modeling is a very broad topic. Shishvan [43] provide a taxonomy of simulation models (see Figure 1.2). Models can be categorized as *deterministic*, where input values are fixed and output values are predictably calculated from these fixed input values; or *stochastic*, where at least one input variable is probabilistic. Models can also be either *static*, where time is not a factor; or *dynamic*, in which time-varying interactions between variables can occur. Finally, dynamic models can either be *continuous*, in which changes to the system occur continuously in response to changes in control variables. These models are typically defined in terms of differential equations. In contrast to continuous simulations, *discrete* models reflect changes to variables instantly in response to events.

There are two main QA activities during the Modeling phase – a check on integrated data, and verification of the model itself. The data check & repair activity, as illustrated in Figure 1.1,

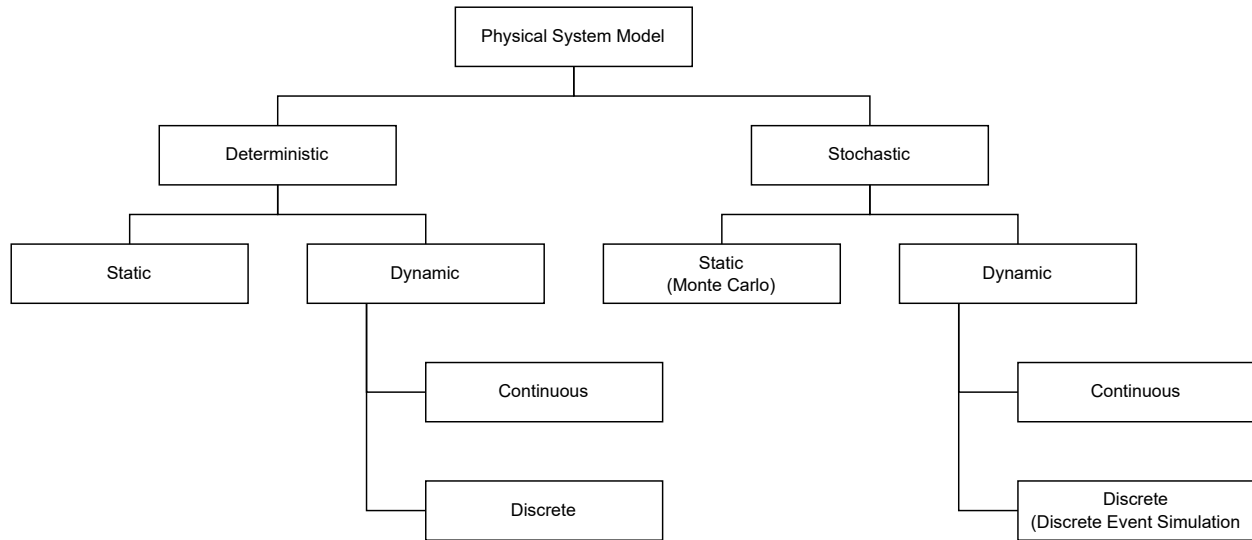


Figure 1.2: A Taxonomy for Model Types

spans the Data Management and Modeling phase because it often uses a subset of the simulation implementation to verify the merged data set.

Validating a simulation refers to ensuring the results of the simulation are correct – for a given set of inputs, the outputs of the simulation reflect the behavior of the underlying physical system. Validating the simulation is critically important, and is not covered very well in the literature. According to [13], “Validation is the most challenging task in simulation. Validation checks whether the results of the system actually match those of the real system.” And according to [4], “The modeling and simulation literature has largely neglected the referent, in the past failing to provide guidance about how to select or describe the information used as the standard for comparison in validation assessment.” Sargent [6] suggests many techniques for verification of simulations, including the following:

- *Historical data validation* uses historical data (if it exists) to compare against simulation output. This strategy needs to use identical input parameters to get the expected output results.
- *Internal validity* calculates expected results from input data and compares the output data against these expected results. This validation strategy is mostly useful for stochastic sim-

ulations to determine whether the range of output values are the expected results from the input parameter ranges.

- *Sensitivity analysis* determines which output values vary the most given changes in input parameters. Higher sensitivity input parameters should have more accuracy.
- *Extreme condition tests* ensures that the simulation behaves reasonably given extremes of input parameters (minimum values, maximum values, zero, etc.).
- *Comparison to other models* results of a simulation are compared with other known good model results, given the same input parameters for both.

Data check & repair is an activity that spans the Data Analysis phase and the Modeling phase, and its goal is to ensure the results of merging data sets are suitable for simulation. It can be viewed as an QA step for merged data. It typically consists of running the merged data through simplified simulations or subsets of the simulation software, identifying issues, fixing the issues in the merge step, and repeating the process until the merged data is of sufficient quality.

For example, as mentioned above, joining data sets across key values can have issues if the formats of the keys are different or the keys themselves are incorrect. A check for this is to find keys of either data set that have no match. These keys must be examined and potential matching keys found manually. There are two potential causes of such a mismatch – either there is a matching key in the other data set, in which case there is some kind of formatting error; or there is no matching key, in which case the record should be flagged and not used in the simulation. If there is a formatting error, it should be addressed. In many cases formatting errors are caused by a systematic problem, and a fix for one of these errors address a whole class of issues.

Another example of a class of problems with merging data is classification. With the AMI data set, there are cases where classification of premises to building type is ambiguous (for more detail, see Chapter 2). It is possible to use heuristic rules to eliminate this ambiguity, however each rule must be developed manually after looking at each case. After some point, this process reaches the point of diminishing returns, where further resolution does not add any value to the simulation,

and, in fact, may start removing value. For example, what does it matter if a random outbuilding is classified as Residential or Commercial if it has very low power usage? In these cases, it is useful to have an UNKNOWN or UNCLASSIFIED classification that can be used to identify all these problem premises, and, if necessary, account for them in the simulation results.

As a final example, configuration data often contains topology information i.e. data that identifies what components are connected to what. With a large set of such configuration data, errors can crop up in the topology data in the form of components (or sets of components) that are not connected to anything. These errors can be identified using network graph algorithms [134]. Again, problematic disconnected nodes should be flagged, corrected, and potentially used in error calculations.

1.3.3 Analysis

During the analysis phase, the results process takes the simulation raw output data and produces useful conclusions. The exact nature of this process is highly dependent on the specific goals of the simulation. Techniques for performing this analysis include aggregating data, statistical techniques (include error bounds calculations), and producing graphs and summary data tables.

For some simulation techniques, in particular Monte Carlo methods, enormous amounts of data can be produced. In this case, it can be useful process the raw data into more useful / performant forms. Techniques to do this include:

- Binary data formats
- Data partitioning
- Parallelization
- Pre-aggregation

These techniques can be implemented by putting the data in a database, which can sometimes be overkill, or using file-based technologies such as Parquet [113], HDF5 [135] or NetCDF [136].

As mentioned above, data sets may contain proprietary or PII data. In order for the data to be publicly released, it must be anonymized. Techniques to do so include generating random values for sensitive keys, or aggregation of values by either time or geography. Data anonymization can happen any time during the overall process. Performing this step just before final report generation facilitates debugging.

1.3.4 Automation

In order to robustly maintain the integrity of the overall simulation process, many of the steps in Figure 1.1 should be automated. It is often the case that some of the data sources will be updated during the lifetime of the simulation. Should this happen, manually re-running the QA steps and repairing data issues would be costly and error-prone. In addition, if the intent of the simulation is to provide results over time, it is worthwhile to automate the entire series of steps, including results processing, anonymization, and report generation.

Automating the QA steps serves an additional purpose – it provides a log of required data changes. This log can be used as an audit of the overall process and can be given to data providers to help them clean up their source data products.

1.4 Document Structure

This dissertation provides three case studies of projects that illustrate the simulation process described above. These case studies are:

- Chapter 2 shows the impact that the COVID-19 shutdown had on the electrical power usage in the city of Fort Collins, Colorado. It uses AMI data, a big data set, merges the data with property records to classify premises and aggregate usage according to the classification. It then presents a temperature-based model to compare 2019 usage to 2020 usage, verifying this model with historical models of power consumption over time.
- Chapter 3 examines the Methane Emissions Estimation Tool (MEET), a toolkit for modeling methane emissions. It uses data from published field campaigns and public data sources,

plus configuration data. The latter is used to create "prototypical sites," a way to mapping thousands of field sites into a manageable number of sites for simulation. This merged data set is then used to run Monte Carlo simulations that provide fine-grained temporal and geographic emission profiles for natural gas production and processing facilities.

- Chapter 4 offers an interesting variation of the simulation process, constraining the time step of the simulation to real-time (i.e., one millisecond of simulation occurs in one millisecond of clock time). The innovation in this project is to split the simulation across separate geographic sites, and techniques were developed to account for varying network delays between the two locations.

Finally, Appendix D describes typical problems found with data validation.

Chapter 2

Electricity Consumption in Commercial Buildings During Covid-19

2.1 Overview

Originally published in [103]

The first case study is a paper examining the impact of COVID-19 shutdowns on electrical power usage in the city of Fort Collins, CO. The paper used AMI data to develop a model of daily power consumption based on temperature – a fairly common modeling technique. What makes this paper unique is that it classifies premises into fine-grained usage categories (schools, restaurant, offices, etc.) to get a unique picture into how the shutdown impacted different commercial segments.

2.1.1 Data Management

Data sources used in the paper were the AMI data for Fort Collins, property classification for Larimer County, and temperature records. The AMI data set provides 15-minute energy consumption values for each premise in Fort Collins. Thus, for one year there are over 35,000 records for each premise, and there are more than 80,000 premises in the city, for a total of more than 2.8 billion records per year. In terms of access to individual premises alone, this provides a significant data management challenge from the sheer mass of the data. The AMI data is used to generate customer bills, and this provides an important financial incentive for the city to validate the data. Appendix A.2 describes the partitioning strategy used to provide performant access to individual premise consumption records by date. Each premise is mapped to a parcel ID maintained by Larimer county, and the parcel ID mapped to information about buildings on the parcel. Property usage data is maintained at the building level, so mapping a premise to precise property usage data

is often ambiguous. This problem is also described in Appendix A.2. For the purposes of this study, such ambiguous classification mappings were excluded.

2.1.2 Modeling

The modeling phase for the paper defined a functional relationship between the maximum temperature for the day and the daily power consumed by the premise. The process for doing this is defined in Section 2.3.2. Verification of this model was performed in two ways. First, the model was verified year-on-year, using historical data for previous years. Second, the COVID shutdown occurred in March, 2020, so January and February of 2020 were used for verification.

2.1.3 Related Works

Appendix A.1 shows the use of the AMI data set to define an alternate model that predicts Photovoltaic (PV) power consumption for an individual premise given only weather data, AMI data, and the approximate date of PV system installation.

Appendix A.2 shows the use of the AMI data and a much deeper modeling technique to perform a solar power hosting capacity analysis for the city of Fort Collins.

Abstract

The COVID-19 pandemic had wide-ranging effects on how people lived, worked and learned. Consequently, electricity use was altered from pre-pandemic patterns. The deviation from expected electricity usage patterns in commercial properties due to the COVID-19 pandemic was analyzed in a medium-sized American city. The focus on a single community a) allows usage to be linked specifically to the timing of public health and executive orders and b) provides a fine grained, detailed understanding of usage in different property classifications (e.g. restaurants, hotels, schools, outpatient medical facilities, offices, and religious organizations). Electricity consumption data from 2019, adjusted for average daily temperature, was utilized to calculate expected use in 2020. Electricity usage was found to be lower than expected for most commercial property

classes, but the timing and magnitude of these effects varied. For example, within the hospitality industry, hotels evidenced a larger and more sustained decrease in usage (-17%) as compared to restaurants (-11%). In addition, usage patterns for outpatient medical facilities can be linked to specific executive orders. Together, a heterogeneous rate of electricity use is found to vary according to specific types of commercial properties.

Practice Relevance

In much of the research regarding electricity usage, utility customers are placed into the broad categories of residential, commercial and industrial. These groups are viewed as homogeneous. In the present study, we examined different property classes within the commercial group during the COVID-19 pandemic. Although we found that electricity usage was lower than expected for most commercial property classes, we also discovered that the timing and magnitude of these effects varied. For example, within the hospitality industry, hotels evidenced a larger and more sustained decrease in usage than did restaurants. In addition, usage patterns for outpatient medical facilities were linked to specific executive orders. Together, these data sources allowed for a detailed examination of COVID-19's effect on electricity use for specific types of commercial properties and it revealed that their response to the pandemic varied widely. Thus, it is an oversimplification to view commercial properties as a single, homogeneous group.

2.2 Introduction

In 2020, many municipal, state, and national governments imposed quarantine shutdown protocols in response to COVID-19. This pandemic response resulted in what has been called 'the biggest psychological experiment in history' [65]. It changed how people worked, learned, shopped – virtually every area of life – and it affected how we used energy.

2.2.1 Energy use during the COVID-19 pandemic

The effect of the pandemic on energy use has been documented in a variety of large-scale studies. Specifically, studies conducted in Europe, South America, Australia, Asia and North America have shown overall reductions in energy use during COVID-related lockdowns (e.g., [63, 64, 70, 77, 84, 96, 88])). In addition, many of these same studies have described increases in residential energy use over the same time period (e.g., [70, 78, 81]). Thus, the overall decrease in energy use is attributable to reduced demand by the industrial and commercial sectors (e.g., [73, 70, 81, 96]). The majority of these studies have focused on energy use in countries (e.g., UK, Australia), large cities (e.g., Ontario, Santiago, Tokyo), or utility service areas with a million or more customers (San Diego Gas and Electric, Midcontinent Independent System Operator) and have used the broadscale utility classifications of residential, commercial and industrial. Together, they provide an excellent view of the impacts of the shutdown. However, [73] have argued that increased granularity, specifically in terms of spatiotemporal scale and level of aggregation, is needed in order to fully understand the effects of COVID-19 on energy use patterns.

A few studies have analyzed data with greater temporal granularity, showing how COVID-19 affected times of peak usage, particularly for the residential sector (e.g., [72, 94, 80, 95]). For example, [79] took a deeper dive into impacts of the pandemic on residential electricity patterns using high-resolution, disaggregated data for 225 housing units from 2018–20. They analyzed shifts in electricity use related to HVAC loads, non-HVAC loads, and whole-home loads and found increased electricity use during periods that occupants would usually be away from home. In addition, a few studies have examined data with greater spatial granularity, demonstrating how the timing and restrictiveness of government orders coincided with changes in electricity usage ([86, 87, 69]). However, very few studies on electricity usage during the pandemic have included a greater level of granularity with regard to customer class.

[76] provided a detailed analysis of electricity usage in higher education buildings, but these patterns were not compared to other sectors. In contrast, [78] , used weather-adjusted monthly utility data in S. Korea across a variety of customer types. They examined year-over-year change

in energy use during the first five months of 2020 and found that educational and research facilities showed the greatest decreases in consumption (16-32%), followed by (from largest to smallest): a) religious, b) hotels, c) sales, d) office, and e) medical. Similarly, [87] analyzed monthly resolution electricity data (January – September, 2020) in Chile across a number of ‘economic sectors’ (a mixture of industrial and commercial classifications). They found that in June 2020 when the quarantine was fully implemented, the hotel and restaurant category showed a 75% decrease relative to 2019. They also observed decreases in energy use for education (schools and universities- which were closed for the duration of the study) that were similar in magnitude to [78] (ranging from 11% to 36%). However, in contrast to other studies which showed reductions in the industrial sector, [97] found that the mining and quarrying sector actually increased their energy use in 2020. The authors explain that these industries were considered a critical service in Chile, which highlights the importance of specific COVID-19 policy decisions for electricity usage patterns during the pandemic.

In the present study, we sought to add to the literature demonstrating the effects of the pandemic on electricity usage by analyzing daily smart meter data from a medium-sized U.S. city from 2019-2020. By focusing on a single community we are able to a) link usage to specifically timed public health and executive orders, and b) provide a detailed picture of usage across different property classifications. As such, we present a nuanced view into the electricity consumption responses of varying business types resulting from the government-mandated changes in operation.

2.2.2 Context for the present study

In the U.S., there are more than 300 small and mid-sized cities and more people live in these cities than in America’s largest cities ([144, 148]). Fort Collins, Colorado is a community of approximately 175,000 residents. The Fort Collins health care system provides services to Northern Colorado, Southern Wyoming, and Western Nebraska. In addition, Fort Collins is known for a thriving micro-brew industry and a plethora of restaurants. In 2016, Fort Collins generated \$274.6

million in revenue from tourism ([39]). Overall, Fort Collins has an economic mix typical of many mid-sized cities in the United States.

Fort Collins has a municipal electricity provider, Fort Collins Utilities (FCU), that serves about 50 square miles of territory and 77,000 customers. In 2016, FCU established a research collaboration with Colorado State University (CSU) to share load interval data and other data sets for the purpose of detailed analyses and the creation of research techniques to analyze these large-scale sources of information. During the pandemic, Fort Collins Utilities performed a detailed analysis of the shutdown’s impact on electric power consumption in the city. This analysis provided a mechanism for tracking actual electric consumption versus expected consumption and was used to monitor the financial impacts of the shutdown. In this paper we use this analysis tool to integrate electricity use with detailed property classifications and the timing of public health orders to demonstrate the impact of COVID on electricity consumption across different sectors. This study adds to the literature on the impacts of COVID-19 on electricity use by providing a more fine-grained analysis than previously available.

2.3 Methods

2.3.1 Sample characteristics

In 2015, Fort Collins transitioned all of its electricity customers to Advanced Metering Infrastructure. These smart meters collect detailed 15-minute interval data for approximately 77,000 electrical accounts. Billing records allow customers (and thus loads) to be accurately categorized into high level groups such as Residential, Commercial, and Industrial. Further, each account is associated with publicly available building data maintained by the Larimer County Assessor. In contrast to traditional utility classifications which are based on their peak demand, the property classifications maintained by the county are based on the type of activity occurring at that premise and are far more specific. Thus, integrating meter data with property classifications enabled us to differentiate the usage patterns of property types such as offices and restaurants.

Account selection for analysis – utility classifications

Fort Collins Utilities provides electric service, metered at 15-min intervals, for approximately 77,100 accounts, with approximately 69,000 of these accounts representing residential customers. Locations that changed function (e.g., a location that was a hair-salon in 2019, then became a fast-food restaurant in January of 2020) and newly constructed buildings were excluded from the analysis. In addition, any location where there was a change of rate class (see below) or property type (see below) was removed from the analysis, ensuring the analysis maintained a constant number of accounts and account-types. After excluding these premises, a total of 76,912 premises were left for use in the analysis.

These remaining premises fell into five distinct rate codes: 1) Residential; 2) GS, 3) GS-25, 4) GS-50 and 5) GS-750. GS-25, GS-50 and GS-750 refer to accounts where the annual average of each month's peak-power demand is below 25 kW, between 25 and 50 kW, between 50 and 750 kW, or above 750 kW, respectively. Roughly speaking, the GS class consists of Small Commercial stores (like "Mom and Pop" stores), GS-25 is comprised of Medium Commercial groups (large retail stores, typically of a national chain franchise), GS-50 is Large Commercial (grocery stores, big-box store, office buildings, etc.) and GS-750 consists of Industrial customers. The GS-750 class was not included in the final analysis of this paper because the group was relatively small, containing only 15 accounts. Also, these accounts tend to be process-centric and their energy consumption is driven more by the production activities in their facilities than by a thermal model, as is intended for the IPMVP Option C model. This resulted in a final sample containing a total of 68,038 accounts in the Residential class; 7,549 in the GS class; 648 in the GS-25 class; and 677 in the GS-50 class.

Account selection for analysis – property assessor classifications

The Larimer County Property Assessor classifies properties according to the buildings (or improvements) on each parcel of land. There are 321 different options of varying detail and redundancy in the county classification system. This list was meticulously reviewed and condensed into 28 classifications. The mapping between these classifications underwent extensive manual

verification and intervention before being automated. The 28 classifications align with ENERGYSTAR Portfolio Manager building types, with the intent that it allows for regional building comparisons using Commercial Buildings Energy Consumption Survey (CBECS) data (Energy Information Administration, 2012). Thus, this classification was used in the present study to refine the ‘commercial’ classification used by the utility and enable a more fine-grained analysis of the effect of public health orders.

As of the time of this writing, when parcels contain multiple buildings, property data can only be connected to utility data at the parcel level, not the individual building level. This means that the county’s improvement data had to be aggregated to the parcel level. To address this, each parcel was assigned the classification for the largest building on the parcel. Unfortunately, this method was not effective for strip malls. These properties are identified by the County as strip malls, but typically contain units with a mix of retail, offices, food service, etc. Consequently, strip malls were excluded from the analysis.

Finally, there were three groups (Transportation, Parking, Residential HOA) whose consumption activities are process driven, mostly independent from thermal based consumption. As such, the IPMVP Option C is not applicable to them and, similar to the decision to exclude GS750 properties, these three groups were not included in the final analysis of this article.

After all of these considerations, the following groups were included in the analysis: Education (Poudre School District), Food Sales (Groceries), Food Service (Restaurants), Healthcare-Outpatient, Lodging, Mixed (combination of residential and commercial units), Office, Public Assembly (movie theaters, music venues, lodges, etc.), Religious, Retail-Mall, and Services (hair or nail salon, etc.). The number of premises within each of these classifications is shown in Table 2.1.

For clarity, in the education group we include only the Poudre School District (PSD) K-12 buildings. We exclude charter schools, community colleges, and local university. By using only Poudre School District (PSD) K-12 buildings, we ensure that all the selected facilities would follow the same lock-down schedule. Additionally, the results of the K-12 group are more widely applicable to other communities, as not all communities are home to large universities.

Table 2.1: Number of parcels associated with each one of the groupings from the ENERGYSTAR Portfolio Manager.

Premise Classification	Premise Count	Actual Consumption 2019 (MWh)	Actual Consumption 2020 (MWh)	% Change from 2019 to 2020
Business in Converted Home	145	7.9	7.6	-4.1
Education	142	46.8	40.3	-13.7
Education (Poudre K12 only)	70	42.5	34.7	-18.4
Food Sales	96	30.6	28.5	-6.8
Food Service	267	63.2	56.2	-11.1
Healthcare	35	6.1	9.3	51.9
Healthcare Inpatient	20	46.3	45.9	-0.8
Healthcare Outpatient	258	26.0	25.4	-2.6
Lodging	117	38.0	31.4	-17.3
Mixed	235	14.4	12.4	-14.5
Office	903	372.3	346.6	-6.9
Outbuilding	199	3.0	3.2	7.1
Parking	25	2.0	1.0	-52.7
Public Assembly	150	35.6	28.7	-19.2
Public Service	22	4.7	4.2	-11.6
Religious	111	16.0	14.0	-12.9
Residential HOA	5	0.1	0.1	-3.4
Residential Manufactured	1601	29.0	29.9	3.1
Residential Multi Family	6599	166.7	165.7	-0.6
Residential Single Family	42807	938.7	983.1	4.7
Retail	375	78.6	73.0	-7.1
Retail Mall	353	42.2	35.7	-15.4
Services	183	20.5	19.6	-4.3
Storage	113	8.2	8.0	-2.1
Transportation	6	0.9	0.8	-8.8
Utility	7	24.4	22.7	-6.9
Healthcare Vet	33	5.3	5.3	0.2

2.3.2 Procedure for modeling expected electricity usage

International Performance Measurement and Verification Protocol (IPMVP)

Many techniques exist to determine the energy impact of system updates or improvements at the building-level. This paper extends one such technique, IPMVP Option C [145] to examine the effects of the pandemic. Energy management professionals are familiar with IPMVP Option C, which is typically used when a specific building has multiple systems updated or improved in an energy efficiency retrofit (e.g., LED lighting installation, window replacements, and/or building automation system optimization) and an evaluation of the energy savings from the retrofit(s) is needed. In such a case, the overall energy consumption of the facility is measured via utility meters, typically over a period of 12 months before (baseline period) and 12 months after (reporting period) the intervention. Weather variations over the baseline and reporting periods are controlled via a linear regression on Heating Degree Days (HDD) and Cooling Degree Days (CDD).

In the current study, the IPMVP methodology was used to examine the effect of the COVID-19 shutdown on electricity usage across the whole City of Fort Collins. This is a valid generalization of the IPMVP methodology because the pandemic created multiple modifications to a “facility,” by changing occupancy patterns of commercial facilities (capacity limits in stores, different operating hours) and residential consumption patterns (working and studying from home, frequency and timing of clothes washing loads, dishwasher loads, TV watching hours/week, number of meals cooked at home, etc.). When considering model selection, it is important to note that IPMVP has four options from which to choose. Option C is recommended to be used when a) the impact is expected to exceed 10% over the baseline period and b) the only available data is 12 monthly utility bills. However, Option C is an appropriate choice, even when the expected deviation is less than 10%, when more granular data are available. In the present study, fifteen-minute interval smart meter data from January 1, 2019 through December 31, 2020 was aggregated into 24-hr windows.¹ Thus, smart meters provided the actual consumption before and during the pandemic

¹ Data for October 26th 2020 was excluded from the analysis because air quality in Fort Collins was significantly, adversely affected by the Cameron Peak and East Troublesome Fires – the two largest wildfires in Colorado history.

while the expected consumption (i.e., consumption in the absence of a pandemic) was estimated by a modification of the IPMVP-Option C method as described below.

Weather normalization procedure

Fort Collins is located in a region with a semi-arid climate with four distinct seasons and low annual precipitation. Summers range from mild to hot (average temperature in July is 73.1°F / 23°C) with low humidity. Winters range from mild to moderately cold (average temperature in December is 31.2°F / -0.4°C). Annual snowfall is about 50 inches (1.3 m) and average precipitation is 15.9 inches (400 mm). The city receives plenty of solar radiation with 300 days of sunshine per year. Given these clearly demarcated weather patterns, meteorological conditions play a significant role in electricity consumption. Temperature is a primary indicator of electricity usage, while solar radiation, humidity, and previous-day-average temperature are second-order corrections to temperature when modeling electricity consumption. In this paper we limited ourselves to using temperature as the sole weather parameter and ignored any hysteresis properties of weather on electricity consumption

In Fort Collins, heating is most commonly provided by gas-burning, forced-air systems. In contrast, cooling is provided by electric air conditioning units. Thus, the relation between temperature and electricity consumption differs for cold and warm temperatures. The electricity consumption of a group of accounts can be parameterized by two temperature splines (above and below the equilibrium point for heating and cooling, 55°F / 13°C) and by using the day-type classification (weekday, weekend) as a secondary parameter. As shown in Figure 2.1, a linear fit is applied to temperatures $\leq 55^\circ\text{F}$ / 13°C (left) and a parabolic fit is applied to temperatures $> 55^\circ\text{F}$ / 13°C (right).

Average daily temperatures for each day in 2020 were plugged into these regression equations to create daily expected consumption, that is, what the energy consumption *would have been* in the absence of a pandemic, by using the average temperature for each day (in the measurement

These fires, which occurred almost simultaneously, resulted in residents being instructed to stay indoors with their windows closed when possible.

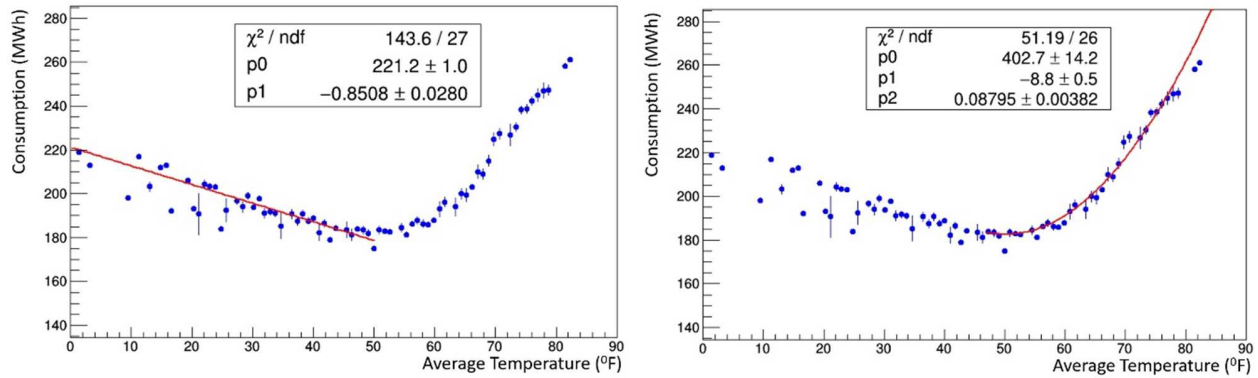


Figure 2.1: Regression of energy daily consumption vs average daily temperature (for weekdays in this case) for a given subset of accounts.

year, 2020). A single daily temperature was used for the entire service territory. By comparing the actual consumption from the metering data with the expected consumption, it is possible to estimate the impact of the COVID-19 pandemic as the ratio of the two values (i.e., as a percentage increase or decrease).

The analysis for the education group required an additional parameter for its model, which was an indicator variable for “in-session” or “in-recess”. This binary variable accounted for summer vacations, spring break, etc. Therefore, we modelled energy usage for day-type (Weekend-Weekday) for both in-session or in-recess periods, based on the IPMVP model, and applied the relevant set of parameters to each day during 2020. We considered the remote teaching period as “in-session”, as the buildings were unoccupied but classes continued.

2.3.3 Timeline of COVID-related events in Fort Collins

Table 2.2 contains a timeline of important COVID-related events in Fort Collins (collected from [143, 112]). As this timeline shows, numerous actions were taken by the Colorado Governor and the Colorado Department of Health and Environment (CDPHE) in March, 2020 which affected every sector. However, over the remainder of the year, there was a de-coupling of these orders which allowed us to examine the effects of these orders on specific sectors.

Table 2.2: Timeline of relevant COVID-related events in Fort Collins, Colorado

Date	Event	Affected Group(s)
3/9/20	First COVID-19 case in Fort Collins	
3/11/20	Gov. Polis declared a state of emergency	
3/15/20	Colorado State University suspended in person learning and closed the university to all but critical personnel. In August, CSU partially opened for hybrid learning, with most classes still online and was fully online from 11/21/20 through the end of the year.	Residential
3/18/20	Gov. Polis suspended in-person instruction for all K-12. PSD returned to hybrid learning (50% capacity) from 10/5/20 – 11/23/20 and was fully online for the remainder of the year.	PSD, residential
3/19/20	CDPHE issued a public health order that limited all mass gatherings to no more than 10 people [started being relaxed 6/1/20]	All except medical
3/19/20	Gov. Polis suspended all elective and non-essential surgeries and medical procedures [through 4/26/20]	Medical outpatient
3/25/20	Gov. Polis issued a stay at home order. On 4/26/20 it was replaced with the less restrictive safer at home order [through 6/1/20]. From June through December, restrictions (e.g., capacity limits) varied according to the number of cases and hospitalizations, but there were no further stay at home orders.	All except medical

2.4 Results

The COVID-19 pandemic arrived in full to Northern Colorado in early March 2020. This means that we can calculate the *expected* consumption for the first two months of 2020 (before the full impact of COVID-19) and compare it against the actual consumption for these two months to validate our model.

2.4.1 Utility classification results

The IPMVP Option C model for customer class was done via GUI in ROOT ([146]), carefully fitting the data and uncertainties to obtain the lowest possible Chi-Square. Figure 2.2 shows a graph of *absolute expected* versus *absolute actual* energy. Since Industrial (GS-750) accounts were excluded from the study, they are not included in Figure 2.2. Despite daily variations due to weather, it is possible to observe that the model follows the actual energy consumption for all classes quite well until mid-March 2020, when the overall electricity use became noticeably lower than predicted. This pattern lasts until June, and we see another drop in the fall. This reduction is driven by the commercial sector which started to have a significant drop in consumption in March. Although this pattern varied in magnitude across the year, it was largely sustained through the end of 2020. In contrast, the residential sector performed largely as expected, seeing only a limited increase in consumption during the summer months.

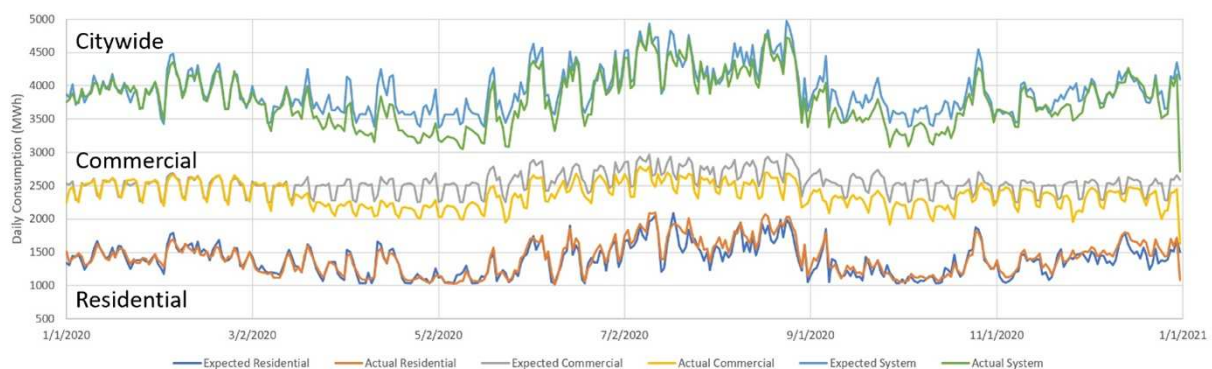


Figure 2.2: Breakdown of Expected vs Actual daily energy consumption broken down by commercial and residential sectors.

Figure 2.3 shows a breakdown by Utility rate code. All groups were normalized to their expected value, as to more clearly show the drop in consumption starting in mid-March 2020. The plot has been smoothed out by creating a 5-day rolling average of the ratios. Clearly the residential group has much larger fluctuations than the commercial but does not show a significant decrease or increase in consumption.

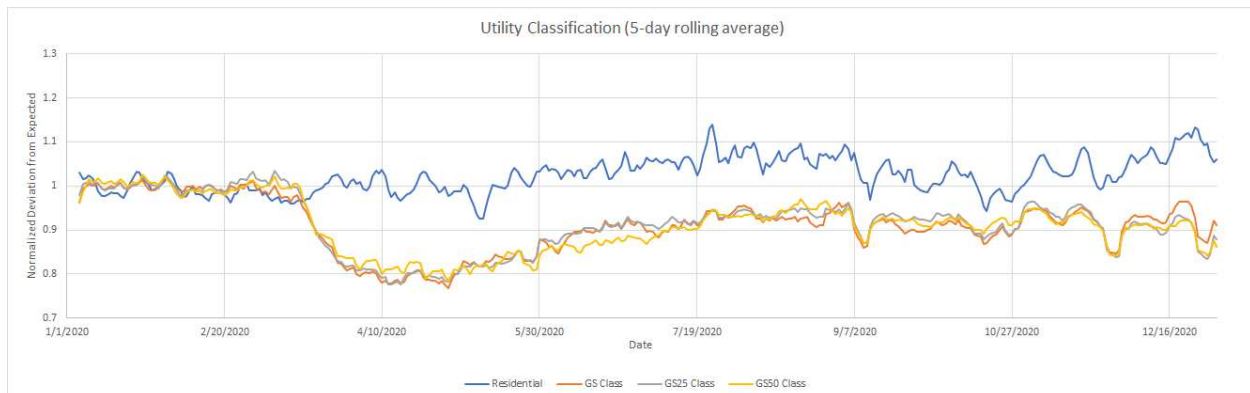


Figure 2.3: Daily consumption deviation from expected.

Table 2.3 presents a monthly summary with customers segmented by Utility rate code. Shaded values are statistically significant, while bold values indicate a decrease of the actual value relative to the expected value. Consistent with findings from other researchers, commercial customers of all sizes showed statistically significant declines in their electricity usage in 2020 compared to the expected value. Moreover, the decreases are most pronounced in April through June when government imposed restrictions were strongest. Residential patterns show less variability, with the summer months demonstrating single-digit increases relative to expected values. We should note that the presence of the university in Fort Collins limits the generalizability of our results in the residential classification. Specifically, in March 2020, not only were CSU classes moved online, but the university essentially told students not to come back after spring break. There was a concern that if tens of thousands of students returned to Fort Collins after traveling that this would fuel the spread the COVID-19 in the community. Thus, many residences were empty, especially

from March to August 2020. This likely accounts for the fact that observed increases in residential use were smaller than those reported by other studies.

Table 2.3: Monthly deviation from expected consumption. Note: Shaded values are deviations that are statistically significant and bold shows reduction relative to the expected value.

Class	Month, 2020											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Residential	0%	-1%	-1%	0%	0%	4%	6%	6%	3%	-1%	3%	7%
GS	-1%	0%	-9%	-21%	-16%	-11%	-8%	-7%	-9%	-10%	-9%	-8%
GS-25	0%	0%	-8%	-21%	-17%	-10%	-8%	-6%	-8%	-9%	-8%	-11%
GS-50	0%	-1%	-7%	-19%	-17%	-13%	-9%	-6%	-8%	-8%	-9%	-11%
System	0%	0%	-3%	-7%	-7%	-3%	-2%	-1%	-2%	-3%	-3%	-1%

These results largely mirror those found by other researchers around the world. It is noteworthy that the rate classes, which represent different sized commercial entities, showed remarkably similar patterns of use. However, as we shall demonstrate, these results obscure the heterogeneity that exists within the commercial sector. In the following section, we examine these entities, not according to their size, but rather by their function.

2.4.2 Detailed property classification results

To develop a deeper understanding of the energy impact of the pandemic, commercial customers were further subdivided into the Property Classification Schema described earlier. Table 2.4 presents a monthly summary with customers broken down by Property Classification. Shaded values are statistically significant, while bold indicates a decrease of the actual value relative to the expected value. These data show that for virtually every class of commercial entity, there were significant decreases from the expected consumption from March 2020 through December 2020. The only exceptions to this were in services (e.g., hair and nail salons) and outpatient medical facilities.

Although almost every group showed significant decreases throughout the year, there was temporal variability consistent with the timing of executive and public health orders. For instance,

Table 2.4: Deviation on energy consumption broken down by different commercial sectors. Note: Shaded values are deviations that are statistically significant and bold shows reduction relative to the expected value.

Property Class	Month, 2020 (First Half)					
	Jan	Feb	Mar	Apr	May	Jun
K-12 Schools	1.0%	-1.0%	-20.0%	-51.0%	-51.0%	-39.0%
Food Sales	0.6%	0.4%	-3.8%	-8.6%	-9.8%	-10.2%
Food Service	-1.3%	0.6%	-10.6%	-28.4%	-24.1%	-13.0%
Healthcare Out	3.1%	3.8%	-3.8%	-13.9%	-4.6%	-2.9%
Lodging	-1.6%	0.4%	-14.6%	-33.7%	-35.5%	-25.8%
Mixed	-2.0%	0.6%	-9.5%	-30.4%	-30.4%	-16.1%
Office	0.1%	-0.3%	-3.5%	-8.0%	-8.3%	-9.5%
Public Assembly	0.1%	0.3%	-16.4%	-37.2%	-35.7%	-26.4%
Religious	4.7%	5.9%	-5.8%	-19.0%	-23.3%	-20.3%
Retail Mall	1.6%	-0.2%	-11.5%	-30.9%	-27.1%	-23.1%
Services	6.7%	7.3%	-1.7%	-21.5%	-12.8%	-5.3%

Property Class	Month, 2020 (Second Half)					
	Jul	Aug	Sep	Oct	Nov	Dec
K-12 Schools	-18.0%	13.0%	-29.0%	-18.0%	-12.0%	-9.0%
Food Sales	-8.4%	-8.3%	-10.0%	-10.0%	-9.7%	-9.4%
Food Service	-9.5%	-8.2%	-9.4%	-8.5%	-11.5%	-19.0%
Healthcare Out	-2.7%	-2.3%	-3.6%	-3.3%	-1.9%	0.1%
Lodging	-17.1%	-12.9%	-17.2%	-13.4%	-14.8%	-13.5%
Mixed	-13.3%	-12.9%	-16.7%	-16.6%	-15.6%	-13.5%
Office	-11.6%	-9.8%	-8.3%	-7.7%	-7.4%	-8.2%
Public Assembly	-22.4%	-20.5%	-18.2%	-18.5%	-17.2%	-19.7%
Religious	-19.1%	-16.8%	-18.3%	-13.2%	-12.9%	-13.9%
Retail Mall	-20.2%	-18.7%	-16.9%	-15.2%	-10.8%	-12.5%
Services	-5.1%	-2.0%	-4.6%	-5.0%	-6.5%	-4.1%

outpatient medical facilities showed a significant decline in usage following Governor Polis’ executive order D 2020 009 which banned elective and non-essential medical procedures, but quickly rebounded after the order was rescinded (see Figure 2.4).

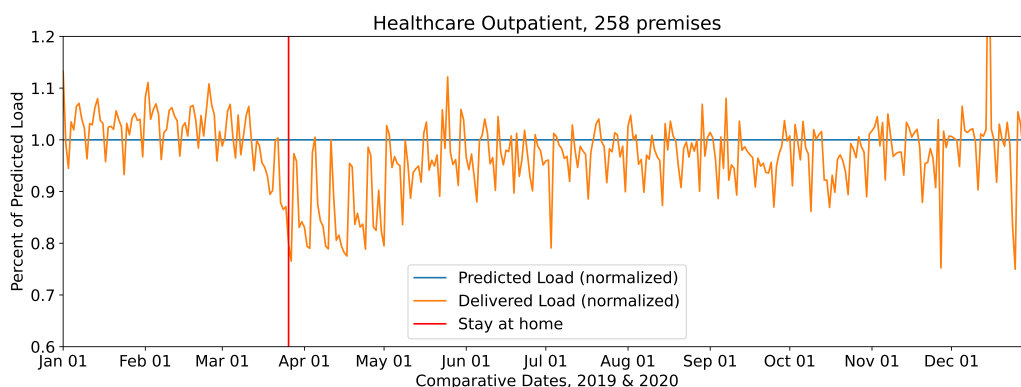


Figure 2.4: Daily consumption deviation from expected for the Healthcare Outpatient classification.

We can trace similar temporal patterns in the data for schools. Figure 2.5 shows actual 2020 usage as a percentage of expected use for the local public school system. K-12 buildings showed substantial reductions in use during the time that instruction was fully online (April, May and September). Smaller decreases were evident when children went back to school in October and November. These smaller reductions are likely due to reduced use of PSD buildings outside of regular school hours because extra-curricular activities were curtailed and the school district did not rent out buildings for community events due to increased cleaning protocols. An increase in usage in August was likely the result of a concentrated effort by district facilities professionals to increase ventilation and filtration in the buildings prior to the resumption of in person learning. This deviation was confirmed with the PSD energy manager (personal communication) who explained that the schools were run as if occupied after being equipped with new MERV-13 filters.

In addition to temporal patterns, our detailed analysis revealed variations in the pandemic’s impact within the commercial sector. The hospitality industry (food service and lodging) showed substantial decreases in electricity usage in April and May (see Figure 2.6 and Figure 2.7). They maintained smaller decreases throughout the rest of the year. Notably, these reductions are larger

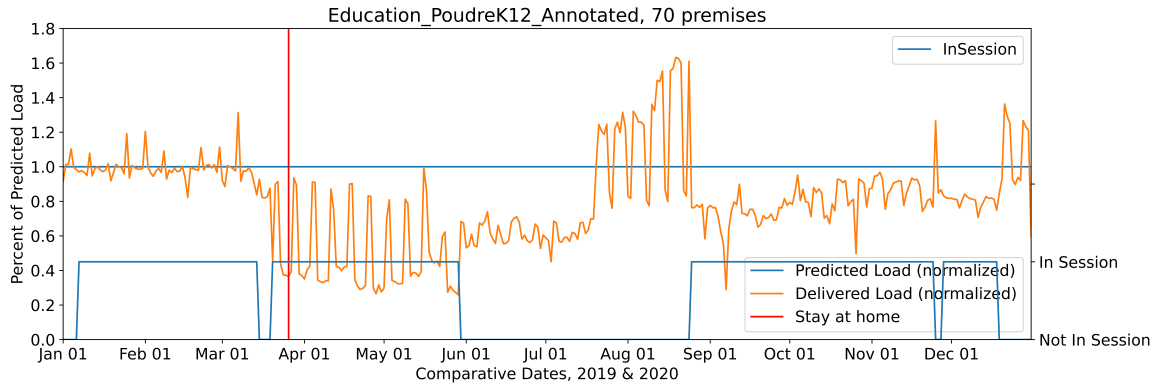


Figure 2.5: Daily consumption deviation from expected for the Poudre School District (K-12 schools) Property Classification.

for the lodging industry (Hotels and Motels but not AirBNB or VRBO-type of units) than for restaurants. This could be due to usage scaling more with occupancy for hotels (e.g., HVAC set points) or, alternatively, it might be due to the availability of a take-out option for restaurants (which would entail the kitchen being in use even when the dining area was not). To distinguish these potential explanations, we examined 2020 sales tax revenue data for Fort Collins. These data support the take-out explanation: restaurants saw a 16.4% decrease compared to 47.2% for lodging (Financial Services, City of Fort Collins, January 11, 2021). Even with a take-out option, however, the restaurant industry showed a noticeable dip in late November and December 2020, which coincides with an increase in COVID-19 cases and hospitalizations in the county.

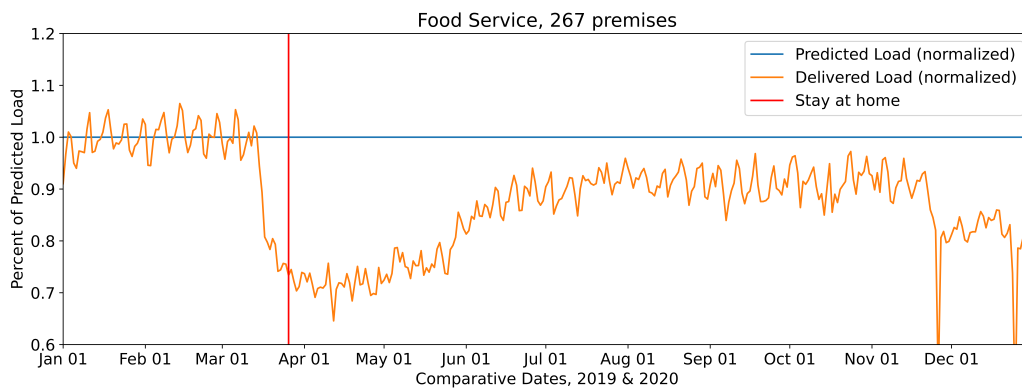


Figure 2.6: Daily consumption deviation from expected for the Food Service Property Classification.

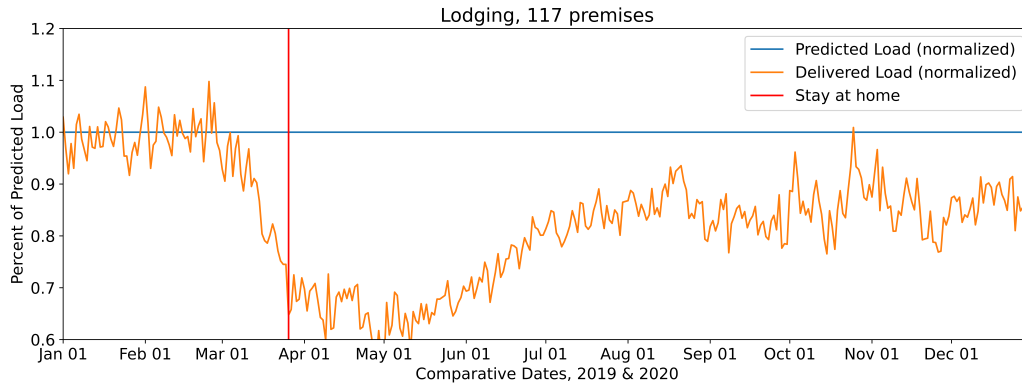


Figure 2.7: Daily consumption deviation from expected for the Lodging Property Classification.

In contrast to the hospitality, healthcare, and educational sectors, which showed substantial temporal variability across 2020, there were other groups that were less variable. The public assembly and religious categories (Figure 2.8 and Figure 2.9, respectively) maintained double-digit reductions for the entire year. Initially, gatherings of 10 or more were prohibited. Even after these prohibitions were lifted, there were substantial limits placed on these activities by the Health Department, and many houses of worship chose to remain closed to protect their congregants.

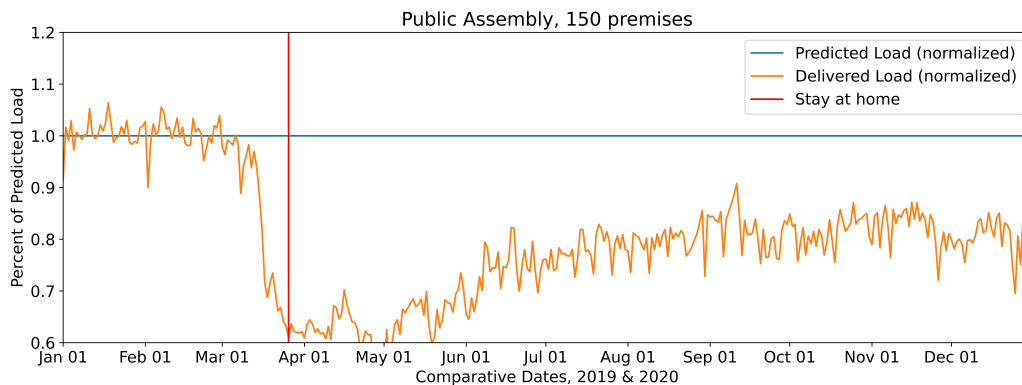


Figure 2.8: Daily consumption deviation from expected for the Public Assembly Property Classification.

The Office group also showed a gradual drop (suggesting that businesses were starting to go remote even prior to the Stay-at-Home order) and it became fairly stable and predictable immediately after the Stay-At-Home order. It is the only group that became steady and predictable starting

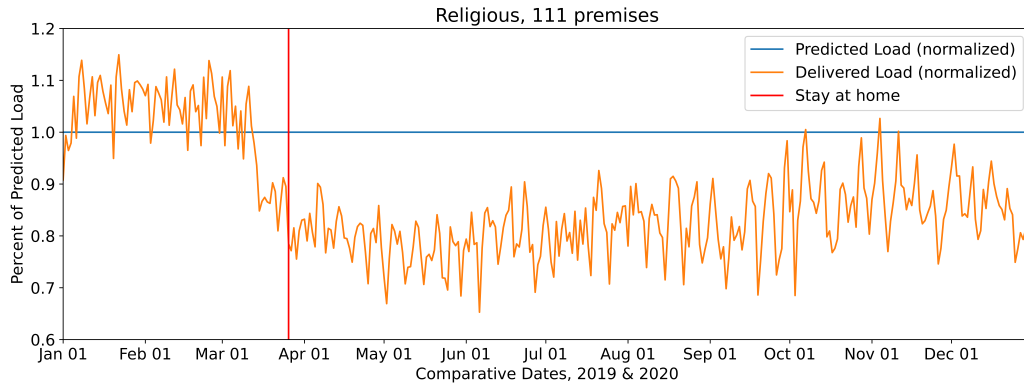


Figure 2.9: Daily consumption deviation from expected for the Religious Property Classification.

on that day (see Figure 2.10), indicating that the majority of remote workers were not back to the office during the period under study.

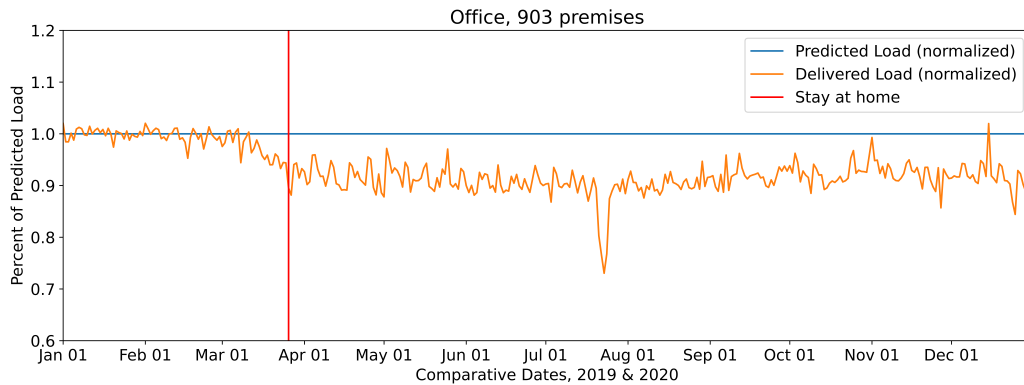


Figure 2.10: Daily consumption deviation from expected for the Office Property Classification.

Finally, the food sales category showed single digit, but significant, decreases for the whole year. This result is surprising because these businesses were classed as critical and their operations were not substantially curtailed by the government. However, some grocery stores limited their hours of operation due to staffing shortages and to give overworked staff time to clean and restock [139].

To summarize, we found that the commercial sector showed significant decreases in electricity consumption compared to what was expected based on 2019 consumption, and that the residential sector showed a small increase. However, groups within the commercial sector demonstrated

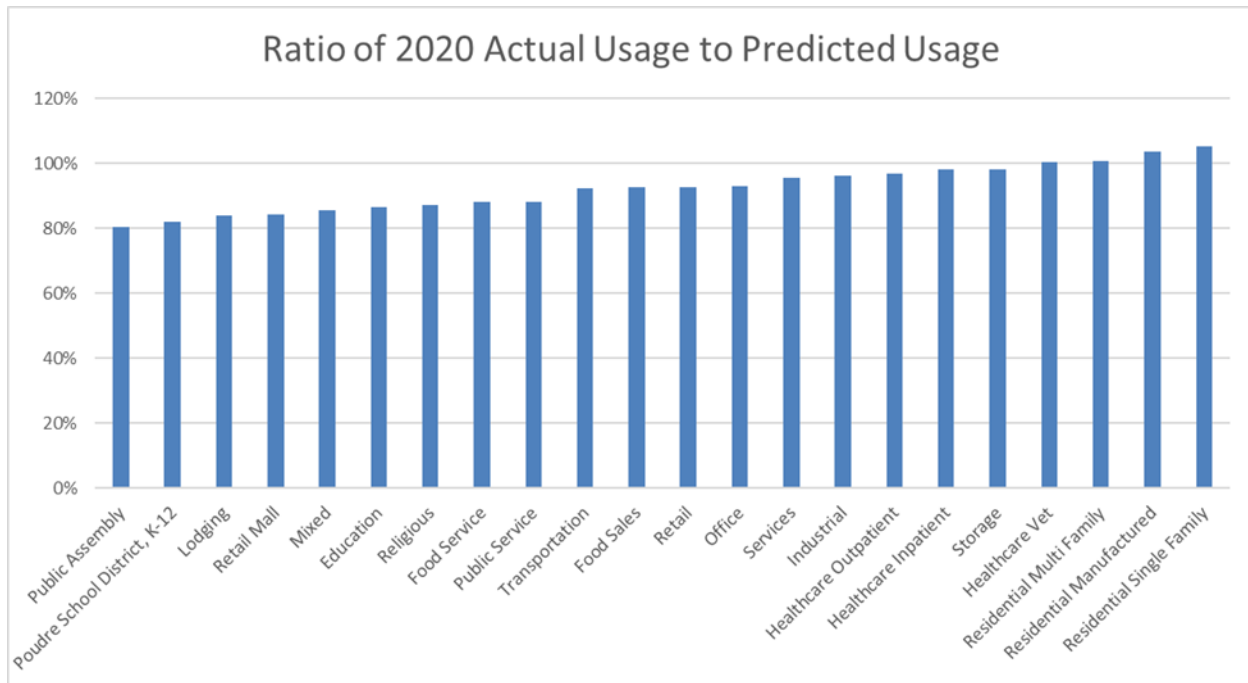


Figure 2.11: The ratio of actual to predicted energy use in 2021 grouped by property classification.

different magnitudes of change (see Figure 2.11) and also varying temporal patterns which often aligned with executive and health orders that impacted them directly.

2.5 Discussion

2.5.1 Findings and implications

In the present study, we focused on the impact of COVID-19 on electricity usage in the commercial sector in a small city. In doing so, we have been able to provide a more detailed and nuanced description of the effect of executive and health orders on hospitality, education, health-care, and other commercial property classes. This work contributes to, and extends, the literature on the energy impacts of the COVID-19 pandemic. To our knowledge, this is the first study of this scope to analyze the impact of the pandemic at such a granular level while grouping meters by both customer and business type.

Taking the view adopted in other studies, we found that from March to June, and again in late fall, overall energy consumption patterns were disrupted. The decreases in the commercial

sectors were significant and sustained. Unlike other studies, we observed only a small consumption increase in the residential sector (in the summer). Thus, the overall effect was a total reduction in energy consumption of 2.5%. This created a revenue shortfall and significant financial cash-flow issues for the utility. As a benchmark, about 6.7% of the annual utility expenses are allocated to staff salaries, therefore a 2.5% reduction in energy consumption represents a significant impact on the utility's financial operations and overall viability of the business.

Although researchers and utility professionals refer to the 'commercial sector' generically, the truth is that there is substantial variability within this group of utility customers. By combining smart meter data with information from the county property assessor, we were able to reveal heterogeneity within the commercial class that was undetectable when focusing purely on rate classifications. For instance, restaurants and hotels are often considered together under the umbrella of 'hospitality,' but these businesses showed very different patterns of electricity usage across the pandemic.

These patterns also shed light on how people responded to the pandemic and generate more research questions. For instance, once restrictions were lifted, energy consumption for the services sector (predominantly hair and nail salons) rebounded, but religious institutions showed sustained decreases in usage. Is this because of choices by the institutions themselves of whether to return to business as usual or are they reflective of people's choices and priorities about the activities they are willing to curtail? In fact, examination of the plots shows that for every group, usage began to decline before Gov. Polis issued a stay-at-home order. Thus, the effects on energy usage observed in this, and other studies, are not purely due to the enactment of formal policies, but also to people's perceptions regarding their safety.

Perceptions of safety apply not only to customers, but also to business owners. The Public Assembly group (theaters, movie theaters, and concert venues) and the office group never fully recovered in 2020, even after restrictions were lifted. Many entertainment venues canceled shows or limited capacity, and many remote workers stayed home throughout 2020, even without explicit policies requiring them to do so.

Some of the variability across groups is likely due to the individual business' ability to adapt to changing policies and perceptions. For instances, healthcare facilities are accustomed to rigorous safety and cleaning protocols and could therefore resume operations quickly once restrictions were lifted. In contrast, Poudre School District was slower to respond, resulting in a decision to delay in-person classes until October 2020. Thus, there are implications for organizational researchers to explore in this regard.

2.5.2 Limitations

Neither the energy usage data nor the property classification were created for the purposes of research. For example, some anomalous load behavior exists that were not rigorously examined:

- Two significant dips in consumption occur in the Food Service segment. The dates for these decreases correspond to Thanksgiving and Christmas, and could possibly be attributed a decrease in holiday customers during the COVID shutdown.
- Consumption in the “Office” segment had a significant dip at the end of July, 2020 compared to July 2019. The data showing this dip is valid, but the actual cause of the decrease was not investigated.

It was found that model reproduced the actual energy consumption during January and February of 2020 quite well. In all cases the actual consumption was only 1% to 2% systematically higher than the modeled consumption across all classes. That difference could be considered an econometric energy impact due to the strong economic growth experienced through 2019 carried over first few months of 2020, as it is consistent with previous year-over-year increases. These observed increases in early 2020 suggest that the decreases we have reported (as compared to expected use) are likely to be underestimates of the actual effect of the pandemic.

We recognize that there are limitations to the generalizability of our findings. First, as mentioned previously, Fort Collins is home to a large university. It is likely that the exodus of the majority of the students in March 2020 impacted the residential usage patterns. Second, the specific results we have obtained (e.g., a drop, partial rebound, and another drop for the food service

industry) should not be viewed as reflective of the patterns likely to be observed in other communities. However, we believe that the mechanisms for this pattern (stay-at-home order drop; cases decline and restrictions lifted rebound; cases increase drop) are likely to generalize to other locations and different-sized communities. These are the key take-aways from this research. While our results may not be directly transferable to other locations, due to a unique industry mix, they do provide insight to the variability of impacts across industries and the effect of COVID-19 policies on those industries.

2.5.3 Conclusion

Often we think of buildings or machinery using energy, but what energy usage during the pandemic demonstrates is the effect of behavior on energy use. People's behavior changed, and energy consumption changed with it. It is difficult to overestimate the impact the COVID-19 pandemic had on our lives. In this paper, a model to predict energy usage based on past data and outdoor temperatures was implemented, with the goal of quantifying the impact of the COVID-19 pandemic on energy usage in a medium-sized city.

Two years of 15-minute interval smart meter data was combined with property classification data to create a rich dataset, with a substantive sample size, that enabled us to examine a wide variety of impacts. The ability to drill down to the level of business type uncovered unique consumption patterns.

This work contributes to the literature on the COVID-19 pandemic by presenting the results of a year-long study of the pandemic's impact on electricity usage patterns of approximately 77,000 premises. The large dataset provides a unique level of granularity and depth that may not be easily replicated in other studies.

2.6 Data Availability Statement

The data used in this research project have not been made available, as Fort Collins Utilities complies with the "Red Flags Rule" from the United States Federal Trade Commission which mitigates Identity Theft. Under this rule, energy consumption cannot be disclosed.

Chapter 3

Modeling air emissions from complex facilities at detailed temporal and spatial resolution: The Methane Emission Estimation Tool (MEET)

3.1 Overview

Originally published in [100]

This second case study describes MEET, a tool that produces fine-grained temporal and geographic estimates of methane emissions. A typical inventory focused emission estimate uses average equipment / site counts and emission factors. This works well for producing a high level value for a particular time period (e.g. annual estimates), but it does not capture the variation in emissions that are typical for actual sites. MEET improves this method by using discrete event simulator (DES) to model equipment behavior and emissions over one-second intervals, which gives a detailed picture of emission patterns over time.

3.1.1 Data Management

MEET ships with published field survey data from recent campaigns (for example [19, 18, 8, 62]). It also includes data from regulatory agencies ([17]). It is easily extensible to include data from other sources, include other basins and even proprietary operator data.

3.1.2 Modeling

MEET has been used to estimate emissions across the Denver-Julesburg Basin (DJ), which includes more than 2000 operating sites [110]. Site parameters in MEET are defined in a "site definition file." To avoid defining one file per site, "prototypical sites" are defined which represent sets of operating sites [109]. In all, thirteen prototypical sites were defined for that study, which

roughly correspond to generations of site designs from the operator point of view. This is the source of Configuration Data described in Figure 1.1, and was validated by in-person meetings with operators.

Significant attention was paid to verifying the models used in the DJ estimate. Multiple cycles were performed where the models were run and the results checked against operator reported figures. Any significant discrepancies were identified, models modified to eliminate these differences, the results re-validated, and the process repeated as necessary.

A follow on release of MEET added the *Mechanistic Models*, or the ability for changes in one piece of equipment to affect the behavior of a downstream piece of equipment. An example is a stuck dump valve in a separator, which allows vapor from the separation process to be passed to downstream equipment, such as a holding tank. This increased vapor in the tank might cause a pressure relief valve to open. These mechanistic models add to the realism of the simulations and their acceptance to the community.

3.1.3 Analysis

The original version of MEET wrote all its results into a combination of .csv and .json files. Analysis software would then have to parse these files before performing any useful work. Revisions to MEET included a step where these results were cleaned up and written to binary formatted files (Parquet, [113]), which decreased the time required to access the results by orders of magnitude. This significantly increased the size of simulations which could be run, both in complexity of sites and numbers of Monte Carlo iterations.

3.1.4 Related Works

- Appendix B.1 provides information on the publicly available data sources (GHGRP) and studies conducted at CSU. It also provides a description of the Monte Carlo simulation process.

- Appendix B.2 gives background information on the equipment modeled in MEET and the state machines used to describe their behavior.

Abstract

Recent attention to methane emissions from oil and gas infrastructure has increased interest in comparing measurements with inventory emission estimates. While measurement methods typically estimate emissions over a few periods that are seconds to hours in length, current inventory methods typically produce long-term average emission estimates. This temporal mis-alignment complicates comparisons and leads to underestimates in the uncertainty of measurement methods. This study describes a new temporally and spatially resolved inventory emission model (MEET), and demonstrates the model by application to compressor station emissions – the key facility type in midstream natural gas operations. The study looks at three common facility measurement methods: tracer flux methods for measuring station emissions, the use of ethane-methane ratios for source attribution of basin-scale estimates, and the behavior of continuous monitoring for leak detection at stations. Simulation results indicate that measurement methods likely underestimate uncertainties in emission estimates by failing to account for the variability in normal facility emissions and variations in ethane/methane ratios. A tracer-based measurement campaign could estimate emissions outside the 95% confidence interval of annual emissions 30% of the time, while ethane/methane ratios could be mis-estimated by as much as 50%. Use of MEET also highlights the need to improve data reporting from measurement campaigns to better capture the temporal and spatial variation in observed emissions.

3.2 Introduction

Methane is a powerful greenhouse gas and atmospheric methane concentrations have increased steadily since 1980 [149]. Evidence indicates that a substantial portion of the increase is due to anthropogenic sources, including fossil fuel development [59, 90, 68]. This increase coincides with increases in natural gas production globally, driving interest in methane emissions from the

natural gas supply chain as well as emissions of hazardous air pollutants and CO₂. Multiple recent studies have also indicated that large emitters from oil and gas (O&G) operations are likely under-represented in current inventories of O&G methane emissions and that these emitters likely originate from abnormal process conditions at O&G production and midstream operations [53, 75, 82, 83]. In response, several organizations independent of regulators or the O&G industry have started large-scale aircraft – and in some cases satellite – campaigns to identify large emitters and publish them in open data portals, including the Environmental Defense Fund’s *Permian Map* project [128], the Jet Propulsion Laboratory’s *Methane Source Finder* [131], and CarbonMapper’s open data portal [116].

Data from these large-scale measurement efforts indicate that a substantial fraction of the detected large emitters behave stochastically: While a similar number of large emitters may be seen on multiple days of measurement, the location and size of the emitters varies substantially. Emissions with highly variable temporal and spatial behavior are poorly captured in current inventory methods, such as those utilized by national inventories (e.g. EPA Greenhouse Gas Inventory (GHGI)) and many modeling studies, including recent work modeling national-scale emissions for transmission and storage (T&S) [33], gathering and processing (G&P) [23, 71], and distribution systems [22]. Since these emitters are poorly captured in inventory methods, it is difficult to compare results of aircraft/satellite surveys (‘top down’) with inventories (‘bottom up’) that represent emissions as averages over extended periods and large regions – typically annual averages at basin-to-continental scales. These challenges are reflected in several comparisons of top-down measurements to bottom-up inventories [50, 31, 32].

The work presented here introduces a new approach to bottom-up inventory methods that increases both spatial and temporal resolution, building upon modeling done for basin scale studies in the Fayetteville Shale region [50] and the Barnett production basin [37] in the USA. This approach supports more accurate comparisons between measurement surveys and inventories by better incorporating normal variability in emissions as well as recent data on large emitters and abnormal process conditions. As an illustration of this method, the work focuses on one sector of the natural

gas supply chain – midstream – and compares to several commonly utilized facility-scale methods to the improved inventory method. The same tool and method could be made utilized for other natural gas sectors or for other anthropogenic source categories that exhibit high variability.

The midstream sector of the USA natural gas supply chain transports and upgrades an estimated 33×10^{12} standard cubic feet (scf), or $934 \times 10^9 \text{ m}^3$, of natural gas from US production basins to distribution companies and industrial consumers using over 400,000 miles (640,000 km) of natural gas pipelines[120] and associated compressor stations and processing facilities.

Compressor stations in all sectors of the natural gas industry receive gas from one or more pipelines originating at well pads or other compressor stations, compress the gas, and inject it into downstream pipelines. Stations often host equipment to support connected pipelines, such as pig launchers and receivers, and metering equipment. In gathering and storage applications, stations often include equipment that upgrades gas quality, including dehydrators and acid gas removal units (AGRU). Compressor equipment found in these sectors is also found at gas processing plants to inject processed gas into transmission pipelines.

The gathering compression stations, compressors at gas processing plants, and the inter-connecting pipelines between these stations represents the sizable G&P sector of the U.S. onshore natural gas industry. The T&S system, which transports processed gas to consumers, consists primarily of transmission stations interconnected by pipelines, and storage stations that buffer gas supplies for use in peak demand periods. These two sectors make up an estimated half[124] of U.S. methane emissions from the onshore natural gas system. Therefore, emissions from these facilities are of strong interest for both air quality and greenhouse gas reasons.

Emissions from these sectors have traditionally been estimated using ‘inventory’ or ‘bottom-up (BU)’ methods, including the EPA GHGI[124]. Inventories estimate emissions by multiplying a measure of activity (‘activity factor’) – such as component or equipment counts, or operating hours – by an estimated emission rate (‘emission factor’) for the activity. For example, connector leak emissions at a facility would be estimated by multiplying the average emission rate for all connectors on a gathering compressor station with the number of connectors on that station. Gen-

erally, both activity and emission factors represent long-term averages of both quantities, and are suitable for calculating total emissions over extended time periods (months to years), for a large area, such as a nation or state, although some recent work has looked at gridded inventories with finer spatial resolution.[34, 50, 37]

Recently, however, there has been increasing interest in BU estimates that are more representative of shorter time periods and higher spatial resolutions. Examples include comparing top-down (TD) aircraft estimates of basin emissions to BU inventories of those basins [36, 25, 21, 12]; comparing TD estimates to inventories based on BU measurements [42]; or comparing downwind facility-level measurements with BU inventories of those facilities [45, 38, 35, 41, 30, 27]. In each case, measurement using downwind or mass balance methods occur during short time periods (typically minutes) and are often constrained to afternoon working hours during the week, while the comparison inventory methods are long-term averages that include no temporal variability or spatial heterogeneity. These differences have stimulated concerns about the representativeness of inventory methods and spurred an interest in understanding the emission variation in bottom-up models [50].

This study describes BU inventory modeling using MEET, an inventory emissions model expressly designed to model time scales ranging from seconds to years, resolved to the spatial resolution of individual facilities or major equipment units. MEET models track any gas species which are available in emissions data, typically CO₂, methane, and other light alkanes, although other species are readily accommodated. A companion paper [74] describes the calculation of gas and fluid compositions that serve as inputs to MEET. Of particular interest is the comparison between short-duration measurements – here represented by full-facility measurement method – and inventory methods, including both variability in emissions and gas species ratios, which are often utilized by TD methods for source attribution (for example, references [25, 21, 42]). Additionally, the study also illustrates how continuous monitoring can be analyzed using temporally-resolved emissions.

To demonstrate these capabilities, we utilize natural gas compressor stations. Compressor stations represent one of the most concentrated and temporally variable sources of GHG emissions in the natural gas supply chain. The large compressors (0.5-50 MW) on these facilities cycle on/off with demand, and other equipment on the stations increases/decreases loading and associated emissions as compressor loads change. In addition, many compressor drivers (typically large engines or turbines) often burn some of the gas passing through the station for power. Driver exhaust gases often represent one of the largest sources of emissions on a station[33, 45, 71, 61]. Compressor equipment also releases short duration, high-rate, emissions as equipment is depressurized when idled (blown down) or started (gas starters). This variability has a substantial impact on emissions which may be measured by short-duration methods, leading to a interest in improving the modeling of compressor station emissions.

3.3 Methods

MEET implements a robust framework for emissions modeling built around a *DES*, a common multi-agent simulation methodology where individual ‘agents’ model the behavior of specific processes and interact with each other via a central event queue. Agent models (hereafter ‘models’) currently implemented in MEET include models for well pads and compressor stations, each of which combine models for major equipment, components, leaks, and operational changes on these facilities. The model framework is extensible to other sectors, equipment and facility types, as well as processes which interact with these facilities, such as leak detection and repair (LDAR) surveys, liquid transport, or equipment modifications. Events in MEET are resolved to the second, but can be queried at time resolutions from seconds to years. MEET resolves spatial dimensions to the major equipment level if those data are available.

An overview of the MEET simulation process is shown in Figure 3.1. The analyst defines activity data and either selects or populates emissions data. MEET runs multiple Monte Carlo (MC) iterations and saves results to event logs. After simulation, the analyst extracts time-series of emissions from the desired MC iterations at the desired time resolution. Using multiple calls to

the time series extraction tools, the analyst can drill down to finer time resolution in periods and facilities of interest. For both simulation and extraction, individual MC iterations and time-period queries are independent and can be executed in parallel on multi-core and cluster servers.

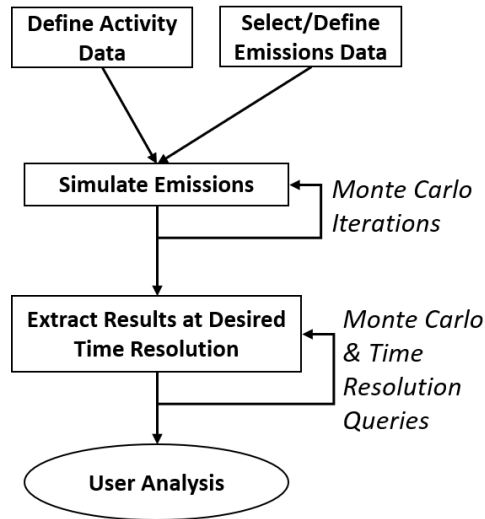


Figure 3.1: Overview of MEET process.

As with all inventory models, emissions and activity data completely define the behavior of the emissions model. However, the introduction of spatial resolution and time varying behavior introduces several unique aspects into MEET’s models, which are discussed below.

Emitter population: The compressor station model, like all facility models in MEET, is organized into a three-level hierarchy matching that used in several recent studies[71, 45, 33] [SI Section B.2.1]:

1. *Station or Facility:* A physical group of equipment, typically surrounded by a fence or some other outer marker.
2. *Major equipment unit:* A major, integrated, equipment unit at the station. Six types of major equipment at compressor stations are modeled for this study: Compressors, dehydrators, acid gas removal units (AGRUs), station separators, tanks, and flares. The limit of each unit of major equipment is defined by the flanges which connect it to the gas transport piping on

the station. All equipment not assigned to another major equipment unit is grouped into a single unit of major equipment termed ‘yard piping,’ as in Zimmerle et al.[62].

3. *Components*: Each major equipment unit includes many components. Nine types of components (threaded and flanged connectors, open-ended lines (OELs), valves, meters, pressure reducing valves (PRVs), regulators, gauges, and gas-driven pneumatic controllers) cover the general-purpose component categories found on all major equipment categories. Compressors include several additional types of compressor-specific components (blowdown and isolation valves, rod packing or shaft seals, pocket vents, gas starter valves, and combined vents). Additional specialized categories, such as station vents, are also modeled [SI Section B.2.1].

Activity data are defined by enumerating individual stations and the major equipment units on the station, while components are instantiated using population distributions for each component type [SI Section B.2.1].

Emission labels: In addition to chemical species, BU emissions simulation also provides insight into the origin and cause of emissions, and allows analysts to compare relative, interacting, emission rates between source types and causes. To support these analyses, MEET labels emissions with their source and cause. *Source* is identified by using user-supplied activity identifiers, which follows the hierarchy above. Currently, MEET supports the three labels for the *cause* of an emission:

1. *Fugitive*: Unintended emissions of uncombusted gas from components; commonly called ‘leaks,’ but also including excess venting and abnormal process conditions that release gas to atmosphere.
2. *Vented*: Emissions of uncombusted gas as part of regular site operations. Examples include venting from gas-powered pneumatic controllers and blowdowns of pressurized equipment.
3. *Combusted*: Emissions from onsite combustion. These emissions include combustion products (this study tracks only CO₂) and unburned fuel entrained in the combustion exhaust,

commonly called ‘combustion slip,’ from engines, flares, heaters, dehydrator still vents and other sources.

For labeling to be effective, activity and emissions data must support the classification of emissions into the appropriate categories. Emissions from a single location may change labels depending upon the operating state of equipment, the emissions behavior (size, duration, etc.), or via interaction with external agents (e.g. leak repair). Several special cases need to be considered.

First, in most venting categories, there is an intended emission rate and duration. Two possibilities exist: (a) Emissions below the intended rate *and* within the intended duration are classified as *vented* emissions; and (b) Emissions above the intended rate or longer than the intended duration are *fugitive* emissions. Unfortunately, few published data sources include sufficient information to separate vented emissions measurements between (a) and (b). For example, low- and intermittent-bleed gas pneumatic controllers are specified with an intended emission rate and, for intermittent bleed controllers, an expected duration for each actuation.[55] However, most data on pneumatic controller emissions provide only total emissions [19, 22, 33] which are not segmented into vented and fugitive components. Luck et al. [56] provide a prototype of how emissions could be segmented, but the data are limited in scope. Ideally, future field studies would collect and publish these data.

Second, properly operating combustion devices typically emit a fraction of the fuel gas uncombusted. For example, flares are specified to have a ‘combustion efficiency’ – typically 98% – which indicates that 2% of the combustible hydrocarbons sent to the flare are released uncombusted. Similarly, dehydrator still vents, engine exhaust, AGRU reboilers, etc. have similar specifications for properly functioning emissions. When malfunctioning, efficiency may drop substantially, and these emissions should be classified as fugitive emissions. As with pneumatic controllers, these data are not generally available.

Due to these data limitations, available input data forces emissions to be coarsely labeled in the current model. SI Table B.2 summarizes labeling for each category of modeled emissions,

SI Section B.2.6 describes episodic venting (gas starters and blowdowns), and SI Section B.2.7 describes non-episodic vented sources.

Time domain behavior: Since emissions models in MEET are DES agents, changes in one model may impact the behavior of other models. For example, a leak detection agent could identify a fugitive emission, trigger a leak repair agent, which in turn mitigates the fugitive emission.

For compressor stations, the operating mode of the on-site compressors represents the single largest driver of temporal variations in emissions, due to the large exhaust emissions of the compressor drivers[45, 71], and the high-rate, short-duration, gas releases when compressors are started or blown down. For emissions purposes, each compressor is at all times in, or transitioning into, one of three principle operating modes: Operating (OP); not operating pressurized (NOP); and not operating depressurized (NOD). State transitions are shown in Figure 3.2. To depressurize the unit, the compressor passes (briefly) through the unit blowdown state, when the blowdown valve is opened and pressurized gas in the compressor is vented to atmosphere. When entering OP mode from either NOP or NOD, the compressor must pass through a short-duration ‘start’ state to start the compressor driver; if a gas starter is present, gas will be vented during this state. In some cases, the start fails (dotted lines), resulting in a return to the previous state. (Failed starts were not implemented for this study due to the lack of data for the frequency at which they occur).

Major compressor valves (isolation, blowdown, starter) change state with the compressor operating mode [SI Section B.2.1, B.2.6]. Emissions through the valve seal when the valve is closed are fugitive emissions. When open, there are no emissions or emissions are intentionally vented emissions. For example, during unit blowdowns, the blowdown valve is opened and emissions are vented (i.e. intended). When the compressor is pressurized (OP and NOP modes), the blowdown valve is closed, and any leaks through the valve seal are fugitive (i.e. unintended) emissions. When the compressor is depressurized (NOD mode), the blowdown valve is typically left open, and causes no emissions; any emissions observed at the blowdown vent are due to leaks in isolation valve seals [SI Table B.1, Figure B.3]. A similar analysis was performed on other emitter categories, as shown in Table 3.1 [More detail in SI Table B.2].

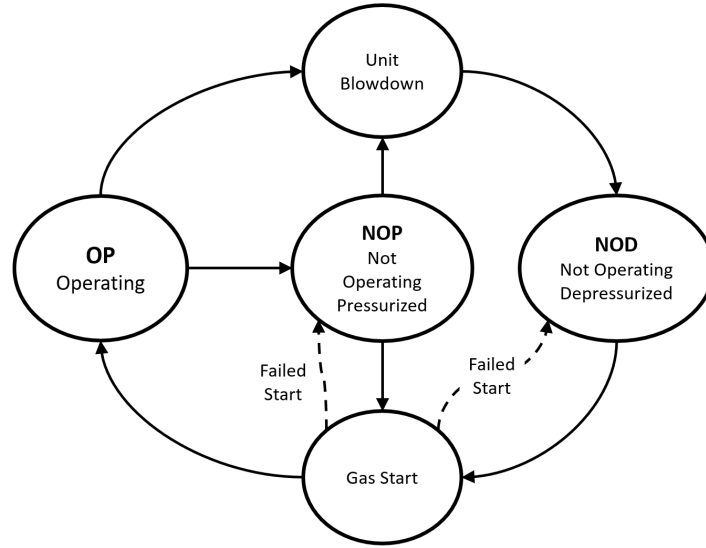


Figure 3.2: Compressor unit operating states. The diagram illustrates the principal operating modes of the compressor and its associated driver. Valid state transitions are shown with arrows. Current model does not implement failed starts.

Intermittent gas pneumatics also cycle between two operating states – a near-zero emission waiting period and a short, high-emitting, actuation. Actuation rates vary between near-zero to frequent actuations[56]. Typically, actuation emission rates peak at 75 slpm (150 scfh) whole gas, and may interact with other processes, such leak detection.

Table 3.1: Emission labels for emitter categories on compressors that change with compressor mode

Emitter Category	OP	NOP	NOD	START	BLOW-DOWN
Components Leaks	F	F		F	F
Gas pneumatics*	V	V		V	V
Blowdown valve	F	F		F	V
Isolation valve		F	F		F
Starter valve	F	F	F	V	F
Rod packing#	V	F	F		
Wet/dry seals	V				
Combined unit vents	F	F		F	
Compressor driver exhaust	C				

Key: *F*: fugitive, *V*: vented, *C*: exhaust, and blank indicates the emitter is inactive, but not mitigated.

* Emissions from intermittent pneumatic controllers may be either fugitive or vented emissions.

Rod packing emissions are drawn from different distributions for each compressor operating mode.

Inter-dependencies between emitters: Gas throughput for a station is determined by the number of compressors operating and the current throughput of each unit. The operation of other equipment on the station, such as dehydrators, flares and tanks, varies with the station throughput. Variations can include both gas throughput for individual units, or the number of units online at any one time. The frequency of dump valve actuations on station separators is also dependent upon throughput and the quantity of liquid in the incoming gas. Unfortunately, little data exists to link the operation of non-compressor equipment to the current throughput of the station. For example, vent emissions from dehydrators depend upon throughput, but also re-circulation rate and other settings that vary somewhat independently of throughput.

Due to lack of public data, MEET's current example inputs do not coordinate compressor operations with other equipment. However, when analysts select input data, they can coordinate the throughput-related loadings between equipment types. Assuming the availability of the required data, future extensions of the model could coordinate throughput for all facility equipment.

Leak models: A primary focus of BU models is an accurate treatment of fugitive (leaks and other unexpected) emissions behavior. MEET uses one leak model for all facility types; it requires three sets of information:

1. Distributions of component/source counts for each component/source category, for each major equipment type. Component count distributions are typically derived from field studies that counted components, or from internal company data.
2. The fraction of each component category that is leaking.
3. Distribution of emission rates for *leaking components*, commonly called a 'leaker factor' or 'leaker distribution.' Note that leaker data *is not* the same as *average* or *population* emission factors.[62]

Since many distributions are highly skewed, empirical distributions from field studies are preferred for both activity and emission data.

Most field studies of leaks are ‘snapshots’ in time that identified and measured individual leaks. In contrast, MEET needs not only the leak size, but also a measure of the frequency at which new leaks appear. To estimate these parameters, MEET makes a number of assumptions to emulate a time-varying leak count with variable leak size. Key steps are discussed below; SI Section B.2.5 provides mathematical detail on the formulation of the model.

First, the model assigns component counts for each component type on each unit of major equipment (e.g. a compressor). Counts are pre-set for some source types [SI Section B.2.1] or drawn from an empirical distribution of counts for others [SI Section B.2.4].

Second, at the start of the simulation, MEET assumes that the fraction of components leaking in each component category equals the fraction observed in field data. MC methods are used to populate these initial leaks. Leak count may vary substantially between MC iterations, but will converge over a sufficiently large number of MC iterations to the fraction of components seen leaking in the field study. This method assures that the facilities modeled in MEET reflect the conditions seen in the field study at the start of the simulation.

Finally, later in the simulation, additional leaks will start. Leaks may also be repaired, either through routine facility maintenance or via specific LDAR processes. To generate additional leaks, an assumption must be made on the frequency at which leaks appear on the simulated equipment. Most emissions models are populated with data from field campaigns that screened for leaks at many field locations (operator records from LDAR programs may also be used). In most cases, the facilities screened for leaks in these studies have been subject to prior LDAR activities. The frequency of LDAR surveys, and the fraction of detected leaks repaired, impact how these data reflect the *generation rate* of new leaks. Unfortunately, these qualifying data are seldom published. Therefore, the model makes several assumptions to produce a realistic leak generation rate.

First, the model assumes that (a) facilities in the field study were subject to periodic LDAR surveys; (b) all leaks were detected; and (c) all detected leaks were fixed shortly after detection, prior to the next survey. Therefore, leaks detected during any survey – including the study’s field campaign – reflect the number of new leaks since the last survey. Since field campaigns typically

measure facilities at arbitrary times, on average they visit facilities half way between LDAR surveys. Leaks detected therefore represent, on average, half the leaks generated at the facility per LDAR survey.

Unfortunately, leak survey frequency for studied facilities is not typically published with field studies. Therefore, MEET assumes that LDAR surveys occur annually, a common frequency, and therefore the field data represent new leaks which appeared in six months. Different survey frequencies can be accommodated, if known, by adjusting the leak generation rate parameters.

Second, if leaks are generated but not repaired, the number of leaks will increase throughout the simulation. This may be the desired behavior, for example, to test the efficacy of several LDAR programs. However, in the default case, where leak repair actions are not independently simulated, MEET simulates repair actions such that leaks are repaired at the same rate as they are generated. Since repair actions are generated independently from leak generation, the number of leaks will vary, but will converge to the desired leak frequency over a sufficiently large MC simulation.

The availability of leak data varies substantially between compressor stations in different industry sectors, and current sources require specific pre-processing to be included in a MEET model [SI Sections B.2.5 to B.2.5].

Gas composition: MEET simulates CO₂, CH₄, and light alkane emissions. During the DES simulation, the event log notes the type of gas emitted (e.g. whole gas, combustion exhaust, etc.) and gas compositions are applied when data is extracted to a time series.

For compressor stations, gas composition data varies between incoming gas, gas after dehydration, and flash gas from liquid storage tanks. However, these data are not generally available for compressor stations. Therefore, the current MEET model utilizes a single gas composition for all *whole gas* emissions from the facility (vented or fugitive), while combusted emissions are computed using models which are specific to the type of equipment [SI Section B.2.3].

Sources not modeled: Several sources are not included in current MEET models for compressor stations, due to the lack of public data to design these models. Combustion emissions from heaters (e.g. AGRU steam heaters, dehydrator reboilers) are not modeled. MEET also assumes all

AGRU vents are routed to flares. Full or partial station blowdowns are not modeled; these events are infrequent. While only marginally interesting for method comparisons, they must be included to estimate annual emissions. Unlike the MEET production facility model, tank flash emissions are not modeled since little information exists on the composition of liquids sent to the tanks. Finally, emissions from dehydrators and flares are drawn randomly from Greenhouse Gas Reporting Program (GHGRP) reports, and, as a result, there may be mismatch between the size of these emitters and other equipment on the compressor station in some MC iterations.

For ease of use, MEET defaults any data not specified by the analyst. For example, if the number of compressors are specified, but the size and type of compressor driver is not specified, these values are drawn randomly from published data, and will vary substantially between MC iterations. Alternatively, if these values are specified, then major compressor emissions, such as compressor driver exhaust, will show only minor variation between iterations.

3.4 Results and Discussion

To demonstrate the model, a set of eight compressor stations was constructed. Seven of the stations are modeled after facilities measured in a recent basin study [45], including the size and number of compressors and the type of compressor driver. After configuring the facility equipment, the run time of the compressors was manipulated to create a variety of operating characteristics across a range of compressor loading (fractional load when running) and utilization (fraction of the time the compressor is running). The eighth station (station ID 1), utilizes compressor types, loads, and utilization taken from T&S data [33]. Key characteristics of the stations are summarized in Table 3.2, and SI Table B.6.

The stations were simulated for a year, using a relatively small number of Monte Carlo iterations (20), since the primary purpose is to illustrate the functionality of the model. Across all iterations, 3131 emitters were active at some time. Large emitters were included by modifying default emissions distributions [SI B.2.9]; the 25 emitters larger than 10 kg/h are summarized in SI

Table 3.2: Key operating characteristics of test stations

Station ID	Sector	Size	Compressor Count	Total Station Power (10 ³ HP)	Total Station Power (MW)	Compressor Loading When Operating	Utilization	Operating Fraction
1	T	-	2	9.78	7.29	90%	Very Low	35%
2	G	Typical	8	12.5	9.29	80%	Typical	79%
3	G	Typical	5	8.88	6.62	98%	Typical	85%
4	G	Small	2	3.55	2.65	104%	High	90%
5	G	Small	2	3.55	2.65	50%	High	90%
6	G	Large	24	40.9	30.5	90%	Typical	79%
7	G	Small	2	6.23	4.65	0%	Zero	0%
8	G	Typical	7	12.5	9.29	90%	Low	58%
All	G&T	-	52	97.8	72.9	86%	Typical	73%

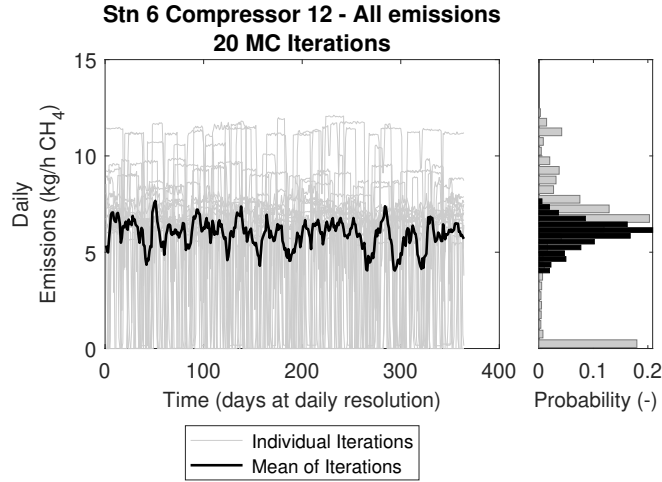
Table B.9. These emitters appeared in 14 Monte Carlo iterations, and on 6 of the eight stations, in at least one iteration, and 6 sources emitted over 100 kg/h.

Daily total emissions were extracted for all emitters, and a set of 26 randomly-selected half-day periods were extracted at one-minute resolution to simulate measurement methods. Figure 3.3 provides an example of daily and minute resolution time series.

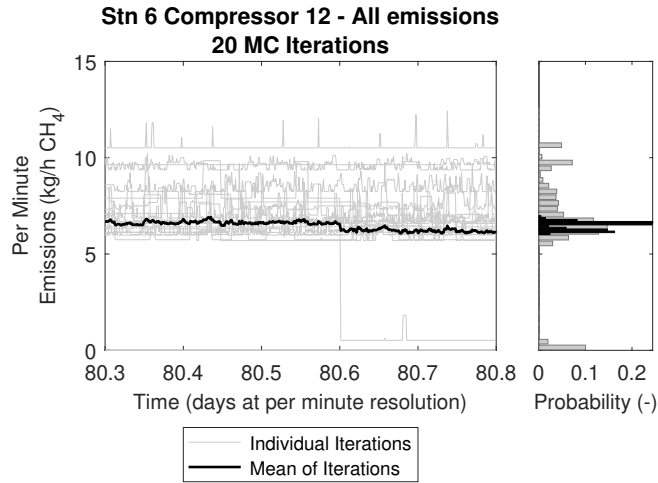
To demonstrate the model, we consider three test cases which model aspects of methods in active use: error bounds when estimating emissions using a downwind tracer flux method, variability in ethane/methane ratio, and the probability that a continuous monitor can identify a fugitive emission given a substantial background of vented and combusted emissions.

Tracer Flux Simulation: Single- and dual-tracer flux methods are well-established methods for estimating facility-scale emissions [45, 33, 38, 30, 46], following methods developed in the 1990s[2] and extended recently with the use of more advanced trace gas analyzers[27]. Briefly, the method utilizes one or more tracer gasses – acetylene and nitrous oxide are common – released at a known rate near a suspected emission source. A mobile instrument transects the emission plume downwind of the release, covering sufficient distance to measure both background and plume concentrations of both tracer and target species. Emission rates are calculated from the ratio of the target species to the known tracer emission rate [SI Section B.2.9].

In this study, tracer measurements were simulated using the one-minute resolution data. Each transect (or ‘plume’) was calculated as the mean of 3 minutes of data with a measurement error



(a) Daily resolution for the year. Day-to-day variability is primarily due to operating times of the compressor and, while load is constant during any one run period, it varies randomly between operating periods.



(b) Minute resolution, midday on day 80. Fine detail illustrates short duration events occurring during the simulation, such as pneumatic actuations and the cycling of emitters with the compressor operating state.

Figure 3.3: Simulation examples for compressor 12 on station 6. Left panel displays the time series. Light gray lines indicate the 20 individual Monte Carlo iterations and the dark line indicates the mean over all iterations. Right panel displays the same data as a histogram of values.

standard deviation of 15%. The plume duration is a conservative assumption (which improves the tracer estimates) by averaging over a longer time than a typical transect. Measurement error is typical of estimated error from prior studies. [30, 27, 24]. A single tracer measurement was

the mean of multiple (2 to 8) plumes, which were spaced 5 to 15 minutes apart. The resulting simulations were ‘best case’ tracer method estimates, with no confounding sources, and neglecting any issues due to spatial misalignment of emission sources relative to tracer release location, poor road access, or challenging wind conditions. Results are summarized in Figure 3.4 and SI Table B.10.

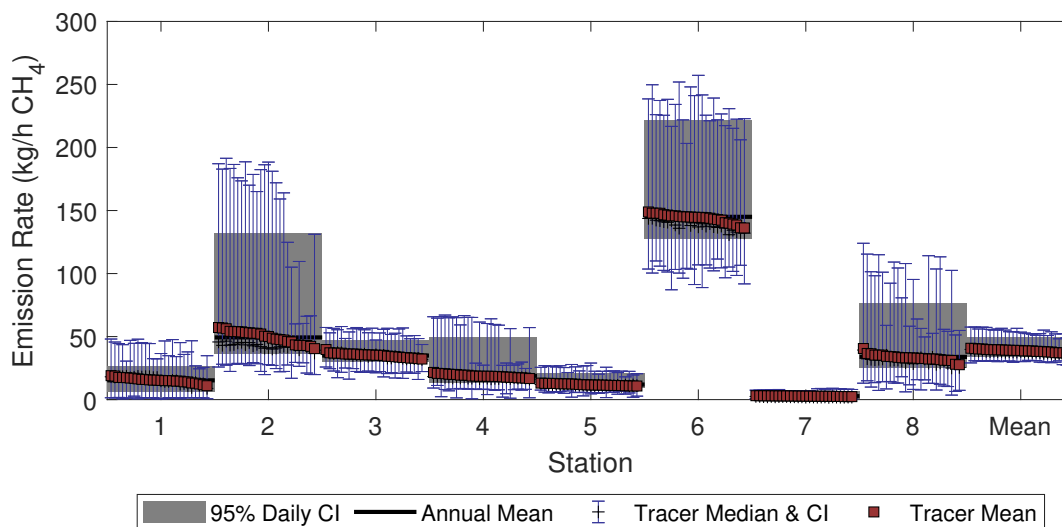


Figure 3.4: Results of tracer simulation. Data are grouped by station, with tracer estimates in declining order by median estimate. Results from all stations combined are shown as a per-station estimate labeled ‘Mean’. Gray blocks represent the confidence interval (CI) for station emissions based upon 365 days of simulated emissions. Tracer estimates for each day are shown as median and mean points and bars with 95% confidence intervals. Confidence interval bars extending outside the gray region represent the probability that a field measurement campaign measuring each of these facilities on one day would estimate emissions outside the confidence interval of annual emissions for these facilities.

The tracer simulation demonstrates that the method correctly estimates emissions for all station types and sizes over a large number of days and simulated measurement events. Errors in mean estimated emissions range from -2% to 0.9% for individual stations, and -0.5% for measured emissions on the group of 8 stations. This illustrates that the method itself produces correct estimates in the mean.

However, no measurement campaign includes a large number of days and measurements; campaigns are economically limited to a few samples of each station [45, 27, 24], typically occurring on one or two days. Therefore, the accuracy of any measurement campaign is more correctly reflected by the accuracy of individual measurements, which is far lower than large ensembles of measurements. For individual stations 30% of measurements are outside the confidence interval of the annual estimate, with more estimates (24%) underestimating emissions than overestimating (5.3%). For the station group, 28% of measurements are outside the annual confidence interval (21% under the lower CI and 7% over the upper CI). Therefore, due only to natural variability in emissions from these midstream sites, a measurement campaign has a one in three chance of estimating emissions outside the 95% confidence interval of annual emissions.

Given that the simulated measurements do not reflect all of the challenges of actual field measurements, these simulations illustrate that the uncertainty in tracer measurements – arguably one of the most accurate of field measurement techniques – is higher than commonly estimated. In particular, error estimates typically include only the measurement process, and do not account for the unknown variability in the facility’s operating state and emissions behavior.

Gas Ratio Impact: Regional measurements often utilize ethane/methane (C2/C1) ratio to attribute greenhouse gas emissions to a mixture of thermogenic and biogenic sources in a region. For production basins and urban areas, a common assumption is that thermogenic natural gas emissions will have a C2/C1 ratio that corresponds to the gas production in the basin, or delivered gas in a distribution system area. The time-resolved simulations of the MEET model allow an estimate of both long-term mean and variability in C2/C1 ratio.

All compressors on the stations simulated here are equipped with lean-burn natural gas engines as drivers. Lean-burn engines are common in gathering and transmission, accounting for 66% of all compressor drivers in a recent gathering study[71, Table S3-19]. When operating, a fraction of the natural gas fuel is emitted, uncombusted, in the engine’s exhaust, an emission source commonly called ‘combustion slip.’ Further, C2+ hydrocarbons are preferentially oxidized in the engine (and

potentially in the exhaust catalyst), leading to excess methane emissions relative to the fuel gas composition.

The C2/C1 ratio based upon daily simulations is shown in Figure 3.5. Station 7, which has no operating compressors, exhibited a C2/C1 ratio matching that of the natural gas transported through the station. In contrast, Stations 3 and 6, with 5 and 24 compressors, respectively, and high compressor utilization and loads, exhibited C2/C1 ratios approximately half that of the gas passing through the station. It is common for gathering stations to operate at relatively high utilization (>80%) and high loads. When operating in this common mode, the stations may produce emissions which have one half the C2/C1 ratio of the field's produced gas. For basin-level estimates, underestimating C2/C1 ratio for any source type is likely to lead to under-attributing emissions to that source, and/or increasing attribution to a source with higher relative ethane emissions. A recent basin-level study that modeled emissions variability indicated that gathering station emissions were similar in size to production facility emissions[50]. In this case, underestimating C2/C1 ratio of gathering by $\approx 50\%$ could underestimate that ratio by 25% for all oil and gas operations, and potentially skew source attribution away from gathering toward production or other gas infrastructure.

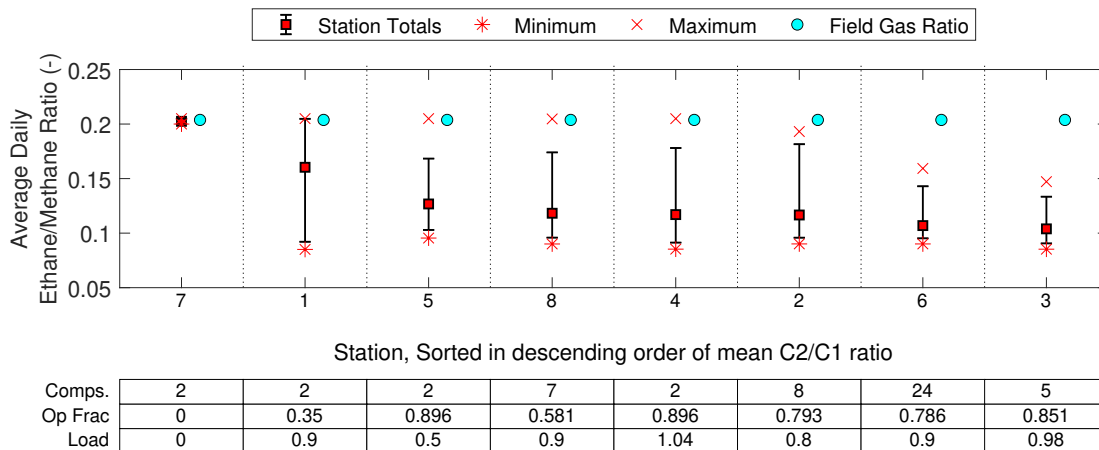


Figure 3.5: The ethane-methane ratio of emissions from the simulated compressor stations. In this simulation, all stations handled natural gas with the same C2/C1 ratio (0.204); variation in the C2/C1 ratio is entirely due to changing activity (vented and combusted emissions) and leak emissions at the station.

A similar analysis conducted on facilities with high liquid production could see an opposite effect; gas dissolved in the liquids and emitted from tank vents could be relatively richer in ethane than the produced gas composition.

Additionally, day-to-day variability in C2/C1 ratio increases with variability in compressor utilization due to higher variability in the number of compressors operating at any one time. Transmission stations generally have lower utilization than gathering stations to provide capacity for surge demands on the transmission system. Correspondingly, Station 1, a transmission station with 2 large engines, displayed the highest day-to-day variability emission. In the current simulation, compressor utilization for transmission is entirely random. A more representative simulation, however, would reflect seasonal variations in transmission demand, with corresponding seasonality in both emissions and C2/C1 ratio.

These results indicate the need for caution when applying current methods of source-attribution in basin-scale studies. Since bottom-up inventories that are commonly utilized as Bayesian priors for source-attribution do not model differences in C2/C1 ratios, or temporal variation in those ratios, attribution may be skewed between O&G sectors, or between O&G and other methane emission sources.

Continuous Monitoring: ‘Continuous monitoring’ refers to the use of permanently installed sensors, coupled with appropriate analytics, to sense emissions from a facility on a near-continuous basis. This type of sensing has drawn increasing attention due to the development of lower-cost sensors with higher sensitivity and improved analytic methods for identifying emissions. Recently, Colorado (a centrally-located U.S. state) passed legislation that will require continuous monitoring at many oil and gas facilities.[54], and proposed regulatory changes by the EPA [89] also encourage continuous monitoring.

The prior simulation examples, above, were primarily concerned with total emissions from a compressor station. In contrast, a successful continuous monitor must identify unwanted emissions in an environment where the other sources may produced planned emissions. To analyze this

scenario, MEET divides emissions into the categories described in the Methods section (‘Emission labels’) and utilized by the EPA and other inventories [125, 124].

The goal of a continuous monitor installation is to identify *fugitive* emissions in an environment where *vented* and *combusted* emissions are substantial. Therefore, in this example we simulate the mix of emissions at the compressor stations to develop an understanding of what sensitivity and analytic performance would be required for a continuous monitor to detect fugitive emissions.

The simulation assumes that knowledge of the facility’s background (*vented* and *combusted*) emissions is known exactly – i.e. combustion slip from compressors, timing, size and duration of blowdowns and gas starts, and gas releases from pneumatic controllers is known precisely and can be ‘tagged’ as vented by the monitoring solution’s analytics. In practice, this type of knowledge is near-impossible to acquire. Therefore, the estimates here provide a rough upper bound on the performance of a continuous monitoring solution [SI Section B.2.9].

To consolidate the complex topic of monitoring performance into a single metric, simulations measure the monitoring solution’s performance using a ‘detection threshold’ defined as “ability to identify fugitive emissions that are X% higher than the *perfectly known* background emissions.” For example, a detection threshold of 20% would indicate that emissions that are $\geq 120\%$ of the known background would trigger a fugitive emissions detection at the station; if fugitives are $< 120\%$ of known background, no detection would be triggered. SI Figure B.5 provides a time series example for Station 6 – a large station with 24 compressors and high utilization – and SI Figure B.6 for Station 7 – a small station with no compressors running. The simulation utilized methane as the sensed gas species, and both stations have fugitive emissions at all times.

For station 6, a monitor with a 20% detection threshold would detect fugitive emissions on 10% of the one-minute samples in the simulation; the low percentage is primarily due to the high background emissions, primarily from combustion exhaust. In contrast, the same solution on station 7, with no combustion exhaust and lower overall activity levels, would detect fugitives 99% of the time. Extending the analysis, Figure 3.6 illustrates results for detection thresholds from 0-100% of total emissions.

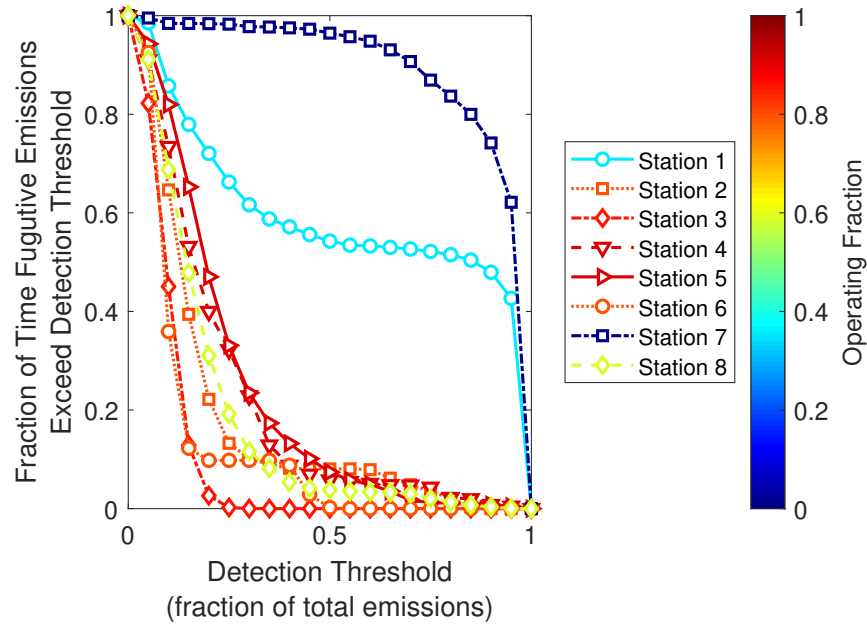


Figure 3.6: Fraction of the time that a continuous monitor (CM) would detect fugitive emissions. The horizontal axis shows the detection thresholds for the CM solution, ranging from 0% (a perfect system that can distinguish *any* fugitives within any amount of non-fugitive emissions) to 100% (all emissions must be fugitive emissions for a detection to take place). A detection threshold of 50% indicates fugitive emissions equal non-fugitive emissions. The vertical axis shows the the fraction of the time the continuous monitor would successfully indicate that fugitive emissions were occurring. The color scale indicates the time-weighted utilization of compressors on the station, from 0 – no compressors operating – to 1 – all compressor operating.

For stations with higher compressor utilization (i.e. all stations except 1 and 7), the fraction of time a continuous monitor would detect fugitive emissions drops dramatically for detection thresholds less than 40%, even for stations with a small number of compressors (stations 4 and 5). Given the uncertainty in operating state of equipment and timing of emission events, knowing background emission rates to within 30% is challenging; actual performance would likely require more sophisticated sensing, better analytics, or multiple gas species to achieve detection thresholds below 40%. Stations 1 and 7 have extended periods with no compressors operating. This increases the time when background emissions are lower, increasing the time (i.e. the probability) when fugitive emissions would be detected.

These results provide a cautionary note regarding the deployment of continuous monitoring solutions. First, applicability of these methods is highly dependent on site type and site operations.

Second, the analytics behind continuous monitors will likely need ongoing, near-real-time, information about site operations to estimate combusted and vented emissions in order to distinguish fugitive from these other sources. Finally, even for very large facilities (i.e. station 6), continuous monitoring successfully detects the largest emitters when they occur, as illustrated by the detection pattern in SI Figure B.7. For this station, all fugitive emitters with rates above 100 kg/h would be detected by a continuous monitor with a 20% detection threshold.

3.5 Conclusions

The seemingly simple extension of BU inventory methods to accommodate time-varying emissions highlights underestimated uncertainty in measurement methods and emissions monitoring. Results of this study indicate systematic underestimation of uncertainty in common facility-scale measurement methods and basin-scale source attribution methods: Studies typically do not account for variability in emission rates and C2/C1 ratio at midstream facilities, and by extension, other O&G facilities. These increased uncertainties are additive to other method uncertainties which are commonly estimated, such as meteorological uncertainty, instrumentation limits, and modeling limitations. Similarly, interest in continuous monitoring at O&G facilities may obscure limitations on their performance: Most are tested in controlled conditions with little-to-no vented or combusted emissions, then deployed at facilities with high levels of vented and combusted emissions, where these methods may be unable to distinguish fugitive emissions from other planned gas releases.

Use of MEET also highlights significant shortfalls in data available from field studies ... including those performed by the authors. Significant data gaps which could be addressed in future field campaigns include: monitoring of equipment to capture time-varying behavior, such as compressor starts and stops; segmentation of emissions into fugitive and vented components; capturing LDAR survey frequency and the time since the last survey during the field campaign; capturing data to relate equipment operation to overall station throughput; and noting the volume and frequency of

unit blowdowns. Each of these data items are of marginal importance to yearly inventory results, but have substantial impact on short-term simulations, as noted in the examples.

Data released in conjunction with component emissions measurements also presents challenges for time series simulation of emissions. Typically, field campaigns note the emission rates of measurements for a component type, and potentially the number of components of that type. However, it is unusual to note the number of components *screened* for leaks – as opposed to total number of components – and the number of detected emissions, including those that could not be measured. These two elements – component count screened and emission count detected – are a necessary ingredient to estimate time series behavior of many emitters [Section B.2.5]. As an illustration, a study finalized during development of MEET [71] provides a prototype of what data should be captured and how it can be presented (see `AvgFactor.xlsx` in study data).

The MEET software will be released as an open source tool in 2021, hosted by the University of Texas (<http://dept.ceer.utexas.edu/ceer/ect/>).

Chapter 4

Real-Time Simulation and Modeling

This case study describes an experiment conducted between National Renewable Energy Laboratory (NREL) and CSU, where a real-time simulated model was split between the two sites.

4.0.1 Modeling

Real-time simulation, where the simulation step time is the same as real, observed time, are highly depending on precise timing. These simulations are typically run on specialized hardware designed specifically for real-time simulations [137, 142]. Also, these simulations often depend on specialized and expensive equipment (wind turbines, grid simulators, etc.), and developing large-scale experiments at different locations require a network [48]. However, introducing a network adds additional challenges in the form of random, unpredictable network delays, and the problem of how to keep distributed simulations synchronized.

One way to address these problems is to build observers into the model itself to maintain and correct the model state across the distributed elements. This technique is discussed in Appendix C.2. Another technique to handle networking related problems is to build a protocol that keeps track of timing related parameters and presents to the observers for model correction. The rest of this chapter describes such a protocol.

4.0.2 Related Works

- Appendix C.1 describes how to achieve precision real-time performance on modern, inexpensive hardware.
- Section Appendix C.2 describes techniques how to correct for network delays in distributed, real-time simulations.

4.1 Networking for Real-Time Distributed Simulation

The full-scale implementation was split between NREL and CSU. Firewall exceptions were required from both institutions to enable data transfer. Traffic originated from the NREL node (the “primary”), was routed through the NREL firewall, and sent to the CSU node (the “secondary”). The NREL side used Network Address Translation (NAT) to allow bi-directional packet flow to the CSU node. The CSU node required a firewall exception, which allowed traffic from the specific IP address of NREL’s NAT firewall to the secondary node.

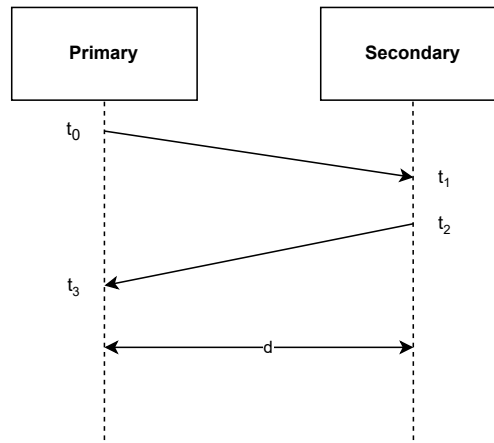


Figure 4.1: Network Timing Calculations

The challenge for this model is that the network delay between the two nodes. This network delay is dependent on many factors, including network load and routing algorithms. It is impossible to determine before a packet is sent. It can be determined after the fact, though conditions can change from one packet to another. To calculate this delay for an individual packet, the primary and secondary include the time of sending and receipt of the packet in the packet itself. These times are identified as t_0, t_1, t_2, t_3 in Figure 4.1. The implementation then uses the Network Time Protocol [1] algorithm to calculate the round-trip delay (d) as shown in (4.1)

$$d = \frac{(t_1 - t_0) + (t_3 - t_2)}{2} \quad (4.1)$$

Note: Mills also provides a formula to calculate the time offset between two nodes – in this case, the difference between t_0 and t_1 . This calculation is not used in the initial implementation, but can be used in subsequent implementations to improve calculation of the delayed terms in the observers.

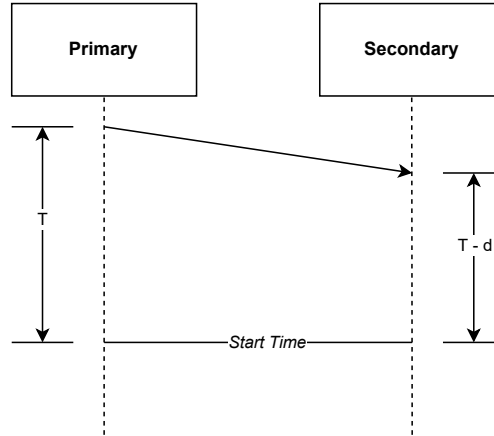


Figure 4.2: Synchronizing the Reference Signal

Changes to the Reference Signal need to be sent to both Observers simultaneously. This is achieved by synchronizing the start of processing in both the primary and the secondary. In this case, the primary determines when to start processing and calculates a time delta (T in Figure 4.2). The primary then adjusts this time delta with the average network delay and sends this calculated modified value to the secondary and then waits for the time interval T . The secondary receives the packet, and waits for the time interval $T - d$, which assures the primary and secondary node start at the same time.

Basic operation of the system is shown in Figure 4.3. On a periodic basis, the primary sends its updated output value to the secondary. Also included in this packet is the next reference signal value – this allows the system to be completely controlled by the primary and not require sharing of the complete reference signal values in the secondary. The secondary then replies to this packet with its updated output value. Also shown is the timing relationship of the output values to the delayed values stored in both Observer A and Observer B. The current implementation simply uses

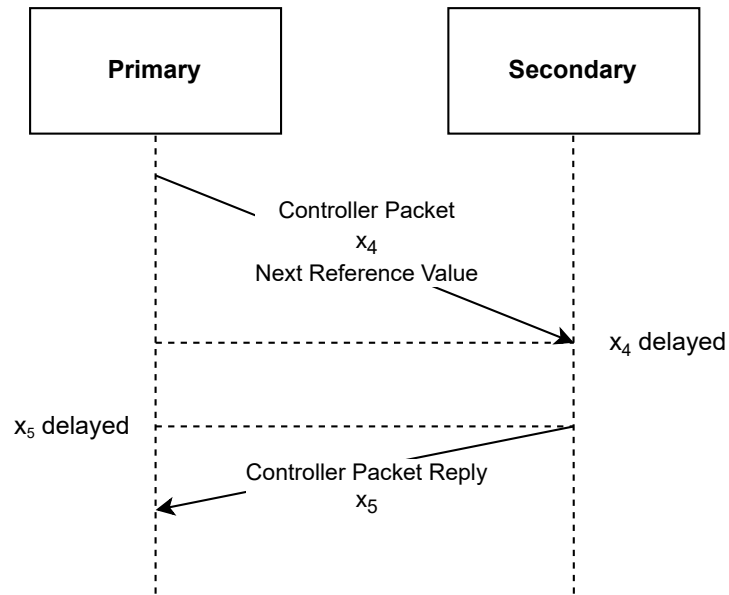


Figure 4.3: Ongoing Timing Adjustment

the stored value from the previous time step for these values. Future implementations can store more values, along with the local timestamps when they were calculated, and combine this with the timestamp values in the packet and the offset calculation to refine and improve the accuracy of the delayed signals.

Chapter 5

Conclusion

Simulation's impact on science, and thus policy, is increasing. This means that attention must be paid to the process by which simulations are developed to ensure high-quality, defensible results.

Two main reasons for developing simulations are when the processes being simulated are expensive or dangerous for experimentation, and when the timing events being examined are either too fast or too slow to be observed regularly.

There has been a proliferation of data that can be used to drive models in recent years. One reason for this proliferation is a dramatic increase in sensor / IoT / I-IoT data, where environmental conditions can be inexpensively and reliably monitored. Another reason for this proliferation is the wide availability of public data sources – including regulatory agencies making reporting data publicly available.

Unfortunately, this proliferation of data has a downside – the quality of it is unknown. Anyone using this data must take responsibility for ensuring that the data values are in range and make sense, missing values are identified, and there are no systemic biases. A second issue with this data proliferation is that data sets are developed in isolation of each other, and care must be taken to merge them together. Keys must be checked for uniqueness and completeness, and merged data sets must be checked to ensure they are consistent.

The increasing importance of simulation, and the issues with proliferation of data, suggest the need for a robust process for defining high-quality simulations. Such a process starts with solid *Data Management* principles, including verifying individual data sets, merging data sets from different sources to provide a comprehensive data set, and an iterative process to identify data quality or merging issues.

The second step of a robust simulation process is the *Modeling* phase, where the simulation itself is actually developed. This phase typically starts with small-scale simulation runs that verify the isolated steps of the process, then expand to larger and larger verifications. Two key mecha-

nisms for validating simulation results are to set the inputs to the simulation to the initial conditions of an observed process, and compare the output of the simulation to the observed results. The other mechanism is to compare the results of the model being developed with a known good model in the same domain.

The third step is the *Analysis* phase, where conclusions are extracted from simulation results. This is in fact the most intellectually challenging step, as pulling out good conclusions from huge amounts of simulation results can be difficult. Another important process during this step is to anonymize the results, which may, in fact, require consideration during the Data Management and Modeling phases.

Automation is critical for the reproducibility of results and to handle data updates – perhaps for re-running the simulations for a new year. During the Data Management phase, a machine-readable "metadata" file should be created that describes where the data came from, when it was created, what QA checks were applied to it, etc. Whenever a data set is updated, the metadata should be updated as well. During the actual modeling, automation can be used for continuous results validation, and an additional metadata block should be created that (a) references the metadata for all input data, and (b) describes specifics of the simulation process itself – who ran it, when it was run, input parameters, software versions, etc.

Finally, during the Analysis phase, automation can help by generating reports from templates and anonymizing the results. And a metadata block should be created to describe specifics of the analysis process. All the metadata blocks can be used for debugging and to provide a chain of custody and transformation of the simulation results.

5.0.1 Future Work

Processes for developing data sets that can be more easily merged should be defined. These processes can include basic data QC practices, management of keys for merging, and electronic access / data search processes.

Since automation plays such an important role for the repeatability of the automation process, software components to support the process can be developed. Example modules that could be useful include metadata management, template-based report generation, generic timeseries manipulation, and tools for verifying simulation output data vs. other data for verification.

Bibliography

- [1] D. L. Mills. “Internet Time Synchronization: The Network Time Protocol”. In: *IEEE Transactions on Communications* 39.10 (Oct. 1991), pp. 1482–1493. ISSN: 0090-6778. DOI: 10.1109/26.103043.
- [2] Brian K. Lamb et al. “Development of Atmospheric Tracer Methods To Measure Methane Emissions from Natural Gas Facilities and Urban Areas”. In: *Environmental Science & Technology* 29.6 (June 1995). tex.ids: lambDevelopmentAtmosphericTracer1995a, pp. 1468–1479. ISSN: 0013-936X. DOI: 10.1021/es00006a007. URL: <http://dx.doi.org/10.1021/es00006a007> (visited on 08/19/2016).
- [3] L. Moresi, M. Gurnis, and Shijie Zhong. “Plate Tectonics and Convection in the Earth’s Mantle: Toward a Numerical Simulation”. In: *Computing in Science & Engineering* 2.3 (May 2000), pp. 22–33. ISSN: 1558-366X. DOI: 10.1109/5992.841793. URL: <https://ieeexplore.ieee.org/abstract/document/841793> (visited on 01/15/2024).
- [4] Dale K Pace. “Modeling and Simulation Verification and Validation Challenges”. In: *JOHNS HOPKINS APL TECHNICAL DIGEST* 25.2 (2004).
- [5] A. J. Koning, S. Hilaire, and M. C. Duijvestijn. “TALYS: Comprehensive Nuclear Reaction Modeling”. In: *AIP Conference Proceedings* 769.1 (May 2005), pp. 1154–1159. ISSN: 0094-243X. DOI: 10.1063/1.1945212. URL: <https://doi.org/10.1063/1.1945212> (visited on 01/15/2024).
- [6] Robert G. Sargent. “Verification and Validation of Simulation Models”. In: *Proceedings of the 2010 Winter Simulation Conference*. Dec. 2010, pp. 166–183. DOI: 10.1109/WSC.2010.5679166. URL: <https://ieeexplore.ieee.org/document/5679166> (visited on 01/11/2024).
- [7] Paolo Padoan, Troels Haugbølle, and Åke Nordlund. “A SIMPLE LAW OF STAR FORMATION”. In: *The Astrophysical Journal Letters* 759.2 (Oct. 2012), p. L27. ISSN: 2041-

8205. DOI: 10.1088/2041-8205/759/2/L27. URL: <https://dx.doi.org/10.1088/2041-8205/759/2/L27> (visited on 01/15/2024).
- [8] David T. Allen et al. “Measurements of methane emissions at natural gas production sites in the United States”. en. In: *Proceedings of the National Academy of Sciences* 110.44 (Oct. 2013), pp. 17768–17773. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1304880110. URL: <http://www.pnas.org/content/110/44/17768> (visited on 09/04/2016).
- [9] François E. Cellier and Jurgen Greifeneder. *Continuous System Modeling*. Springer Science & Business Media, Mar. 2013. ISBN: 978-1-4757-3922-0.
- [10] Jack David Flicker and Jay Johnson. *Photovoltaic Ground Fault and Blind Spot Electrical Simulations*. Tech. rep. SAND2013-3459. Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), June 2013. DOI: 10.2172/1089985. URL: <https://www.osti.gov/biblio/1089985> (visited on 01/15/2024).
- [11] T Dessa Glasser. “Leveraging Data for Financial Stability Monitoring”. In: *Journal of Banking Regulation* 14.3 (July 2013), pp. 195–208. ISSN: 1750-2071. DOI: 10.1057/jbr.2013.11. URL: <https://doi.org/10.1057/jbr.2013.11> (visited on 01/15/2024).
- [12] Anna Karion et al. “Methane emissions estimate from airborne measurements over a western United States natural gas field”. en. In: *Geophysical Research Letters* 40.16 (Aug. 2013), pp. 4393–4397. ISSN: 1944-8007. DOI: 10.1002/grl.50811. URL: <http://onlinelibrary.wiley.com/doi/10.1002/grl.50811/abstract> (visited on 09/18/2016).
- [13] Susmita Bandyopadhyay and Ranjan Bhattacharya. *Discrete and Continuous Simulation: Theory and Practice*. CRC Press, June 2014. ISBN: 978-1-4665-9639-9.
- [14] Peter Fritzson. *Principles of Object Oriented Modeling and Simulation with Modelica 3.3*. 1st ed. John Wiley & Sons, Ltd, 2014. DOI: 10.1002/9781118989166. URL: <https://onlinelibrary.wiley.com/doi/10.1002/9781118989166> (visited on 01/11/2024).

- [15] I S Jha, Subir Sen, and Vineeta Agarwal. “Advanced Metering Infrastructure Analytics — A Case Study”. In: *2014 Eighteenth National Power Systems Conference (NPSC)*. Guwahati, India: IEEE, Dec. 2014, pp. 1–6. ISBN: 978-1-4799-5141-3. DOI: 10.1109/NPSC.2014.7103882. URL: <https://ieeexplore.ieee.org/document/7103882> (visited on 02/11/2024).
- [16] Ramyar Rashed Mohassel et al. “Application of Advanced Metering Infrastructure in Smart Grids”. In: *22nd Mediterranean Conference on Control and Automation*. June 2014, pp. 822–828. DOI: 10.1109/MED.2014.6961475. URL: <https://ieeexplore.ieee.org/abstract/document/6961475> (visited on 02/07/2024).
- [17] OAR US EPA. *Greenhouse Gas Reporting Program (GHGRP)*. Other Policies and Guidance. June 2014. URL: <https://www.epa.gov/ghgreporting> (visited on 02/07/2024).
- [18] David T. Allen et al. “Methane Emissions from Process Equipment at Natural Gas Production Sites in the United States: Liquid Unloadings”. In: *Environmental Science & Technology* 49.1 (Jan. 2015), pp. 641–648. ISSN: 0013-936X. DOI: 10.1021/es504016r. URL: <http://dx.doi.org/10.1021/es504016r> (visited on 08/19/2016).
- [19] David T. Allen et al. “Methane Emissions from Process Equipment at Natural Gas Production Sites in the United States: Pneumatic Controllers”. In: *Environmental Science & Technology* 49.1 (Jan. 2015), pp. 633–640. ISSN: 0013-936X. DOI: 10.1021/es5040156. URL: <http://dx.doi.org/10.1021/es5040156> (visited on 08/19/2016).
- [20] NASA Earth Science Data Systems. *ICARTT File Format Standards V2.0* | *Earthdata*. Program. Apr. 2015. URL: <https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/icartt-file-format> (visited on 01/13/2024).
- [21] Anna Karion et al. “Aircraft-Based Estimate of Total Methane Emissions from the Barnett Shale Region”. In: *Environmental Science & Technology* 49.13 (July 2015). tex.ids: kari-onAircraftBasedEstimateTotal2015a, pp. 8124–8131. ISSN: 0013-936X. DOI: 10.1021/acs.est.5b00217. URL: <http://dx.doi.org/10.1021/acs.est.5b00217> (visited on 09/04/2016).

- [22] Brian K. Lamb et al. “Direct Measurements Show Decreasing Methane Emissions from Natural Gas Local Distribution Systems in the United States”. In: *Environmental Science & Technology* 49.8 (Apr. 2015). tex.ids: lambDirectMeasurementsShow2015a, pp. 5161–5169. ISSN: 0013-936X. DOI: 10.1021/es505116p. URL: <http://dx.doi.org/10.1021/es505116p> (visited on 08/19/2016).
- [23] Anthony J. Marchese et al. “Methane Emissions from United States Natural Gas Gathering and Processing”. In: *Environmental Science & Technology* 49.17 (Sept. 2015), pp. 10718–10727. ISSN: 0013-936X. DOI: 10.1021/acs.est.5b02275. URL: <http://dx.doi.org/10.1021/acs.est.5b02275> (visited on 08/19/2016).
- [24] Austin L. Mitchell et al. “Measurements of Methane Emissions from Natural Gas Gathering Facilities and Processing Plants: Measurement Results”. In: *Environmental Science & Technology* 49.5 (Mar. 2015). tex.ids: mitchellMeasurementsMethaneEmissions2015a, pp. 3219–3227. ISSN: 0013-936X. DOI: 10.1021/es5052809. URL: <http://dx.doi.org/10.1021/es5052809> (visited on 08/19/2016).
- [25] J. Peischl et al. “Quantifying atmospheric methane emissions from the Haynesville, Fayetteville, and northeastern Marcellus shale gas production regions”. en. In: *Journal of Geophysical Research: Atmospheres* 120.5 (Mar. 2015). tex.ids: peischlQuantifyingAtmosphericMethane2015a, 2014JD022697. ISSN: 2169-8996. DOI: 10.1002/2014JD022697. URL: <http://onlinelibrary.wiley.com/doi/10.1002/2014JD022697/abstract> (visited on 09/04/2016).
- [26] C. Reimsbach-Kountaze. *The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis*. OECD Digital Economy Papers 245. Paris, Jan. 2015. DOI: 10.1787/5js7t9wqzvg8-en. URL: https://www.oecd-ilibrary.org/science-and-technology/the-proliferation-of-big-data-and-implications-for-official-statistics-and-statistical-agencies_5js7t9wqzvg8-en (visited on 01/14/2024).
- [27] J. R. Roscioli et al. “Measurements of methane emissions from natural gas gathering facilities and processing plants: measurement methods”. In: *Atmos. Meas. Tech.* 8.5 (May

- 2015), pp. 2017–2035. ISSN: 1867-8548. DOI: 10.5194/amt-8-2017-2015. URL: <http://www.atmos-meas-tech.net/8/2017/2015/> (visited on 08/19/2016).
- [28] R. Subramanian et al. “Methane Emissions from Natural Gas Compressor Stations in the Transmission and Storage Sector: Measurements and Comparisons with the EPA Greenhouse Gas Reporting Program Protocol”. In: *Environmental Science & Technology* 49.5 (Mar. 2015), pp. 3252–3261. ISSN: 0013-936X. DOI: 10.1021/es5060258. URL: <http://dx.doi.org/10.1021/es5060258> (visited on 09/04/2016).
- [29] Michael van Steen. “Regulatory Big Data: Regulator Goals and Global Initiatives”. In: *Moody’s Analytics Risk Perspectives V* (May 2015). URL: <https://www.moodyanalytics.com/risk-perspectives-magazine/risk-data-management/regulatory-spotlight/regulatory-big-data-regulator-goals-and-global-initiatives> (visited on 01/14/2024).
- [30] Tara I. Yacovitch et al. “Mobile Laboratory Observations of Methane Emissions in the Barnett Shale Region”. In: *Environmental Science & Technology* 49.13 (July 2015), pp. 7889–7895. ISSN: 0013-936X. DOI: 10.1021/es506352j. URL: <http://dx.doi.org/10.1021/es506352j> (visited on 09/04/2016).
- [31] Daniel Zavala-Araiza et al. “Reconciling divergent estimates of oil and gas methane emissions”. en. In: *Proceedings of the National Academy of Sciences* 112.51 (Dec. 2015), pp. 15597–15602. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1522126112. URL: <http://www.pnas.org/content/112/51/15597> (visited on 08/19/2016).
- [32] Daniel Zavala-Araiza et al. “Toward a Functional Definition of Methane Super-Emitters: Application to Natural Gas Production Sites”. In: *Environmental Science & Technology* 49.13 (July 2015). tex.ids: zavala-araizaFunctionalDefinitionMethane2015a, pp. 8167–8174. ISSN: 0013-936X. DOI: 10.1021/acs.est.5b00133. URL: <http://dx.doi.org/10.1021/acs.est.5b00133> (visited on 09/04/2016).
- [33] Daniel J. Zimmerle et al. “Methane Emissions from the Natural Gas Transmission and Storage System in the United States”. In: *Environmental Science & Technology* 49.15 (Aug.

- 2015), pp. 9374–9383. ISSN: 0013-936X. DOI: 10.1021/acs.est.5b01669. URL: <https://doi.org/10.1021/acs.est.5b01669> (visited on 07/22/2019).
- [34] Joannes D. Maasakkers et al. “Gridded National Inventory of U.S. Methane Emissions”. In: *Environmental Science & Technology* (Nov. 2016). ISSN: 0013-936X. DOI: 10.1021/acs.est.6b02878. URL: <http://dx.doi.org/10.1021/acs.est.6b02878> (visited on 11/20/2016).
- [35] Mark Omara et al. “Methane Emissions from Conventional and Unconventional Natural Gas Production Sites in the Marcellus Shale Basin”. In: *Environmental Science & Technology* 50.4 (Feb. 2016), pp. 2099–2107. ISSN: 0013-936X. DOI: 10.1021/acs.est.5b05503. URL: <http://dx.doi.org/10.1021/acs.est.5b05503> (visited on 08/19/2016).
- [36] J. Peischl et al. “Quantifying atmospheric methane emissions from oil and natural gas production in the Bakken shale region of North Dakota”. en. In: *Journal of Geophysical Research: Atmospheres* 121.10 (May 2016), 2015JD024631. ISSN: 2169-8996. DOI: 10.1002/2015JD024631. URL: <http://onlinelibrary.wiley.com/doi/10.1002/2015JD024631/abstract>.
- [37] David T. Allen, Felipe J. Cardoso-Saldaña, and Yosuke Kimura. “Variability in Spatially and Temporally Resolved Emissions and Hydrocarbon Source Fingerprints for Oil and Gas Sources in Shale Gas Production Regions”. en. In: *Environmental Science & Technology* 51.20 (Oct. 2017), pp. 12016–12026. ISSN: 0013-936X, 1520-5851. DOI: 10.1021/acs.est.7b02202. URL: <http://pubs.acs.org/doi/10.1021/acs.est.7b02202> (visited on 01/20/2019).
- [38] Clay Bell et al. “Comparison of methane emission estimates from multiple measurement techniques at natural gas production pads”. en. In: *Elem Sci Anth* 5.0 (Dec. 2017). tex.ids: bellComparisonMethaneEmission2017a. ISSN: 2325-1026. DOI: 10.1525/elementa.266. URL: <http://www.elementascience.org/articles/10.1525/elementa.266/> (visited on 06/28/2018).

- [39] Pat Ferrier. *Tourism study hints at why people visit Fort Collins*. en-US. Sept. 2017. URL: <https://www.coloradoan.com/story/money/business/2017/09/27/tourism-study-hints-why-people-visit-fort-collins/705070001/> (visited on 12/06/2023).
- [40] Fortinet. *Understanding the IoT Explosion and Its Impact on Enterprise Security*. Tech. rep. Feb. 2017. URL: <https://www.fortinet.com/content/dam/fortinet/assets/white-papers/WP-Understanding-The-IoT-Explosion-And-Its-Impact-On-Enterprise-Security.pdf> (visited on 02/05/2024).
- [41] Anna M. Robertson et al. “Variation in Methane Emission Rates from Well Pads in Four Oil and Gas Basins with Contrasting Production Volumes and Compositions”. In: *Environmental Science & Technology* (June 2017). ISSN: 0013-936X. DOI: 10.1021/acs.est.7b00571. URL: <http://dx.doi.org/10.1021/acs.est.7b00571>.
- [42] Stefan Schwietzke et al. “Improved Mechanistic Understanding of Natural Gas Methane Emissions from Spatially Resolved Aircraft Measurements”. In: *Environmental Science & Technology* (May 2017). ISSN: 0013-936X. DOI: 10.1021/acs.est.7b01810. URL: <http://dx.doi.org/10.1021/acs.est.7b01810>.
- [43] Masoud Soleymani Shishvan and Jörg Benndorf. “Operational Decision Support for Material Management in Continuous Mining Systems: From Simulation Concept to Practical Full-Scale Implementations”. In: *Minerals* 7.7 (July 2017), p. 116. ISSN: 2075-163X. DOI: 10.3390/min7070116. URL: <https://www.mdpi.com/2075-163X/7/7/116> (visited on 02/12/2024).
- [44] Peter Tozzi and Jin Ho Jo. “A Comparative Analysis of Renewable Energy Simulation Tools: Performance Simulation Model vs. System Optimization”. In: *Renewable and Sustainable Energy Reviews* 80 (Dec. 2017), pp. 390–398. ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.05.153. URL: <https://www.sciencedirect.com/science/article/pii/S1364032117307761> (visited on 02/05/2024).

- [45] Timothy L. Vaughn et al. “Comparing facility-level methane emission rate estimates at natural gas gathering and boosting stations”. en. In: *Elem Sci Anth* 5.0 (Nov. 2017). ISSN: 2325-1026. DOI: 10.1525/elementa.257. URL: <http://www.elementascience.org/articles/10.1525/elementa.257/> (visited on 12/05/2017).
- [46] Tara I. Yacovitch et al. “Natural gas facility methane emissions: measurements by tracer flux ratio in two US natural gas producing basins”. en. In: *Elem Sci Anth* 5.0 (Nov. 2017). tex.ids: yacovitchNaturalGasFacility2017a. ISSN: 2325-1026. DOI: 10.1525/elementa.251. URL: <http://www.elementascience.org/articles/10.1525/elementa.251/> (visited on 12/05/2017).
- [47] Hugh Boyes et al. “The Industrial Internet of Things (IIoT): An Analysis Framework”. In: *Computers in Industry* 101 (Oct. 2018), pp. 1–12. ISSN: 0166-3615. DOI: 10.1016/j.compind.2018.04.015. URL: <https://www.sciencedirect.com/science/article/pii/S0166361517307285> (visited on 02/07/2024).
- [48] Antonello Monti et al. “A Global Real-Time Superlab: Enabling High Penetration of Power Electronics in the Electric Grid”. In: *IEEE Power Electronics Magazine* 5.3 (Sept. 2018), pp. 35–44. ISSN: 2329-9215. DOI: 10.1109/MPEL.2018.2850698. URL: <https://ieeexplore.ieee.org/abstract/document/8458285/authors#authors> (visited on 01/03/2024).
- [49] Emiliano Sisinni et al. “Industrial Internet of Things: Challenges, Opportunities, and Directions”. In: *IEEE Transactions on Industrial Informatics* 14.11 (Nov. 2018), pp. 4724–4734. ISSN: 1941-0050. DOI: 10.1109/TII.2018.2852491. URL: <https://ieeexplore.ieee.org/abstract/document/8401919> (visited on 02/07/2024).
- [50] Timothy L. Vaughn et al. “Temporal variability largely explains top-down/bottom-up difference in methane emission estimates from a natural gas production region”. en. In: *Proceedings of the National Academy of Sciences* (Oct. 2018), p. 201805687. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1805687115. URL: <http://www.pnas.org/content/early/2018/10/24/1805687115> (visited on 10/29/2018).

- [51] Daniel Zimmerle, Arthur Felicio B. Santos, and Gerald P. Duggan. “DC Approximate Models for Modeling Minigrid Systems”. In: *2018 IEEE Global Humanitarian Technology Conference (GHTC)*. Oct. 2018, pp. 1–7. DOI: 10.1109/GHTC.2018.8601930.
- [52] Somudeep Bhattacharjee and Champa Nandi. “Implementation of Industrial Internet of Things in the Renewable Energy Sector”. In: *The Internet of Things in the Industrial Sector: Security and Device Connectivity, Smart Environments, and Industry 4.0*. Ed. by Zaigham Mahmood. Computer Communications and Networks. Cham: Springer International Publishing, 2019, pp. 223–259. ISBN: 978-3-030-24892-5. DOI: 10.1007/978-3-030-24892-5_10. URL: https://doi.org/10.1007/978-3-030-24892-5_10 (visited on 02/07/2024).
- [53] Riley M. Duren et al. “California’s methane super-emitters”. en. In: *Nature* 575.7781 (Nov. 2019), pp. 180–184. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1720-3. URL: <https://www.nature.com/articles/s41586-019-1720-3> (visited on 11/28/2019).
- [54] Stephen Fenberg et al. *CONCERNING ADDITIONAL PUBLIC WELFARE PROTECTIONS REGARDING THE CONDUCT OF OIL AND GAS OPERATIONS, AND, IN CONNECTION THEREWITH, MAKING AN APPROPRIATION*. Apr. 2019. URL: <https://leg.colorado.gov/bills/sb19-181>.
- [55] Benjamin Luck et al. *Methane Emissions from Gathering and Boosting Compressor Stations in the U.S. Supporting Volume 1: Multi-Day Measurements of Pneumatic Controller Emissions*. en. Tech. rep. Colorado State University, Apr. 2019, p. 120.
- [56] Benjamin Luck et al. “Multiday Measurements of Pneumatic Controller Emissions Reveal the Frequency of Abnormal Emissions Behavior at Natural Gas Gathering Stations”. In: *Environmental Science & Technology Letters* 6.6 (Apr. 2019), pp. 348–352. DOI: 10.1021/acs.estlett.9b00158. URL: <https://doi.org/10.1021/acs.estlett.9b00158> (visited on 05/15/2019).

- [57] Mark Philips. *IoT in Oil & Gas: Exploring the New Digital Frontier*. Oct. 2019. URL: <https://www.iiot-world.com/industrial-iot/connected-industry/iot-in-oil-gas-exploring-the-new-digital-frontier/> (visited on 02/07/2024).
- [58] Fei Tao et al. “Digital Twin in Industry: State-of-the-Art”. In: *IEEE Transactions on Industrial Informatics* 15.4 (Apr. 2019), pp. 2405–2415. ISSN: 1941-0050. DOI: 10.1109/TII.2018.2873186. URL: https://ieeexplore.ieee.org/abstract/document/8477101?casa_token=er-VQwaZU-UAAAAA:wGbza6tEmyLedUUNeUTM94hKplZhHPIQn3zteAYaaCDP2Dcv1rg4uxB3VGQbi61XCv6jB-mlXQ (visited on 01/17/2024).
- [59] Alexander J. Turner, Christian Frankenberg, and Eric A. Kort. “Interpreting contemporary trends in atmospheric methane”. en. In: *Proceedings of the National Academy of Sciences* 116.8 (Feb. 2019). Publisher: National Academy of Sciences Section: Perspective, pp. 2805–2813. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1814297116. URL: <https://www.pnas.org/content/116/8/2805> (visited on 12/18/2021).
- [60] U.S. Department of Energy. *Smart Grid Investment Grant Program: Overview: Recovery Act: Archive* | SmartGrid.Gov. Dec. 2019. URL: https://www.smartgrid.gov/archive/recovery_act/overview/smart_grid_investment_grant_program (visited on 02/07/2024).
- [61] Daniel Zimmerle et al. *Characterization of Methane Emissions from Gathering Compressor Stations: Final Report*. en. Tech. rep. Colorado State University, Apr. 2019, p. 62. URL: <http://dx.doi.org/10.25675/10217/194544>.
- [62] Daniel Zimmerle et al. *Methane Emissions from Gathering and Boosting Compressor Stations in the U.S.: Supporting Volume 3: Emission Factors, Station Estimates and National Emissions*. en. DOE Report. Fort Collins, Colorado: Colorado State University, Apr. 2019, p. 63. URL: <http://dx.doi.org/10.25675/10217/194541>.
- [63] Azzam Abu-Rayash and Ibrahim Dincer. “Analysis of the electricity demand trends amidst the COVID-19 coronavirus pandemic”. eng. In: *Energy Research & Social Science* 68 (Oct. 2020), p. 101682. ISSN: 2214-6296. DOI: 10.1016/j.erss.2020.101682.

- [64] Duzgun Agdas and Prabir Barooah. “Impact of the COVID-19 Pandemic on the U.S. Electricity Demand and Supply: An Early View From Data”. In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 151523–151534. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3016912. URL: <https://ieeexplore.ieee.org/document/9169615> (visited on 12/05/2023).
- [65] Lydia Denworth. *The Biggest Psychological Experiment in History Is Running Now*. en. July 2020. URL: <https://www.scientificamerican.com/article/the-biggest-psychological-experiment-in-history-is-running-now1/> (visited on 12/05/2023).
- [66] Gerald P. Duggan, Daniel Zimmerle, and Sonu Upadhyay. “Big Data Analytics for Power Distribution Systems Using AMI and Open Source Tools”. In: *2020 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*. Oct. 2020, pp. 1–5. DOI: 10.1109/TD39804.2020.9299952.
- [67] David Jones et al. “Characterising the Digital Twin: A Systematic Literature Review”. In: *CIRP Journal of Manufacturing Science and Technology* 29 (May 2020), pp. 36–52. ISSN: 1755-5817. DOI: 10.1016/j.cirpj.2020.02.002. URL: <https://www.sciencedirect.com/science/article/pii/S1755581720300110> (visited on 01/16/2024).
- [68] Alexei V. Milkov et al. “Using global isotopic data to constrain the role of shale gas production in recent increases in atmospheric methane”. en. In: *Scientific Reports* 10.1 (Mar. 2020), p. 4199. ISSN: 2045-2322. DOI: 10.1038/s41598-020-61035-w. URL: <https://www.nature.com/articles/s41598-020-61035-w> (visited on 12/18/2021).
- [69] Stephen Snow et al. “Drivers behind Residential Electricity Demand Fluctuations Due to COVID-19 Restrictions”. en. In: *Energies* 13.21 (Jan. 2020). Number: 21 Publisher: Multidisciplinary Digital Publishing Institute, p. 5738. ISSN: 1996-1073. DOI: 10.3390/en13215738. URL: <https://www.mdpi.com/1996-1073/13/21/5738> (visited on 12/05/2023).
- [70] Jemma Townson. *Commercial down v residential up: COVID-19’s electricity impact*. en-AU. Apr. 2020. URL: <https://www.energynetworks.com.au/news/energy-insider/2020-ene>

rgy-insider/commercial-down-v-residential-up-covid-19s-electricity-impact/ (visited on 12/05/2023).

- [71] Daniel Zimmerle et al. “Methane Emissions from Gathering Compressor Stations in the U.S.” en. In: *Environmental Science & Technology* 54.12 (June 2020), pp. 7552–7561. ISSN: 0013-936X, 1520-5851. DOI: 10.1021/acs.est.0c00516. URL: <https://pubs.acs.org/doi/10.1021/acs.est.0c00516> (visited on 10/14/2020).
- [72] Ben Anderson and Patrick James. “Covid-19 lockdown: impacts on GB electricity demand and CO2 emissions”. en-US. In: 2.1 (Feb. 2021). Number: 1 Publisher: Ubiquity Press, pp. 134–149. ISSN: 2632-6655. DOI: 10.5334/bc.77. URL: <https://journal-buildingscities.org/articles/10.5334/bc.77> (visited on 12/05/2023).
- [73] Casey D. Burleyson et al. “Multiscale effects masked the impact of the COVID-19 pandemic on electricity demand in the United States”. In: *Applied Energy* 304 (Dec. 2021), p. 117711. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.117711. URL: <https://www.sciencedirect.com/science/article/pii/S0306261921010631> (visited on 12/05/2023).
- [74] Felipe J. Cardoso-Saldaña et al. “A Searchable Database for Prediction of Emission Compositions from Upstream Oil and Gas Sources”. In: *Environmental Science & Technology* 55.5 (Mar. 2021). Publisher: American Chemical Society, pp. 3210–3218. ISSN: 0013-936X. DOI: 10.1021/acs.est.0c05925. URL: <https://doi.org/10.1021/acs.est.0c05925> (visited on 04/14/2021).
- [75] Daniel H. Cusworth et al. “Intermittency of Large Methane Emitters in the Permian Basin”. In: *Environmental Science & Technology Letters* 8.7 (July 2021). Publisher: American Chemical Society, pp. 567–573. DOI: 10.1021/acs.estlett.1c00173. URL: <https://doi.org/10.1021/acs.estlett.1c00173> (visited on 08/11/2021).
- [76] Xuechen Gui et al. “The impact of COVID-19 on higher education building energy use and implications for future education building energy studies”. In: *Energy and Buildings* 251 (Nov. 2021), p. 111346. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2021.111346.

- URL: <https://www.sciencedirect.com/science/article/pii/S0378778821006307> (visited on 12/05/2023).
- [77] Peng Jiang, Yee Van Fan, and Jiří Jaromír Klemeš. “Impacts of COVID-19 on energy demand and consumption: Challenges, lessons and emerging opportunities”. In: *Applied Energy* 285 (Mar. 2021), p. 116441. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.116441. URL: <https://www.sciencedirect.com/science/article/pii/S030626192100009X> (visited on 12/05/2023).
- [78] Hyuna Kang et al. “Changes in energy consumption according to building use type under COVID-19 pandemic in South Korea”. In: *Renewable and Sustainable Energy Reviews* 148 (Sept. 2021), p. 111294. ISSN: 1364-0321. DOI: 10.1016/j.rser.2021.111294. URL: <https://www.sciencedirect.com/science/article/pii/S1364032121005815> (visited on 12/05/2023).
- [79] Emily Kawka and Kristen Cetin. “Impacts of COVID-19 on residential building energy use and performance”. In: *Building and Environment* 205 (Nov. 2021), p. 108200. ISSN: 0360-1323. DOI: 10.1016/j.buildenv.2021.108200. URL: <https://www.sciencedirect.com/science/article/pii/S0360132321006016> (visited on 12/05/2023).
- [80] Desen Kirli, Maximilian Parzen, and Aristides Kiprakis. “Impact of the COVID-19 Lock-down on the Electricity System of Great Britain: A Study on Energy Demand, Generation, Pricing and Grid Stability”. en. In: *Energies* 14.3 (Jan. 2021). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 635. ISSN: 1996-1073. DOI: 10.3390/en14030635. URL: <https://www.mdpi.com/1996-1073/14/3/635> (visited on 12/05/2023).
- [81] Moncef Krarti and Mohammad Aldubyan. “Review analysis of COVID-19 impact on electricity demand for residential buildings”. In: *Renewable and Sustainable Energy Reviews* 143 (June 2021), p. 110888. ISSN: 1364-0321. DOI: 10.1016/j.rser.2021.110888. URL: <https://www.sciencedirect.com/science/article/pii/S1364032121001829> (visited on 12/05/2023).

- [82] Thomas Lauvaux et al. “Global Assessment of Oil and Gas Methane Ultra-Emitters”. In: *arXiv:2105.06387 [physics]* (Apr. 2021). arXiv: 2105.06387. URL: <http://arxiv.org/abs/2105.06387> (visited on 11/26/2021).
- [83] David R. Lyon et al. “Concurrent variation in oil and gas methane emissions and oil price during the COVID-19 pandemic”. English. In: *Atmospheric Chemistry and Physics* 21.9 (May 2021). Publisher: Copernicus GmbH, pp. 6605–6626. ISSN: 1680-7316. DOI: 10.5194/acp-21-6605-2021. URL: <https://acp.copernicus.org/articles/21/6605/2021/> (visited on 08/23/2021).
- [84] Daniel Mehlig, Helen ApSimon, and Iain Staffell. “The impact of the UK’s COVID-19 lockdowns on energy demand and emissions”. en. In: *Environmental Research Letters* 16.5 (Apr. 2021). Publisher: IOP Publishing, p. 054037. ISSN: 1748-9326. DOI: 10.1088/1748-9326/abf876. URL: <https://dx.doi.org/10.1088/1748-9326/abf876> (visited on 12/05/2023).
- [85] Paul Overberg and Kevin Hand. “How to Understand the Data Explosion”. In: *Wall Street Journal* (Dec. 2021). ISSN: 0099-9660. URL: <https://www.wsj.com/articles/how-to-understand-the-data-explosion-11638979214> (visited on 01/14/2024).
- [86] Jean Rouleau and Louis Gosselin. “Impacts of the COVID-19 lockdown on energy consumption in a Canadian social housing building”. In: *Applied Energy* 287 (Apr. 2021), p. 116565. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2021.116565. URL: <https://www.sciencedirect.com/science/article/pii/S0306261921001124> (visited on 12/05/2023).
- [87] I. Santiago et al. “Electricity demand during pandemic times: The case of the COVID-19 in Spain”. In: *Energy Policy* 148 (Jan. 2021), p. 111964. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2020.111964. URL: <https://www.sciencedirect.com/science/article/pii/S0301421520306753> (visited on 12/05/2023).
- [88] A. Tingting Xu et al. “Impact of the COVID-19 pandemic on the reduction of electricity demand and the integration of renewable energy into the power grid”. In: *Journal of*

- Renewable and Sustainable Energy* 13.2 (Apr. 2021), p. 026304. ISSN: 1941-7012. DOI: 10.1063/5.0045825. URL: <https://doi.org/10.1063/5.0045825> (visited on 12/05/2023).
- [89] OAR US EPA. *EPA Proposes New Source Performance Standards Updates, Emissions Guidelines to Reduce Methane and Other Harmful Pollution from the Oil and Natural Gas Industry*. en. Other Policies and Guidance. Nov. 2021. URL: <https://www.epa.gov/controlling-air-pollution-oil-and-natural-gas-industry/epa-proposes-new-source-performance> (visited on 12/18/2021).
- [90] Zhen Zhang et al. “Anthropogenic emissions are the main contribution to the rise of atmospheric methane (1993-2017)”. In: *National Science Review* (Nov. 2021), nwab200. ISSN: 2095-5138. DOI: 10.1093/nsr/nwab200. URL: <https://doi.org/10.1093/nsr/nwab200> (visited on 12/18/2021).
- [91] Hutama A. Bramantyo, Bagus Satrio Utomo, and Efrilia M. Khusna. “Data Processing for IoT in Oil and Gas Refineries”. In: *Journal of Telecommunication Network (Jurnal Jaringan Telekomunikasi)* 12.1 (Mar. 2022), pp. 48–54. ISSN: 2654-6531. DOI: 10.33795/jartel.v12i1.300. URL: <https://jartel.polinema.ac.id/index.php/jartel/article/view/300> (visited on 02/07/2024).
- [92] Shenzhen Clou. *Advanced Metering Infrastructure and Smart Grid*. Nov. 2022. URL: <https://www.smart-energy.com/industry-sectors/smart-grid/advanced-metering-infrastructure-and-smart-grid/> (visited on 02/07/2024).
- [93] Office of the Commissioner. “Unleashing the Power of Data”. In: *FDA* (Sept. 2022). URL: <https://www.fda.gov/science-research/focus-areas-regulatory-science-report/unleashing-power-data> (visited on 01/14/2024).
- [94] Ramit Debnath et al. “Lockdown impacts on residential electricity demand in India: A data-driven and non-intrusive load monitoring study using Gaussian mixture models”. In: *Energy Policy* 164 (May 2022), p. 112886. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2022

- .112886. URL: <https://www.sciencedirect.com/science/article/pii/S0301421522001112> (visited on 12/05/2023).
- [95] Arthur Lin Ku et al. “Changes in hourly electricity consumption under COVID mandates: A glance to future hourly residential power consumption pattern with remote work in Arizona”. In: *Applied Energy* 310 (Mar. 2022), p. 118539. ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2022.118539. URL: <https://www.sciencedirect.com/science/article/pii/S0306261922000265> (visited on 12/05/2023).
- [96] Tomasz Rokicki et al. “Changes in Energy Consumption and Energy Intensity in EU Countries as a Result of the COVID-19 Pandemic by Sector and Area Economy”. en. In: *Energies* 15.17 (Jan. 2022). Number: 17 Publisher: Multidisciplinary Digital Publishing Institute, p. 6243. ISSN: 1996-1073. DOI: 10.3390/en15176243. URL: <https://www.mdpi.com/1996-1073/15/17/6243> (visited on 12/05/2023).
- [97] Miguel Sánchez-López et al. “The diverse impacts of COVID-19 on electricity demand: The case of Chile”. In: *International Journal of Electrical Power & Energy Systems* 138 (June 2022), p. 107883. ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2021.107883. URL: <https://www.sciencedirect.com/science/article/pii/S0142061521010978> (visited on 12/05/2023).
- [98] Fei Tao et al. “Digital Twin Modeling”. In: *Journal of Manufacturing Systems* 64 (July 2022), pp. 372–389. ISSN: 0278-6125. DOI: 10.1016/j.jmsy.2022.06.015. URL: <https://www.sciencedirect.com/science/article/pii/S0278612522001108> (visited on 01/16/2024).
- [99] Varun Thakur and Abhinav Mogha. “IIOT in Utility Operations Paper”. In: *ISUW 2021*. Ed. by Reji Kumar Pillai, B. P. Singh, and N. Murugesan. Lecture Notes in Electrical Engineering. Singapore: Springer Nature, 2022, pp. 45–59. ISBN: 9789811687273. DOI: 10.1007/978-981-16-8727-3_5.
- [100] Daniel Zimmerle et al. “Modeling Air Emissions from Complex Facilities at Detailed Temporal and Spatial Resolution: The Methane Emission Estimation Tool (MEET)”. In: *Sci-*

- ence of The Total Environment* 824 (Feb. 2022), p. 153653. DOI: 10.1016/j.scitotenv.2022.153653.
- [101] DataProt. *Internet of Things Statistics for 2023 - Taking Things Apart*. Apr. 2023. URL: <https://dataprot.net/statistics/iot-statistics/> (visited on 02/05/2024).
- [102] Antonia Du Bois. *Managing the Explosion of Healthcare Data to Harness Its Power*. May 2023. URL: <https://www.news-medical.net/health/Managing-the-Explosion-of-Healthcare-Data-to-Harness-its-Power.aspx> (visited on 01/14/2024).
- [103] Gerald P. Duggan et al. “Electricity Consumption in Commercial Buildings During Covid-19”. In: *Buildings and Cities* 4.1 (Oct. 2023), pp. 851–866. ISSN: 2632-6655. DOI: 10.5334/bc.361. URL: <https://journal-buildingscities.org/articles/10.5334/bc.361> (visited on 11/26/2023).
- [104] IABAC. *IoT Data Engineering: Harnessing the Power of the Internet of Things*. Nov. 2023. URL: <https://iabac.medium.com/iot-data-engineering-harnessing-the-power-of-the-internet-of-things-532fc6494cd7> (visited on 02/05/2024).
- [105] George Jackson. *What Is a Physical System? Discover the Fundamental Aspects of Physical Systems [Expert Guide!]* July 2023. URL: <https://physics-network.org/what-is-a-physical-system-discover-the-fundamental-aspects-of-physical-systems/> (visited on 01/13/2024).
- [106] Arvind P. Ravikumar et al. “Measurement-Based Differentiation of Low-Emission Global Natural Gas Supply Chains”. In: *Nature Energy* 8.11 (Nov. 2023), pp. 1174–1176. ISSN: 2058-7546. DOI: 10.1038/s41560-023-01381-x. URL: <https://www.nature.com/articles/s41560-023-01381-x> (visited on 03/11/2024).
- [107] Britannica. *Big Data | Definition, Uses, Tools, Challenges, & Facts | Britannica*. Jan. 2024. URL: <https://www.britannica.com/technology/big-data> (visited on 01/13/2024).
- [108] Deloitte. *Internet of Things in Oil & Gas | Deloitte US*. 2024. URL: <https://www2.deloitte.com/us/en/pages/consulting/articles/iot-digital-oil-and-gas.html> (visited on 02/07/2024).

- [109] Winrose Mollel, Daniel Zimmerle, and Arthur Santos. “Using Prototypical Sites to Model Methane Emissions in Colorado’s Denver-Julesburg Basin Using Mechanistic Emission Estimation Tool”. In: *In Preparation* (2024).
- [110] Arthur Santos et al. “Using Measurement-Informed Inventory to Assess Emissions in the Denver-Julesburg Basin”. In: *In Preparation* (2024).
- [111] OA US EPA. *Biden-Harris Administration Announces Proposed Rule to Reduce Wasteful Methane Emissions from the Oil and Gas Sector to Drive Innovation and Protect Communities*. News Release. Jan. 2024. URL: <https://www.epa.gov/newsreleases/biden-harris-administration-announces-proposed-rule-reduce-wasteful-methane-emissions> (visited on 02/07/2024).
- [112] 2020 Executive Orders | Colorado Governor Jared Polis. URL: <https://www.colorado.gov/governor/2020-executive-orders> (visited on 12/05/2023).
- [113] Apache. *Apache Parquet*. URL: <https://parquet.apache.org/> (visited on 01/22/2024).
- [114] asq. *What Is Statistical Process Control? SPC Quality Tools* | ASQ. URL: <https://asq.org/quality-resources/statistical-process-control> (visited on 03/26/2024).
- [115] builtin. *Data Validation — Overview, Types, How To Perform* | Built In. URL: <https://builtin.com/data-science/data-validation> (visited on 01/13/2024).
- [116] CarbonMapper. *Methane, CO2 Data | Global Open Portal | Carbon Mapper*. en-US. URL: <https://carbonmapper.org/data/> (visited on 12/18/2021).
- [117] CFI. *Data Validation*. URL: <https://corporatefinanceinstitute.com/resources/data-science/data-validation/> (visited on 01/13/2024).
- [118] Colorado. *Oil and Natural Gas Annual Emission Inventory Reporting* | Department of Public Health & Environment. URL: <https://cdphe.colorado.gov/ongaeir> (visited on 01/15/2024).
- [119] datagov. *Open Energy Data at DOE | Resources.Data.Gov*. URL: <https://resources.data.gov/resources/fdspp-doe-open-energy/> (visited on 01/14/2024).

- [120] EIA. *U.S. Natural Gas Marketed Production*. URL: <https://www.eia.gov/dnav/ng/hist/n9050us2a.htm> (visited on 02/18/2019).
- [121] energygov. *Open Data Catalogue*. URL: <https://www.energy.gov/data/articles/open-data-catalogue> (visited on 01/14/2024).
- [122] energygov. *Who Uses Open Data?* URL: <https://www.energy.gov/data/articles/who-uses-open-data> (visited on 01/14/2024).
- [123] *Envirofacts* (<https://enviro.epa.gov/index.html>). en. URL: <https://enviro.epa.gov/index.html> (visited on 10/31/2019).
- [124] US EPA. *Inventory of U.S. Greenhouse Gas Emissions and Sinks 1990-2016*. Tech. rep. EPA 430-R-18-003.
- [125] US EPA. *Subpart W Reporting Form for RY2017*. URL: <https://ccdsupport.com/confluence/display/help/Reporting+Form+Instructions> (visited on 02/04/2019).
- [126] FERC. *Data Sources | Federal Energy Regulatory Commission*. URL: <https://www.ferc.gov/industries-data/resources/data-sources> (visited on 01/14/2024).
- [127] FERC. *Data Strategy | Federal Energy Regulatory Commission*. URL: <https://www.ferc.gov/data-strategy> (visited on 01/14/2024).
- [128] Environmental Defense Fund. *EDF Permian Map Data Viewer*. en. URL: <https://www.permianmap.org> (visited on 12/18/2021).
- [129] *Greenhouse Gas Reporting Program, Subpart W*. en. URL: <https://www.law.cornell.edu/cfr/text/40/98.236> (visited on 01/04/2019).
- [130] IBM. *What Is Model-Based Systems Engineering (MBSE)?* URL: <https://www.ibm.com/topics/model-based-systems-engineering> (visited on 01/11/2024).

- [131] Jet Propulsion Laboratory. *Methane Source Finder*. URL: [https://methane.jpl.nasa.gov/map#b=ESRI_World_Imagery&l=VISTA_2017_CNG_FUELING_STATIONS\(1\),VISTA_2017_FEED_LOTS\(1\),VISTA_2017_DIGESTER\(1\),VISTA_2017_PETRO_REFINE\(1\),VISTA_2017_LANDFILL\(1\),VISTA_2017_COMPOSTING_SITES\(1\),VISTA_2017_OILGAS_WELLS\(1\),VISTA_2017_OILGAS_FACILITY_BOUNDARIES\(1\),VISTA_2017_LNG_FUELING_STATIONS\(1\),VISTA_2017_LIVESTOCK_DAIRIES\(1\),VISTA_2017_NAT_GAS_PROC_PLANT\(1\),VISTA_2017_COMPRESSOR\(1\),VISTA_2017_POWER_PLANT\(1\),VISTA_2017_OILGAS_FIELDS\(1\),VISTA_2017_NAT_GAS_STORE_FIELDS\(1\),AVIRIS\(1\),VISTA_2017_WASTEWTR_TREAT_PLNT\(1\),AVIRIS_SOURCES\(1\)&vm=2D&ve=-107,28.1031615598542,-98,36.82168523696883&pl=false&pb=false&tr=true&d=2021-12-15&tlr=days](https://methane.jpl.nasa.gov/map#b=ESRI_World_Imagery&l=VISTA_2017_CNG_FUELING_STATIONS(1),VISTA_2017_FEED_LOTS(1),VISTA_2017_DIGESTER(1),VISTA_2017_PETRO_REFINE(1),VISTA_2017_LANDFILL(1),VISTA_2017_COMPOSTING_SITES(1),VISTA_2017_OILGAS_WELLS(1),VISTA_2017_OILGAS_FACILITY_BOUNDARIES(1),VISTA_2017_LNG_FUELING_STATIONS(1),VISTA_2017_LIVESTOCK_DAIRIES(1),VISTA_2017_NAT_GAS_PROC_PLANT(1),VISTA_2017_COMPRESSOR(1),VISTA_2017_POWER_PLANT(1),VISTA_2017_OILGAS_FIELDS(1),VISTA_2017_NAT_GAS_STORE_FIELDS(1),AVIRIS(1),VISTA_2017_WASTEWTR_TREAT_PLNT(1),AVIRIS_SOURCES(1)&vm=2D&ve=-107,28.1031615598542,-98,36.82168523696883&pl=false&pb=false&tr=true&d=2021-12-15&tlr=days) (visited on 12/18/2021).
- [132] kahnacademy. *Interquartile Range (IQR) (Video)* | Khan Academy. URL: <https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6th/v/calculating-interquartile-range-iqr> (visited on 03/26/2024).
- [133] khanacademy. *How to Calculate Z-scores (Formula Review) (Article)*. URL: <https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/z-scores/a/z-scores-review> (visited on 03/26/2024).
- [134] NetworkX. *NetworkX — NetworkX Documentation*. URL: <https://networkx.org/> (visited on 02/19/2024).
- [135] OGC. *OGC Hierarchical Data Format Version 5 (HDF5®) Standard*. URL: <https://www.ogc.org/standard/hdf5/> (visited on 01/22/2024).
- [136] OGC. *OGC Network Common Data Form (netCDF) Standards Suite*. URL: <https://www.ogc.org/standard/netcdf/> (visited on 01/22/2024).
- [137] OpalRT. *Opal-RT*. URL: <https://www.opal-rt.com/> (visited on 03/02/2024).
- [138] OpenEI. *Data* | Open Energy Information. URL: <https://openei.org/wiki/Data> (visited on 01/14/2024).

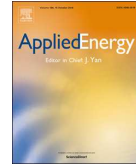
- [139] Rebecca Powell. *Coronavirus shopping surge causes King Soopers stores to change hours of operation*. en-US. URL: <https://www.coloradoan.com/story/news/2020/03/14/coronavirus-surge-prompts-king-soopers-grocery-stores-change-hours/5052028002/> (visited on 12/05/2023).
- [140] pydata.org. *Pandas - Python Data Analysis Library*. URL: <https://pandas.pydata.org/> (visited on 03/26/2024).
- [141] Larry B. Rainey and Andreas Tolk. “Overview and Introduction to Modeling and Simulation Support for System of Systems Engineering Applications”. In: (). URL: <https://onlinelibrary.wiley.com/doi/chapter-epub/10.1002/9781118501757.ch1> (visited on 01/11/2024).
- [142] RTDS. *Real-Time Simulation with the RTDS Simulator*. URL: <https://www.rtds.com/> (visited on 03/02/2024).
- [143] *State-by-State COVID Guidance: Colorado*. en. URL: <https://www.huschblackwell.com/colorado-state-by-state-covid-19-guidance> (visited on 12/05/2023).
- [144] statistica. *Number of U.S. cities, towns, villages by population size 2019*. en. URL: <https://www.statista.com/statistics/241695/number-of-us-cities-towns-villages-by-population-size/> (visited on 12/05/2023).
- [145] Denis Tanguay. *International Performance Measurement and Verification Protocol (IP-MVP)*. en-GB. URL: <https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp> (visited on 12/05/2023).
- [146] ROOT team. *ROOT: analyzing petabytes of data, scientifically*. en. URL: <https://root.cern/> (visited on 12/05/2023).
- [147] Texas. *Data Sets Available for Download*. URL: <https://www.rrc.texas.gov/resource-center/research/data-sets-available-for-download/> (visited on 01/15/2024).
- [148] *The 300 Largest Cities in the United States by Population 2023*. URL: <https://worldpopulationreview.com/us-cities> (visited on 12/05/2023).

- [149] NOAA US Department of Commerce. *Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases*. EN-US. URL: https://gml.noaa.gov/ccgg/trends%5C_ch4/ (visited on 12/18/2021).
- [150] USEIA. *Homepage - U.S. Energy Information Administration (EIA)*. URL: <https://www.eia.gov/index.php> (visited on 01/14/2024).
- [151] USEPA. *40 CFR Part 98 Subpart W – Petroleum and Natural Gas Systems*. URL: <https://www.ecfr.gov/current/title-40/part-98/subpart-W> (visited on 02/18/2024).

Appendix A

Advanced Metering Infrastructure Modeling

A.1 A Method to Estimate Residential PV Generation from Net-Metered Load Data and System Install Date



A method to estimate residential PV generation from net-metered load data and system install date



Wendell Stainsby, Daniel Zimmerle*, Gerald P. Duggan

Department of Systems Engineering, Colorado State University, United States

HIGHLIGHTS

- Time periods paired by weather provide solid before/after load change estimates.
- Large AMI data (850 premises/4 yrs) drive PV estimates to $\pm 20\%$ of physical models.
- Method identifies PV impacts that are difficult to model with physical models.
- Generalizable to other load changes, such as tariff changes or efficiency programs.

ARTICLE INFO

Keywords:

PV generation
Advanced metering infrastructure
Disaggregation
Behind-the-meter

ABSTRACT

In the USA, residential photovoltaic (PV) systems are often configured for net metering “behind-the-meter”, where PV energy generation and building energy demand are reported as a combined net load to advanced metering infrastructure (AMI) meters, impeding estimates of PV generation. This work presents a methodology for modeling individual array and system-wide PV generation using only weather data, premise AMI data, and the approximate date of PV installation – information available to most distribution utilities. The study uses 36 months of data spanning nearly 850 homes with installed PV systems in Fort Collins, Colorado, USA. The algorithm estimates building energy consumption by comparing time periods before PV installation with similar periods after PV installation that have common weather and activity characteristics. Estimated building energy consumption is then compared with AMI meter data to estimate otherwise unobservable solar generation. To assess accuracy, modeled outputs are compared with directly metered PV generation and white-box physical models of PV production. Considering aggregate, utility-wide, generation estimates for the three year study period, the proposed method estimates over 75% of all days to within $\pm 20\%$ of established physical models. The method estimates more effectively in summer months when PV generation peaks and is of most interest to utilities. The model often outperforms physical models for days with snow cover and for arrays with shading or complex multi-roof implementations. The model also supports day-ahead PV prediction using forecasted weather data.

1. Introduction

By 2019, the U.S. had 70 GW_{DC} of solar capacity [1] with around 20% of that being residential rooftop photovoltaic (PV) systems [2]. Between 2014 and 2018, the compound annual growth rate of small scale (under 1 MW) photovoltaic (PV) generation in the United States was 28.2% [3]. The 2 millionth solar array was installed in early 2019 [4] and within 5 years over 2.5% of all U.S. homes will have PV[5]. Additionally, the National Renewable Energy Laboratory has found that the technical solar potential of suitable residential buildings was over 700 GW_{DC}, which could generate 926 TWh/year, a quarter of the United

States’ total annual electricity consumption [6].

The growth of solar has increased interest in better estimates of real-time generation from PV systems for use by distribution utilities [7]. Larger commercial and utility-scale solar deployments are usually directly metered, providing actual generation, typically at high time resolution and with minimal latency. Thus, both system owners and electrical system operators know the actual energy produced by large solar arrays. These data can be used to forecast expected generation.

However, residential solar systems in the USA are typically net-metered by the distribution operator, with 38 states having net-metering policies by 2019 [8]. In traditionally metered systems with PV,

* Corresponding author at: 430 N. College Avenue, Fort Collins, CO, United States.

E-mail address: Dan.Zimmerle@colostate.edu (D. Zimmerle).

<https://doi.org/10.1016/j.apenergy.2020.114895>

Received 7 November 2019; Received in revised form 12 March 2020; Accepted 24 March 2020

Available online 07 April 2020

0306-2619/ © 2020 Elsevier Ltd. All rights reserved.

system operators have access only to net load for the billing period, obscuring any information about the PV generation. When available, digital smart meter data generally reports only net-load or net-generation (generation net of load) for each meter reading. Since generated PV offsets concurrent load in a net-metered building, it is difficult to separate behind-the-meter (BTM) generation from load. Therefore, distribution utilities are essentially blind to PV production at the vast majority of residential solar sites. This lack of insight into solar generation complicates solar forecasting and the balancing of dispatchable generation with demand [9].

Directly estimating generation for residential arrays is also complex. Modeling requires numerous variables, including weather, physical orientation of the array, system size and electrical characteristics. In addition, ever-changing external factors such as seasonal tree shading, snow coverage, ground reflectance, and panel soiling also impact power output. If these data are available, there are well established physical models that utilize these inputs and current or forecasted weather to estimate solar generation [10]. These tools are often utilized for estimating the output of large solar arrays and have achieved accuracy to within $\pm 10\%$ of the actual generation of the array [11,12].

However, as noted in [13], distribution utilities often do not have access to this specific data for each residential array, either because the data were not collected during the permitting process or because the *as installed* configuration of the system does not match the reported configuration. In addition, system or site modifications over time may also impact PV production.

Shaker et al. [14] suggest using the output from a small number of known PV arrays to extrapolate power production across a greater number of PV systems within the same area. However, this type of modeling requires the utility to know the total capacity of residential systems within the area, and implicitly assumes that the small number of monitored arrays represent all variation in system configurations for the unmonitored arrays.

Several authors have proposed machine learning methods to estimate PV production [15]. These methods rely on training a model using separately metered PV and load on residences [16], specific system/site data [17], or use PV production from nearby metered arrays as a surrogate for residential PV production [13].

The key challenges with these proposed methods include *a)* acquiring long time series of local generation data for training, *b)* understanding if the solar resource used for training is representative of current and future PV installations, *c)* acquiring the skills to train the model - likely a challenge for most distribution utilities, and *d)* depending upon the method, having reasonably accurate records of the location, size and *as-built* physical characteristics of individual PV arrays.

Many of the models described above were developed and tested using only small data sets, including a single 5-kW array [16], 8 houses over the span of 10 days [7], 15 days of training data from one PV system [18], or only superficially describe training methods and data sets [17]. Another method requires metered PV output from over 300 sites [19], an arduous data set to replicate. These approaches report an root mean squared errors of between 20% and 50%, and usually require access to data from representative systems where PV is metered independently from load. Because these works use uncommonly available data sets and typically only train and/or model from a few PV systems over short periods of time, further work to improve net-metered PV modeling is appropriate.

The difficulty of modeling a portfolio of residential assets using physical models stimulated interest in estimating PV output based on smart meter (AMI) data. This type of estimate is essentially an effort to disaggregate PV generation from the net-generation/net-load measurements reported by the smart meter. This work proposes an alternative method that will estimate PV production using open-source python software and three data sources that are commonly accessible by distribution utilities: *a)* historical and, if desired, forecasted, weather

data, *b)* net-metered, sub-hourly, AMI data, and *c)* the approximate month of PV system installation. Contact the corresponding author for software examples.

This method does not require training using known PV generation, nor does it attempt to estimate array-specific parameters [20] or a representative system's geometry, as attempted by other work [21]. We use this method to estimate individual system and aggregated generation using 1–4 years of AMI data from nearly 850 net-metered residential systems. By using the known basic configurations of these systems, the method's performance is tested against a physical model to compare generation estimates.

The paper is organized as follows: Section 2 discusses data sources, Sections 3 and 4 discuss the methodology and testing of the AMI model against known generation, and Section 5 provides an overview of the physical model used for model comparisons. Sections 6 and 7 present results, discussion and conclusions. Throughout this work, we designate the mean of any quantity using a bar (e.g. $\bar{P}$), and the median using a caret (e.g. $\hat{P}$).

2. Description of data sources

For this analysis, we utilize data from Fort Collins, Colorado, USA, a high-elevation community in the central USA with highly variable weather conditions. Electricity service is provided by the City of Fort Collins Utilities (FCU), a city-owned municipal distribution company with a sole-provider contract with Platte River Power Authority.

2.1. AMI data from Fort Collins utilities

Beginning in 2014, FCU installed AMI meters on existing premises, and the AMI system became operational at the beginning of 2015 [22]. FCU had approximately 76,000 premise meters in mid-2019.

The AMI meters report on 15-min intervals, and all residential PV systems are net-metered. FCU provided all AMI data to Colorado State University (CSU) researchers - a total of approximately 11 billion meter readings. Four data fields are of interest for this research: premise ID number, which can be tied to geospatial and housing data, timestamp, delivered energy, and received energy. Delivered energy refers to energy supplied to the building by the distribution grid during a 15-min interval. Received energy refers to self-generated energy (in this case from PV) exported by the building back to the distribution grid during a time interval.

The power balance for a single premise is:

$$d_i = l_i - g_i \quad (1)$$

where d_i is the power delivered to the premise at any instant i , l_i is the premise load, and g_i is the PV output power. During one meter reporting period, the meter samples d many times, and d may be both positive - energy delivered to the premise - or negative - generation received by the utility. The meter integrates these individual measurements over the meter's reporting period and reports two quantities to the utility, delivered energy:

$$\Delta_t = \sum_{i \in t} \max(d_i, 0) \cdot dt \quad (2)$$

and received energy:

$$\Pi_t = \sum_{i \in t} |\min(d_i, 0)| \cdot dt \quad (3)$$

where the subscript t refers to one meter reporting period - 15 min in this study - and dt refers to the (unknown) internal sampling period of the meter. Energy quantities for load ($P_l = \sum l_i \cdot dt$) and generation (P_g) are similarly defined.

Fig. 1 schematically illustrates the recorded AMI data. Solid boxes

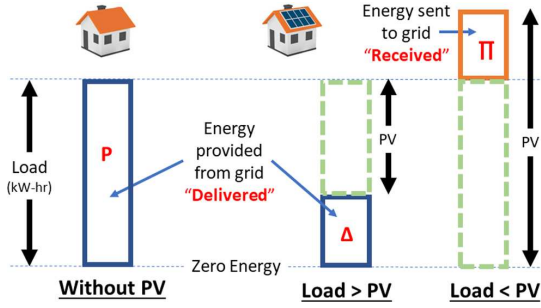


Fig. 1. AMI Data Visualization. Box P represents the delivered energy to the premise before PV installation, which is the total building load when no generation assets are present. Box Δ represents energy delivered to the premise when some load is offset by unmeasured PV generation (dashed green boxes). Box Π represents excess generated energy received by the distribution grid from the premise when generation exceeds load.

represent metered readings and dotted boxes represent unknown values.

2.2. Weather data

Weather data from a weather station located on the main campus of CSU [23] – a reasonably central location to FCU's distribution system – was used for this study. The weather station provides 10-min measurements of temperature ($^{\circ}\text{C}$), wind speed (m/s), and solar irradiance (watts/m^2). Fort Collins snowfall data was obtained from the National Oceanic and Atmospheric Administration's Climate Data online portal [24]. The database provided daily snowfall (inches) and accumulated snowfall (inches).

2.3. Solar assets in Fort Collins

FCU provided records of each photovoltaic system installed in the city's distribution network. This dataset provided the following information: premise ID, installed capacity (kW_{DC}), date of installation, inverter manufacturer, inverter model and quantity, array azimuth and tilt, and the electric service rate class of the premise.

Between 1987 and April 2019, 1,389 PV systems were installed within the City, totaling a capacity of 13.1 MW_{DC} . Approximately 93% of PV systems were installed on residential premises, with an average installed capacity of 5.5 kW_{DC} . Residential premises accounted for 7.3 MW_{DC} , or approximately 55% of total installed capacity. Around 285 older systems, mostly installed before 2015 are missing array tilt and/or azimuth values in the City's records; these values were estimated using satellite and street-level imagery.

2.4. Confidentiality

This analysis utilized the full data set including per-premise AMI readings. This type of data is available to most utilities but is seldom made publicly available. Interested researchers can contact the corresponding author for aggregated and/or anonymized data similar to that used here.

3. AMI model methodology

3.1. Framework

The underlying assumption behind this model is that, given similar weather conditions, building load (i.e. l_i) does not change due to PV system installation. Using this assumption, estimates of load from pre-

PV periods can be used to estimate behind-the-meter PV generation for post-PV periods.

We find this assumption works well for the study area. However, some caution should be used when applying to other systems. Some recent behavioral research has observed an increase in energy usage (termed the 'solar rebound effect') after PV is installed in a household [25]. These observations were made in a region with high cooling/low heating building energy needs (the opposite of Fort Collins), and also found that solar consumers in liberal areas (like Fort Collins) experience lower rebound.

Using this assumption and referring to the notation in Fig. 1, the basic approximation from this model for PV generation, G , is $G = P - \Delta + \Pi$ for any time period. The values Δ and Π are measured and recorded in the AMI data. The model's algorithm matches similar timestamps from pre- and post-PV periods to estimate P , the total load for the building. While applied here to PV production, reconstructing this baseline building consumption from net-metered AMI data can be utilized to estimate demand changes for a wide range of changes in customer behavior, and represents the core contribution this work.

Weather impacts both building energy demand and solar generation. Buildings exhibit temperature-dependent demand driven by heating or cooling [26] – periods of similar temperature should have similar heating/cooling loads [27]. Solar irradiance is also correlated with temperature and is the input energy for PV. For these reasons, temperature and global horizontal irradiance (GHI) are considered in the pairing of timestamps for reconstructing building load. Using this guide, we pair every timestamp t , with a set of comparable timestamps, Θ_t , i.e. each period t maps to a set of comparable timestamps, $\Theta_t = \{\Theta_t\}$. These sets are the same for all premises.

As each premise, k , has a unique date of PV installation, time stamps are split into before- and after- PV installation sets. Denoting the installation time as $t_{k,I}$, β includes all time stamps before PV installation, i.e. where $t < t_{k,I}$ and α includes all time stamps after installation, i.e. $t > t_{k,I}$. We also avoid time stamps immediately before or after the installation date, to allow for differences between the actual and the reported dates to activate the PV system. Using a buffer of b days before, and a days after (in this work $b = a = 20$ days), the definitions of β and α are defined for each premise, k , as

$$\beta_k = \{t: t \in t < (t_{k,I} - b)\} \quad (4)$$

For each $\beta_{k,i} \in \beta_k$ we identify the set of comparable timestamps after installation:

$$\Theta_{k,i}^{(\beta)} = \{\Theta_t: t \in \beta_{k,i}\} \quad (5)$$

Similarly the set of times *after* PV install:

$$\alpha_k = \{t: t \in t > (t_{k,I} + a)\} \quad (6)$$

have a similar mapping,

$$\Theta_{k,i}^{(\alpha)} = \{\Theta_t: t \in \alpha_{k,i}\} \quad (7)$$

where the parenthetical superscripts indicate *all time stamps in the indicated set* for premise k . These form two 2D data structures, dictionaries where each time stamp in α_k and β_k holds a set of comparable timestamps, $\Theta_{k,i}$.

Weather information is used to create Θ_t for all timestamps, t , (Section 3.2 below) and then this set is divided into subsets to populate the dictionaries as described in Eqn. (4)–(7). Similarly, two AMI readings (received and delivered) are also attached to each time stamp in each set for each premise. Finally, the algorithm estimates building consumption by comparing AMI data for one time stamp with the set of AMI data identified by the before- or after- installation comparable timestamps. Once building consumption is estimated, PV generation can be estimated. Fig. 2 below depicts the block diagram of this methodology.

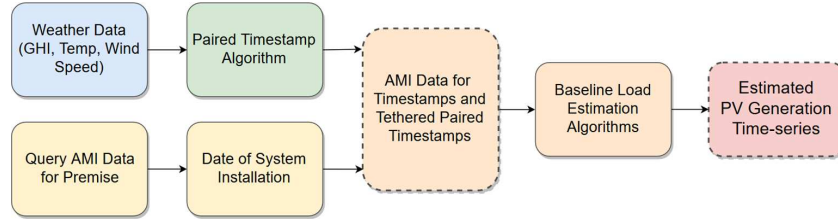


Fig. 2. Block diagram of AMI model.

3.2. Paired timestamps

After interpolating the 10-min weather data to 15-min granularity, each timestamp has measured, average, GHI, wind speed, and temperature for the 15-min interval. For computational efficiency, PV modeling is performed only for periods when the sun is at least one degree above the horizon, as determined by standard solar positioning algorithms [28].

Comparable timestamps must be within 15 days of the day-of-year of the target timestamp. This creates a 30-day range to identify comparable timestamps, and emphasizes days that are similar in seasonal weather, with similar day length, shading, and solar patterns.

From this 30-day subset, the following requirements are added:

1. Comparable timestamps must match the same portion of the week - weekday or weekend - as the target timestamp.
2. Major American holidays are removed since much of the population deviates from daily routines on these dates and likely make any premise's energy usage non-comparable to other non-holiday timestamps with similar weather/temporal conditions. Removed holidays are: New Year's, Memorial, Independence, and Labor Days; Thanksgiving, Black Friday, Christmas Eve & Day, and Boxing Day (day after Christmas).
3. Comparable timestamps must be within ± 4 h of the target timestamp's hour-of-day.
4. The standard deviation of the temperature (σ_{temp}) and irradiance (σ_{irr}) of this remaining set are calculated and timestamps within 30% of σ_{temp} and 40% of σ_{irr} are included in the set. This criteria removes weather outliers from the sets.

The specific parameters described above were developed qualitatively. Differences in load between weekend, weekday, and holiday periods are well documented in the literature, and utilized here as well.

The remaining criteria were set qualitatively by comparing model accuracy to known PV generation sources mentioned discussed in Section 4. The 4-h window limits comparison to similar parts of the day. Tests indicated that ranges larger than ± 4 h increased error between the model and recorded data. While wider windows may select similar weather, solar irradiation angle and other model conditions change markedly over longer periods and negatively impact model performance. The time range and the 30% and 40% temperature and irradiance standard deviations were selected qualitatively by iterating the model and observing its performance. Similar tuning could be completed in other localities, if necessary.

The period of analysis from 1/1/2015 to 3/31/2019 contains over 72,000 timestamps. Upon completion of the comparable timestamp algorithm, the median and mean number of similar timestamps for each target timestamp was 118 and 129, respectively, with a range of 0 to 460.

3.3. Calculating baseline energy load

The baseline equation for estimated energy generation in kilowatt-hours (kWh) is:

$$G_{k,t} = P_{k,t} - \Delta_{k,t} + \Pi_{k,t} \quad (8)$$

in reference to Fig. 1, where $G_{k,t}$ and $P_{k,t}$ are the unknown generation and building load, respectively, and $\Delta_{k,t}$ and $\Pi_{k,t}$ are the known delivered energy and received energy, from premise k for each timestamp, t .

Before PV is installed, there is no generation, thus $\Delta^{(\beta)} = P^{(\beta)}$ and $\Pi^{(\beta)} = 0$. $\Pi^{(\beta)}$ should always be zero before PV was installed because there is no generation at the premise; a non-zero value is an error, typically an incorrectly recorded date of installation or a malfunctioning smart meter.

After PV is installed, either or both $\Delta_{k,t}^{(\alpha)}$ or $\Pi_{k,t}^{(\alpha)}$ could be non-zero for a timestamp with non-zero load; both can be nonzero if the reporting period saw two-way power flow during the timestamp.

The fundamental assumption behind the algorithm is that the load for any timestamp *after* PV is installed will be similar to the load for comparable timestamps *before* PV was installed – i.e. $P_{k,t}^{(\theta_{\alpha})} \approx P_{k,t}^{(\alpha)}$ for each timestamp in α . In practice, the algorithm uses the mean ($\bar{P}_{k,t}$) or median ($\hat{P}_{k,t}$) of $P_{k,t}^{(\theta_{\alpha})}$ as a central estimate of the load for any given timestamp. Additionally, if an individual timestamp in α_k was tethered to a $\Theta^{(\alpha_{k,t})}$ set of fewer than three timestamps, the central estimate of the prior timestamp is used.

Application of Eq. 8 must account for several special conditions. Typically, overall premise demand drops after a premise has installed PV. However, for short durations (such as 15-min timestamp periods) consumption may increase, leading to a negative first term in Eq. (8), and potentially an artificially low or even negative estimate for PV generation. To understand this irregularity, an analysis identified the timestamps where delivered energy ($\Delta_{k,t}$) exceeded estimated building load ($P_{k,t}$). Of calculations from 846 residential premises spanning 1–3.25 years, this irregular condition occurs, on average, in 8.3% of timestamps.

To control for these irregularities, the following corrective actions were taken, in order, stopping when the irregularity was handled, i.e. when an equation was satisfied. Starting from an assumption of estimated energy output of 0 kWh, and dropping the subscripts for convenience, generation is estimated to be:

$$G = \begin{cases} \hat{P} - \Delta + \Pi & : \hat{P} > \Delta \\ \bar{P} - \Delta + \Pi & : \bar{P} > \Delta \\ \hat{P} - \hat{P}^{(\alpha)} + \Pi & : \hat{P} > \hat{P}^{(\alpha)} \\ \Pi & : \Pi > 0 \end{cases} \quad (9)$$

Because mean values are influenced by outliers to a greater degree than the median value, the median and mean will usually differ within P . Because of outlier influence, $\hat{P}$ is the preferred measure of building load, with $\bar{P}$ used in cases in which $\hat{P}$ is less than Δ (in Eq. (9)). When both $\hat{P}$ and $\bar{P}$ are less than Δ (thus not satisfying the first two requirements in the equation set above), $\hat{P}^{(\alpha)}$ (the median delivered energy from the after-install readings) is used. Finally, a last resort calculation is to solely use the received (Π) value at the timestamp as estimated generation.

4. Comparison to actual generation

Recorded generation from four net-metered PV systems was used to

Table 1

Actual generation data sources.

Site Name	Capacity (kW _{DC})	Premise Type	Granularity	Duration
House G	5.04	Residential	Daily	13 months
House Z	6.00	Residential	Hourly	2 months
Timberline	20.59	Non-residential	Hourly	8 months
Church	53.5	Non-residential	Daily	12 months

validate the accuracy and robustness of the AMI model. Table 1 details these net-metered systems.

These premises are a diverse and representative mix of net-metered solar assets, spanning a range of installed capacity and building types, with different usages and load profiles. Residential buildings typically have two daily peaks; a small morning peak and a more substantial peak in late afternoon and early evening. They also exhibit different energy needs during the work week and on weekends. The Timberline Recycling Facility is a city-owned building that is open 360 days a year and is served by a general service net-metered electricity connection. The load profile at this premise is likely be characterized by consistent energy usage during business hours every day of the week, with major fluctuations driven by HVAC needs. In contrast, the energy needs of the Church are likely low during the week and vary on weekends depending on the different services and one-off events that the site hosts.

We use absolute percentage error, ε_s , to compare modeled to actual output. Error on each timestamp is:

$$\varepsilon_s = \frac{G_s - A_s}{A_s} \quad (10)$$

where G_s is the AMI model's generation, and A_s is the actual recorded generation, for each time period s . We compute $\bar{\varepsilon}_k$ as the mean of ε_s and $\hat{\varepsilon}_k$ as the median of ε_s over the set of $s \in \alpha_k$, or for daily and hourly subtotals taken from α_k .

Time-series of modeled and actual generation from three of the four premises are shown in Figs. 3 (House G), 4 (House Z) and 5 (Timberline). Comparison is hourly for Timberline and House Z, and daily for House G and the Church, which were only reported daily. The residential test cases were characterized by daily $\bar{\varepsilon}$ between -4% to -26% , meaning that the AMI model underestimates generation relative to actual recordings. Hourly generation was not available for House G, but for House Z the hourly $\hat{\varepsilon}$ was -26.5% .

Estimations agreed better over periods longer than daily intervals. Over the course of one year at House G, the total actual generation was 6,570 kWh while the AMI model estimated 6,492 kWh, a difference of -1.2% . For the two months analyzed for House Z, the actual generation (885 kWh) was higher than the AMI model estimate (673 kWh), a difference of -23.8% . However, it is known that the occupant of House Z, an energy engineer affiliated with this research, was also aggressively shifting and minimizing loads for time-of-day pricing (TOD) periods

that started coincident with the PV install at this premise. Therefore, the two impacts are difficult to separate. This represents a limitation to the model, but also indicates that the model can be used for purposes other than PV output estimation, such as estimating the impact of TOD pricing on customer behavior. Finally, weather during winter months has higher variability and present more challenges to the model when estimating short time periods for an individual premise (installed Feb. 2019).

The model performed best at Timberline at both hourly and daily resolution, as shown in Fig. 5. This suggests that consumption patterns are very consistent at this municipally-operated commercial site and did not diverge after PV system commissioning. Over the course of the eight months analyzed, actual generation was 16,607 kWh, while the AMI model estimated 16,735 kWh, a difference of $+0.8\%$.

5. Physical model used for validation

FCU does not have access to actual generation data for the nearly 1,300 residential PV systems in Fort Collins, but does have basic array characteristics recorded when the systems were installed. This type of data may be unavailable to many distribution utilities, due to record keeping processes or other policies. A key strength of the proposed model is that it can estimate PV production *without* this data. However, for the purposes of this study, access to physical model data allows a bulk comparison of the proposed model to physical models to build confidence in the approach.

The physical model is based upon PV performance models developed by Sandia National Laboratory [12] and the PV Performance Modeling Collaborative (PVMC) [29], as well as NOAA [30], and NREL open-source online resources [31]. The model, "PV-STEM" [28], is specifically designed to utilize inputs that are readily available to utilities: Array nameplate capacity, tilt and azimuth of the array, and, possibly, inverter make and model to lookup inverter efficiency. While more complex models utilize several components of solar radiation (typically direct normal and diffuse horizontal irradiance), PV-STEM uses only GHI, which is readily accessible from many weather stations.

To assess its accuracy, PV-STEM was tested using the recorded data described previously, and several longer recordings on systems that were not connected to AMI [28]. Using as a reference a commercial array with three years of data – a similar duration to the AMI data set used here – results indicate the model is accurate to within $\pm 6\%$ for average annual production, to within $[+6\%/-15\%]$ for daily aggregations, and to within $\pm 35\%$ for hourly aggregations.

While the physical model is based upon recorded system characteristics, it is subjected to similar errors as other physical models, including inaccurate system characteristics, differences between *as designed* and *as installed* systems, and post-installation changes, such as shading changes, soiling, or aging effects.

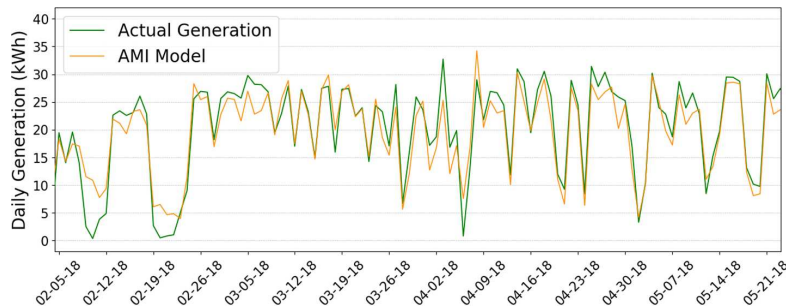


Fig. 3. House G Daily Generation – 3 months of estimated daily generation aligned with recorded generation. Daily $\bar{\varepsilon} = -3.9\%$, i.e. the AMI model underestimated daily generation by around 4%.

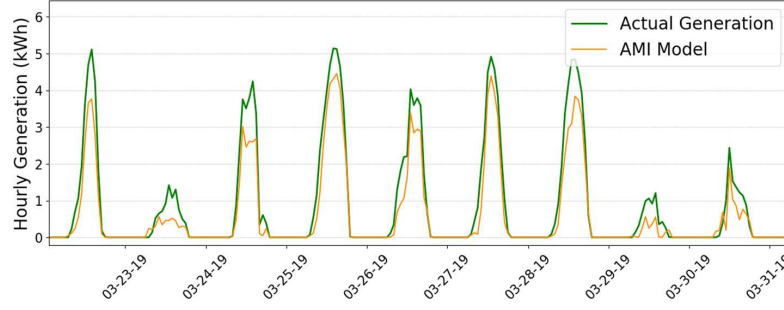


Fig. 4. One Week of Hourly Generation Comparison for House Z – hourly modeled and recorded hourly PV generation. The model underestimates during both sunny and cloudy (days 2,8,9) periods $\hat{\epsilon} = -26.5\%$ at hourly, and $\hat{\epsilon} = -25.8\%$ at daily, granularity.

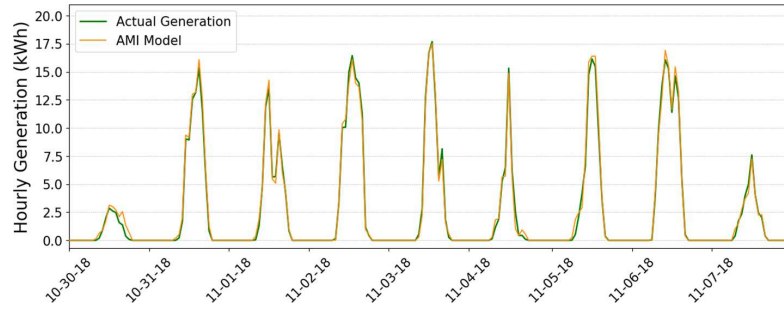


Fig. 5. Timberline Hourly Generation – An hourly $\hat{\epsilon} = 1.7\%$, a daily $\hat{\epsilon} = 0.4\%$.

6. Results

Since both the AMI and PV-STEM models are production estimates there is no baseline on which to calculate relative error. Therefore comparison is based upon the difference in the estimates normalized by the mean magnitude of the estimates:

$$r_s = 2 \left(\frac{G_s - M_s}{G_s + M_s} \right) \quad (11)$$

where s is an element of α_k , G_s is the AMI model's generation estimate, and M_s is the PV-STEM model's generation estimate, during s . Differences aggregated over a specified duration (such as hourly, daily, monthly) are presented as the mean ($\bar{r}_k$) and median ($\hat{r}_k$) for premise k . A positive r_s equates to the AMI estimate being greater than the PV-STEM estimate. All premises were modeled through March 31, 2019. Fig. 6 provides an example of PV-STEM and AMI models plotted together, and is a representative example of the generally close agreement seen for most systems; ($G_s - M_s$) is the difference between the two lines

in the figure at each point.

6.1. Monthly aggregation

For individual premises, a useful modeling duration for analysis is “monthly generation”. Most utility customers in the U.S. receive a monthly bill with a total kWh of consumed energy during approximately the previous 30 days, and the Energy Information Administration reports monthly electricity usage statistics [32]. This kWh total is converted to a dollar amount through a single or multi-tiered billing structure. A residential customer may want to compare their actual monthly generation with predicted monthly generation to understand system performance and expectations for the payback period of the array. The AMI model joins with other physical models as a tool to draw conclusions about expected monthly performance.

For insight into the monthly agreement, the monthly average error, $\bar{r}_k$, was calculated from the all available individual monthly errors for each premise, and a 90% empirical confidence interval (CI) was

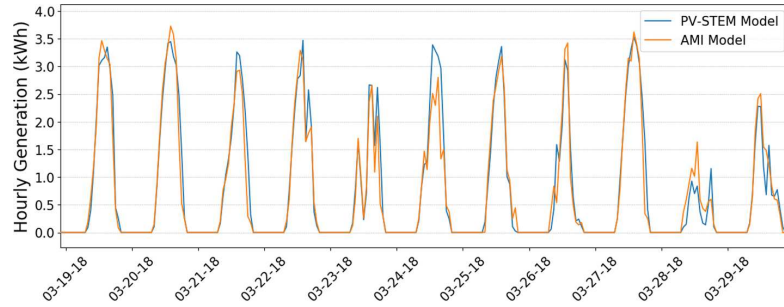


Fig. 6. Example of agreement between PV-STEM and AMI models. 11 days of hourly data for residential premise with a 3.9-kW array installed in February 2017.

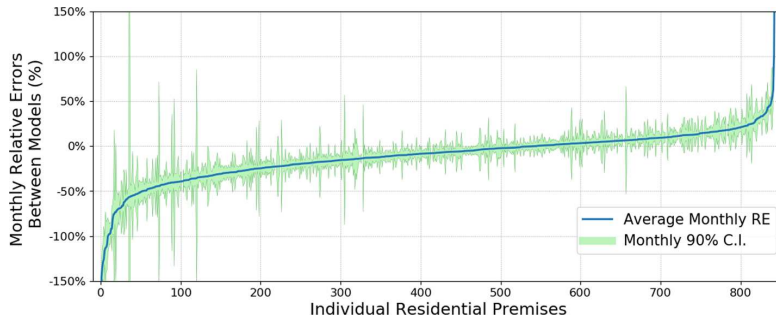


Fig. 7. Average Monthly Error Between Models. The x-axis is ordered by ascending average monthly error. Positive error represents the AMI model estimating more generation than the PV-STEM model. The green shaded area is the 90% empirical confidence interval of monthly error for each premises. 846 residential premises are evaluated.

calculated for the distribution of monthly errors. Fig. 7 shows results for all 861 premises – 536 premises had negative monthly $\bar{r}$ while 325 premises each had positive monthly $\bar{r}$, indicating a slight tendency for the PV-STEM model to estimate more generation than the AMI model.

The figure indicates broad agreement across the majority of premises, while a smaller number of premises, intentionally placed at each end of the curve, indicate large relative disagreements for a minority of premises. These larger disagreements may be attributed to several factors. First, the PV-STEM model mis-estimates output on numerous systems because system information, such as recorded array characteristics (size, tilt, and azimuth) are incorrect. Further, changes to the system such as array modifications or upgrades may not have been recorded in the FCU's records. Using satellite photography and some street level observations, the study team confirmed that numerous systems with large disagreements had an installed panel configuration that did not match the data recorded in FCU records.

Second, PV-STEM and other physical models only partially account for snow coverage reductions. While physical models readily model the reduced insolation occurring during snowfall, these models have difficulty modeling the reduction in PV output after snow events when solar insolation is high but the panels are covered by accumulated snow. These modeling reductions are complex and irregular across events. While a general reduction has been built into the PV-STEM model (and most other physical models) these reductions have only limited sensitivity to different snow and array conditions.

Finally, physical models typically have little-to-no knowledge of array shading from buildings or trees, which may also change over time. Accounting for variable shading at hundreds of premises is not feasible for most utilities. Both the snow- and shading-reductions cause PV-STEM to overestimate production, while errors in physical parameters may cause over- or under-estimation of production.

In contrast, the AMI model uses comparisons of building consumption, which intrinsically captures reduced generation from structural factors – incorrect physical parameters or shading – and provides some insight into the snow reductions. Using these differences, a utility that has physical parameters for all installed systems could identify and update flawed data by comparing physical and AMI estimates.

6.2. System-wide residential aggregation

Distribution operators are primarily interested in the aggregate PV production from residential systems. This section discusses the relative agreement of the two models aggregated across all premises at daily and hourly granularity. Data is presented monthly and color-coded to highlight seasonal variation. Data from January, February, and March 2016 are not shown because these months had fewer than 30 premises, the threshold set to produce credible statistics on a monthly basis. In comparison, months in 2018 include over 500 premises.

Fig. 8 summarizes daily and hourly system-wide aggregations. While this analysis is retroactive – looking at historical values – the analysis provides insights into how the proposed method could be

coupled with weather forecasts for planning utility operations. In that context, daily values are primarily of interest for day-ahead planning and hourly analysis provides insights for intra-day operations.

Daily Aggregation: In every month of analysis, the two models agreed to within $\pm 20\%$ for at least 50% of days in the month, and in 10 months, at least 50% of days agreed to within $\pm 10\%$. Further, in 21 of 36 months the 90% CI – a measure of the variation between daily estimates – lies within $\pm 20\%$ relative difference. Critically, 9 of the 10 highest producing months were within this group, the months likely of most interest to a distribution utility. Overall, the agreement between the proposed model and accepted physical models builds confidence in the proposed approach to PV estimation.

Even at good precision, on most days PV-STEM estimates higher generation than AMI – 25 months contained at least 75% of days when PV-STEM estimated greater generation than the AMI model. Among these 25 months were the 12 most-productive solar months. The red line of total relative error represents the system-wide sum of all AMI estimated generation relative to the sum of all PV-STEM estimated generation. The total relative error is (-7.6%), meaning that total generation across applicable premises was skewed toward the PV-STEM model estimating more generation than the AMI model.

Winter months show less agreement between the two models than other seasons – 8 of 12 winter months have a CI that exceeds $\pm 20\%$ and, in the 14 the months where the CI exceeded $\pm 20\%$, eight were winter months, four were in the fall or spring, and two were summer months. Winter months also include the periods with the lowest PV production – 8 of the 12 lowest producing months are winter months. Possible reasons for model disagreement and PV-STEM overestimation are explored in the next subsection.

Hourly Aggregation: The smaller period for hourly aggregations produces greater disagreement between models (as is typical for statistical estimates at higher temporal resolution). Hourly aggregations (lower panel, Fig. 8 display larger disagreements in the median and wider CIs than daily aggregations (top panel, Fig. 8). However, the median disagreement and wider CIs indicate different disagreements: Quartiles are skewed positive, indicating that the AMI model's hourly estimates are higher than the PV-STEM model's hourly estimates when differences are largest. In contrast, all months have a negative hourly medians, meaning that the central estimate of each month tends indicates PV-STEM overestimating relative to the AMI model. This dichotomy is due to higher absolute errors when the AMI model is overestimating relative to PV-STEM (positive values) coupled with a higher frequency of smaller overestimates by PV-STEM relative to the AMI model (negative values). Months such as March 2017 and September 2016 have clear visual representations of this occurrence.

Explanations for this type of disagreements could include: (a) the physical model consistently overestimates on some fraction of the premises, due to shading, incorrect system parameters or similar issues which “feed” the physical model incorrect assumptions; (b) episodic reductions in PV production, such snow on PV panels, that are captured by the AMI model, but not by the physical model; (c) small but

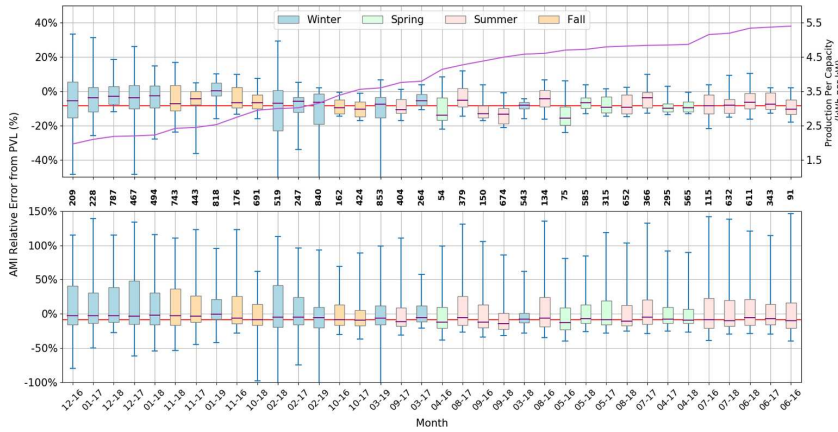


Fig. 8. Comparison of the physical and AMI models. The top plot shows daily relative errors between the two models and the bottom plot contains hourly relative errors; both are grouped by calendar month. The y-axis shows the relative error (r in %). Positive errors indicate a higher estimate from the proposed AMI model relative to the physical (PV-STEM) model. The boxes show the 25th, median, and 75th quartile of each month's set of relative errors. The bold numbers between the boxes are the number of premises available for analysis during that month. The right y-axis shows the estimate of energy production, normalized by installed capacity during the month. The boxes are ordered left to right from least to highest normalized monthly production. The red horizontal line represents the total relative error between the models for all data included in the chart.

persistent changes in customer load behavior are not well represented by the fundamental assumptions in the AMI model.

6.3. Snow losses

The inherent nature of the AMI model positions it to be well-suited at predicting generation reduction due to snow coverage. Unlike physical PV models, the AMI data describes what is happening within the building's load, and is not restricted to the behavior of the PV system. Net-load/generation is inclusive of panel snow coverage and other losses, by-passing the various physical model external variables that must be factored in (often with high difficulty).

Because the AMI model utilizes similarities in building energy load before and after PV installation, a lack of generation from snow is captured within the AMI data. Referring to Eq. (8), if snow is covering the array and no generation is offsetting building consumption, the delivered energy (Δ) should then be similar to the delivered energy from the comparable periods before the premise had PV (P). The difference in these terms would be small; adding the near-zero grid-received (Π) energy would result in a low, and reasonably accurate, estimate for PV generation. Fig. 9 below displays model performance in the days during and following a heavy snow event.

The AMI model does not use precipitation in the timestamp matching process. When rain or snow occurs the irradiance will significantly decrease, thus matching periods of precipitation with other high-cloudiness/precipitation periods.

The AMI model is responsive to the effects of snow coverage during and after the snowfall event. While by no means perfect, the AMI model's structure produces generation estimates that are not "fooled" by high irradiance periods while an array lies under snow. Thus the

proposed model provides a competent alternative to prior snowfall modeling approaches discussed in [33–35].

6.4. Limitations

By design, the proposed model relies on only AMI and weather data - two data streams that are readily accessible to any utility. However, the total reliance on AMI data requires that a premise be excluded from analysis if it lacks 12 months of pre-installation data. Therefore, older systems installed before or during the first year of AMI data collection (2015), or systems on newly constructed buildings cannot be analyzed without extensions to the proposed method. Possibilities include using post-installation data to develop a proxy for pre-installation data or using AMI data from similar premises. These methods were not investigated for this study and could potentially required substantial additional software or human control.

Uncertainty is also derived from sources external to the model. The model operates under the assumption that energy consumption habits do not significantly change once PV is installed. While this work identified no significant behavioral changes, some social science research suggests that this assumption may fail in some cases.

Changes in the premise itself (e.g. efficiency upgrades or changes in appliances) or use of the premise (e.g. different owners or different activities onsite) also impact the load patterns of the premise. Further, the utility is unlikely to know about these events. For these changes, model accuracy is assisted by three factors. First, PV system output (≥ 5 kW on average) is large, relative to other, common, daytime performance changes. For example, EV charging is of similar size (1–6 kW typical), but is more likely to occur during lower nighttime electricity rates, when vehicles are at home. Many other efficiency changes are

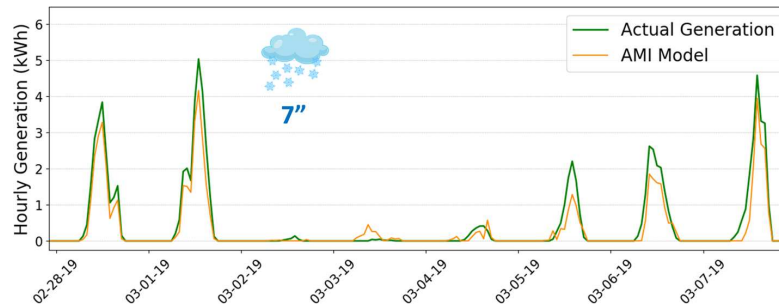


Fig. 9. AMI model performance at House Z during snow event. On March 2, 2019, Fort Collins received over 7 inches of snow.

typically small (<1 kW) relative to PV system output, split between daytime and nighttime load, and often preferentially occur during the evening peak load. Second, aggregated across a distribution system, changes in premise ownership produce relatively unbiased aggregate impacts on load (as is demonstrated by the performance results presented earlier, which do not control for premise ownership.) For example, new owners may increase or decrease load, and aggregated across the distribution system, these changes are likely to cancel out. Finally, with the exception of large changes, such as the purchase of an EV, efficiency upgrades tend to happen incrementally over time, as appliances are added or replaced in a household. This continuous, incremental change, is swamped by the sudden, daytime-only, introduction of PV production to a household. However, it is possible that widespread changes, such as a widely-adopted efficiency upgrade occurring in a short period of time, could produce biased changes which would impact the accuracy of the proposed model's results.

Finally, changes in the utility operation itself may have the largest – and intended – impact on load patterns. In October 2018, FCU implemented time-of-use electricity rates in Fort Collins, which included a greater than 3:1 differential between on- and off-peak electricity pricing. For systems installed near the time of the TOD implementation, the rate structure change significantly impacts the fundamental assumption in the AMI model – i.e. that consumption does not change after PV is installed. While too little data exists at the time of this study, early evidence suggests that daily aggregation of estimated generation from AMI data is more resilient to the effects of inter-hour pricing than hourly modeling. Understanding the behavioral impact of this price change provides numerous other areas of future work with the City of Fort Collins.

7. Discussion and conclusion

The use of AMI data provides another method for estimating PV production in distribution systems. This method requires minimal data about the PV systems in question (approximate installation date) and readily accessible AMI and weather data. As a result, it represents a robust estimate of PV production that is independent of physical models driven by either specific system information or estimated system-wide system information. Comparing estimates provides a method of identifying problems and raising the confidence of all estimates.

High-confidence estimates of renewable generation are needed in the delicate and dynamic balancing of load with intermittent and dispatchable generation by grid operators. Estimating PV generation from thousands of small, behind-the-meter systems, is an important and non-trivial task using existing physical models or algorithms that require actual generation data. Thus, as rooftop solar deployment increases, the ability to disaggregate invisible generation and total load from net-metered AMI data will be necessary for better management of the distribution of electricity.

The “matching timestamp” methodology proposed here allows comparisons of between sets of load data by correcting for weather-driven and time-driven loads, which often account for a large portion of load variation within one premise. As a result, this method can be used to analyze any intervention expected to have an impact on load behavior, such as rate changes or efficiency incentive programs. The caveat to these applications is that the load changes must be large enough, and/or the sample size (number of premises) large enough, to make load profile changes statistically significant. Regarding sample size, modern AMI systems provide large quantities of data – an ongoing analysis of FCU's time-of-day rates has access to over 18,000 residential premises, even after eliminating those with occupancy changes, premises with PV, premises with electric heat, and premises that have participated in efficiency or demand response programs. Utilizing AMI data for these purposes extracts additional, largely hidden, value from these crucial systems.

CRedit authorship contribution statement

Wendell Stainsby: Software, Formal analysis, Writing - original draft, Investigation. **Daniel Zimmerle:** Funding acquisition, Investigation, Supervision, Writing - review & editing. **Gerald P. Duggan:** Software, Supervision, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank the City of Fort Collins Utilities for the large amount of time and data that they provided to our team. John Bleem was instrumental in the development of the PV-STEM model. This work is supported by the National Science Foundation Sustainability Research Networks award number 1444745 “Integrated Urban Infrastructure Solutions for Environmentally Sustainable, Healthy, and Livable Cities”.

References

- [1] Mayes F. More than half of small-scale PV generation comes from residential rooftops. U.S. Energy Information Administration; 2016.
- [2] Wood Mackenzie Power & Renewables. U.S. Solar market insight, executive summary: 2018 year in review; March 2019.
- [3] U.S. Energy Information Administration. Electric power monthly with data for february 2019; April 2019.
- [4] Solar Energy Industries Association. United States surpasses 2 million solar installations; May 9, 2019. <https://www.seia.org/news/united-states-surpasses-2-million-solar-installations>.
- [5] Roselund C. Room for residential solar to reach 41 GW by 2025.
- [6] Gagnon P, Margolis R, Melius J, Phillips C, Elmore R. Rooftop solar photovoltaic technical potential in the United States. A detailed assessment. National Renewable Energy Laboratory; 2016.
- [7] Tabone M, Kiliccote S, Kara EC. Disaggregating solar generation behind individual meters in real time. In: Proceedings of the 5th conference on systems for built environments, BuildSys '18. New York, NY, USA: ACM; 2018, p. 43–52.
- [8] Cleveland M, Andersen G. State net metering policies. <http://www.ncsl.org/research/energy/net-metering-policy-overview-and-state-legislative-updates.aspx>.
- [9] Tarroja B, Mueller F, Eichman JD, Samuels S. Metrics for evaluating the impacts of intermittent renewable generation on utility load-balancing. *Energy* 2012;42(1):546–62.
- [10] Dobos AP. PVWatts version 5 manual, report. National Renewable Energy Laboratory; 2014.
- [11] Menicucci D. PVFORM: A new approach to photovoltaic system performance modeling. In: Presented at the 18th IEEE photovoltaic specialists conference, Las Vegas, Nevada; 21 Oct. 1985.
- [12] King DL, Boyson WE, Kratochvill JA. Photovoltaic array performance model; December 2004.
- [13] Iyengar S, Sharma N, Irwin D, Shenoy P, Ramamritham K. A cloud-based black-box solar predictor for smart homes. *ACM Trans Cyber-Phys Syst* 2017;1(4):21:1–21:24.
- [14] Shaker H, Zareipour H, Wood D. Estimating power generation of invisible solar sites using publicly available data. *IEEE Trans Smart Grid* 2016;7(5):2456–65.
- [15] Chen D, Irwin D. Black-box solar performance modeling: comparing physical, machine learning, and hybrid approaches. *SIGMETRICS Perform Eval Rev* 2017;45(2):79–84.
- [16] Yang H, Huang C, Huang Y, Pai Y. A weather-based hybrid method for 1-day ahead hourly forecasting of PV power output. *IEEE Trans Sustainable Energy* 2014;5(3):917–26.
- [17] Mohan R, Cheng T, Gupta A, Garud V. Solar energy disaggregation using whole-house consumption signals; 2014.
- [18] Chow S, Lee E, Li D. Short-term prediction of photovoltaic energy generation by intelligent approach. *Energy Build* 2012;55:660–7.
- [19] DeBenedictis A, Hoff T, Price S, Woo C. Statistically adjusted engineering (SAE) modeling of metered roof-top photovoltaic (PV) output: california evidence. *Energy* 2010;35(10):4178–83.
- [20] Chen D, Irwin D. SunDance: black-box behind-the-meter solar disaggregation. Proceedings of the Eighth International Conference on Future Energy Systems, e-Energy '17. 2017. p. 45–55.
- [21] Wang Y, Zhang N, Chen Q, Kirschen DS, Li P, Xia Q. Data-driven probabilistic net load forecasting with high penetration of behind-the-meter pv. *IEEE Trans Power Syst* 2018;33(3):3255–64.
- [22] U.S. Department of Energy. Recovery act selection for smart grid investment grant awards; May 22, 2019. https://www.energy.gov/sites/prod/files/edg/media/SGISelections_Category.pdf.

- [23] Colorado State University. Fort Collins weather station; May 22, 2019. <https://ccc.atmos.colostate.edu/autowx/>.
- [24] NOAA National Centers for Environmental Information. Climate data online; June 3, 2019. <https://www.ncdc.noaa.gov/cdo-web/>.
- [25] Qiu Y, Kahn ME, Xing B. Quantifying the rebound effects of residential solar panel adoption. *J. Environ. Econ. Manage.* 2019;96:310–41.
- [26] Haas R, Auer H, Biermayr P. The impact of consumer behavior on residential energy demand for space heating. *Energy Build* 1998;27(2):195–205.
- [27] Yuan S, Stainsby W, Li M, Xu K, Waite M, Zimmerle D, et al. Future energy scenarios with distributed technology options for residential city blocks in three climate regions of the United States. *Appl Energy* 2019;237:60–9.
- [28] Bleem JR, Duggan GP, Stainsby W, Zimmerle D. PV Modeling as a community resource. In: Proceedings – American solar energy conference; August 2019.
- [29] Stein JS, Holmgren WF, Forbess J, Hansen CW. Pvlib: Open source photovoltaic performance modeling functions for matlab and python. In: 2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC); 2016, p. 3425–30.
- [30] NOAA Global Monitoring Division. General solar position calculations; 2019.
- [31] Reda I, Andreas A. Solar position algorithm for solar radiation applications. *Sol Energy* 2004;76(5):577–89.
- [32] U.S. Energy Information Administration. Electric sales, revenue, and average price; October 12, 2018.
- [33] Marion B, Schaefer R, Caine H, Sanchez G. Measured and modeled photovoltaic system energy losses from snow for Colorado and Wisconsin locations. *Sol Energy* 2013;97:112–21.
- [34] Brench BL. Snow-covering effects on the power output of solar photovoltaic arrays, report. U.S. Department of Energy; December 1979.
- [35] Freeman JM, Ryberg DS. Integration, validation, and application of a PV snow coverage model in SAM, report. National Renewable Energy Laboratory; August 2017.

A.2 Big Data Analytics for Power Distribution Systems using AMI and Open Source Tools

Big Data Analytics for Power Distribution Systems using AMI and Open Source Tools

Gerald P. Duggan, Daniel Zimmerle

Energy Institute

Colorado State University

Fort Collins, CO 80524

Email: jerry.duggan@colostate.edu, dan.zimmerle@colostate.edu

Sonu Upadhyay

Smart Grid and Systems Operations

Fort Collins Utilities

Fort Collins, CO 80521

Email: supadhyay@fcgov.com

Abstract—Advanced meters now account for almost half of the electric meters in the United States. In addition to reducing the costs of meter reading, these meters provide a vast increase in the amount of consumption data available to the utility. Many organizations, however, lack the expertise to analyze this data, so it languishes unused. This paper describes steps taken to make AMI data usable for day-to-day planning tasks at a medium-sized distribution utility. The proposed approach can be implemented on commodity hardware, providing a cost effective and secure platform for advanced analysis tasks.

Index Terms—Advanced Metering Infrastructure, Big Data Analytics, Power Distribution Systems, Power Flow

I. INTRODUCTION

Historically, load data has only been readily available to distribution utilities at the head-end of feeders, where recording meters were typically installed for billing and load monitoring. Since the acquisition of load data elsewhere in the feeder has been labor-intensive, analysis of distribution systems historically scaled head-end measurements across all transformers on a feeder, with little modeling of the variation in loads across a feeder. For example, power flow analysis is often performed using only the peak load day / hour, with loads allocated to transformers based on peak load collected at the feeder head-end or a limited number of points along a feeder. Lack of location-specific data produces an incomplete understanding of the system behavior, with little insight into the impacts of changing customer behavior such as the impact of distributed renewable energy or electric vehicle adoption.

The deployment of Advanced Metering Infrastructure (AMI) provides a significant improvement in the amount of data available to utilities. The Federal Energy Regulatory Commission (FERC) estimates that, as of 2016, approximately half the electric meters in the United States are smart meters [1]. Because of this, AMI is an active research topic. Prior studies have provided an overview of the features and requirements for AMI data management [2]–[4] usage of AMI data to develop load profiles [5], [6], using AMI data for load forecasting and demand response [7], [8], and dynamic state estimation using AMI data [9], [10]. In contrast, this study focuses on making AMI data available for routine analysis and decisions support by a broad spectrum of utility personnel.

While AMI data *should* provide necessary data for location-specific understanding of distribution systems and operations,

access to these data is often difficult. Many utilities have contracted with third party information management companies to collect, store and analyze AMI data. Data required for analysis must then be requested through service contracts and may incur extra charges and/or take significant time to acquire. In contrast, work performed at Colorado State University (CSU) illustrates that AMI data can be readily stored and accessed locally utilizing widely available and inexpensive tools. Once accessible to utility staff, data can be readily queried to answer advanced questions using location-specific inputs.

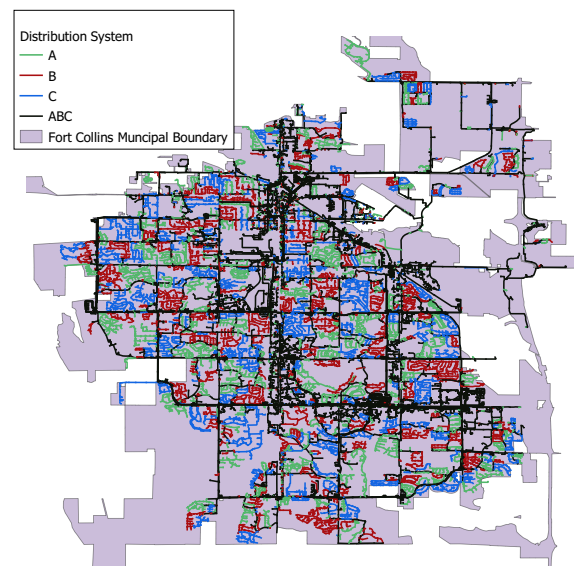


Fig. 1: Fort Collins Distribution System. Single and three phase lines are shown, color coded by phase.

Since 2016, Fort Collins Utilities (FCU) has provided both electrical system and AMI data to CSU. This data has grown to include over 4.5 years of 15-minute interval AMI data, electrical models for all feeders, and location of all distributed generation. Common coding with county property records allows connection of FCU customer premises to local property records, which provides geospatial-locations for most data and

allows socio-economic data, such as U.S. Census data, to be linked to meter data. While the size of this data appears daunting, the work described in this paper utilized hardware and software well within the budget of any small to medium sized utility. The server computer costs <\$10k, and open source software was used for major software components: Postgres database, QGIS for mapping, OpenDSS for power flow analysis, and Python for scripting.

This paper discusses methods to store AMI and ancillary data in open-source database tools, and demonstrates the power of the storage system using a case study – a demonstration of automated power flow analysis using AMI data for location-specific loading. Section II describes the relevant data sets and details how to upload and curate the data. We focus primarily on the large volume of AMI data, which must be structured to support both ad-hoc and systematic queries. Section III details the case study. Finally, Section IV describes current research activities at CSU based on this data set, and provides suggestions for future directions.

II. DATA DESCRIPTION

Fort Collins, a small city in northeastern Colorado, is located ≈ 60 miles north of Denver and has a population of 165,000 people and an area of 57 square miles (Figure 1). FCU Light and Power is a municipal distribution utility that provides electric service to over 70,000 homes and businesses. A high-level view of the data provided from FCU to CSU is given in Figure 2. Elements of this data model are:

- **System Topology**, detailed information for the electrical assets (lines, transformers, breakers, etc.), necessary configuration data, and cabling specifications between them. All data is geospatially located. The data contains sufficient detail to map assets, such as Figure 1, and to generate power flow models.
- **Feeder Current Data**, hourly average current readings taken at the feeder head-end, typically at the substation. Feeder-level data readings provide the aggregated current for the feeder; substation-level data readings provide the aggregated current and power flow figures for all feeders serviced at the substation.
- **AMI** tables provide average power delivered to, and received from, individual premise meters at 15-minute intervals. There are four complete years of this data for a total of over 10.5 billion records. Only real power is reported for residential premises; real power and apparent power are reported for commercial customers..
- **Property Records** from the county property record database provide data on each parcel, including assessed value, number of rooms, house stories, etc.
- **Census Block Group Data**, available from the U.S. Census Bureau, provide demographic information, including breakdowns by age, race, and income.
- **Transformer to Premise** and **Premise to Parcel** mapping tables provide links between the system topology data and the AMI data (indexed by premise) and Property and Census data (indexed by parcel number).

The data provided to CSU is considered sensitive, as it contains Personally Identifiable Information (PII) and data that can be considered Critical Energy Infrastructure Information (CEII) [11]. As such, it is covered by a Non-Disclosure Agreement (NDA) between the two entities. Terms of the agreement include restricting data access to approved personnel, physical security of the server storing the data, and transfer of the data from FCU to CSU through secure communication mechanisms only.

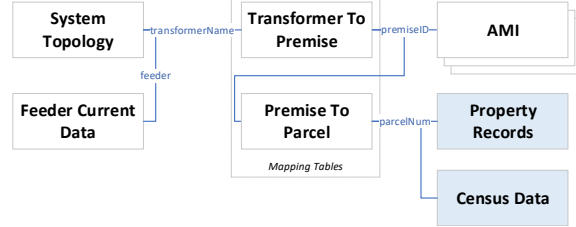


Fig. 2: Data model overview. Blocks represent major data components. White backgrounds indicate data from FCU, shaded backgrounds represent data from external sources.

The data transfer process from FCU to CSU has three distinct steps: raw data upload, data aggregation, and data validation.

A. Raw Data Upload

Data transfer between FCU and CSU exchanges Comma Separated Variable (CSV) files via secure FTP. CSV files are used because they are a ubiquitous format, easily produced and parsed. The process on the FCU side is automated, requiring only manual initiation, in order to guarantee data consistency. System Topology files change infrequently and are uploaded every six months, while AMI data files are uploaded monthly.

At CSU, Python scripts import the CSV files into a Postgres database. These scripts break the system topology files into a core table which describes the elements that are electrically connected and individual tables that contain data specific to the element types (e.g. cable lengths). An example of the core topology table is shown in Table I.

TABLE I: Sample Topology Data

Node	Node Type	Parent	Node Data
S1	SUBSTATION	-	...
L1	UNDERGROUNDLINE	S1	...
T1	TRANSFORMER	L1	...

Postgres supports partitioning large tables, where a logical master table describes the data layout, and one or more of the columns are used to specify physical tables which actually store data. These physical tables are created and stored separately, which aids data upload and maintenance. Queries are made to the master table, and Postgres automatically forwards the query to subtable(s) that overlap the range defined in the query.

The AMI raw data files are broken into separate files for residential and commercial premises. These files are large – 10 GB per month for residential data and 0.5 GB per month for commercial data. Because of the size of these files, the Postgres tables for AMI readings are partitioned by month (see Figure 3). Python code was written to stream the CSV file contents to the DB partition subtable, as opposed to reading the file into memory all at once. The partition scheme keeps monthly data uploads independent of each other, and the streaming implementation keeps computer resource usage (in particular, RAM) to a minimum. This enables multiple AMI data files to be uploaded in parallel. Once both the residential and commercial CSV files are uploaded for a particular month, indexes are created on both the premise ID and timestamps.

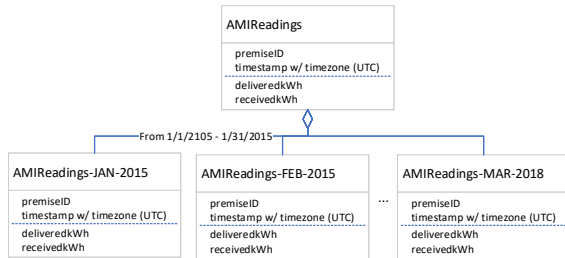


Fig. 3: AMI table partitioning. Individual months are loaded into separate physical tables (bottom of figure), and then mapped into a logical table for unified queries.

One individual table of AMI readings, corresponding to one month of data for both residential and commercial premises, consumes about 22 GB of disk space, including space for indices. To enhance retrieval speed, the entire AMI database is stored on Solid State Disks. At the time of the writing of this paper, a 2 TB enterprise quality SSD costs \$1300, which translates to a cost of \$14 for storing one month of AMI data.

B. Data Aggregation

The basic structure of the AMI readings table is one reading per premise at 15-minute intervals. Researchers often require this data aggregated either by time or location (transformer or feeder). Data aggregation creates summary tables, grouping by convenient time intervals – 15-minute, 1 hour, and 1 day – and by location – individual premise, transformer, and feeder.

C. Data Validation

Data from FCU is not scrubbed and requires data validation processes to assure each table is internally consistent and properly related to other tables. Two mechanisms are used. The first is to use intrinsic properties of the data to check consistency. An example, the nominal distribution topology is radial – i.e. a tree structure with the substation connection as the root node – and can be checked using standard tree traversal algorithms: Every node is referenced in some path from the root node and no paths are cyclic. The second

technique for data validation is to check the contents and keys of one table versus another.

The system topology files provided by FCU are the files used by the utilities as input into their analysis software. At the start of the project, CSU had licenses to the same software. Initial validation of the models was then performed by both entities running models against the same load specifications and ensuring identical results. As the models on the CSU side were stored in the database first, then output in the software's input format, this ensured that the "round-trip" path through the database introduced no errors. The database was then used to generate OpenDSS models, which were again validated by producing the same results for the same load specifications.

The model verification process spot-checks AMI data. In addition, the authors validated the AMI data by comparing it against monthly billing statements – data which are available on-line. While this provides a first-order check for errors, additional inconsistencies and anomalies have been found through regular use. The team is now implementing a systematic data checking process. Identified errors will be corrected in the AMI system itself, therefore improving future data quality for all uses. This type of automated consistency checking is one advantage of routinely accessing AMI data for internal research purposes.

III. CASE STUDY – POWER FLOW ANALYSIS

In the spring of 2019, FCU performed a study to identify and detail the distribution system impacts as a result of Photo-Voltaic (PV) penetration. Three feeders with varied topology were selected for the study. The first feeder did not have any distributed generation (DG) assets; the second feeder had the minimum amount of existing DG; and the third feeder had the maximum amount of DG, exceeding limits set by a previous study. Specific loads used for each feeder were the minimum load days during daylight hours to record the impact of added solar generation, and the peak solar generation day (as determined by output from a municipal solar farm) for the baseline for solar generation.

The FCU has defined conditions required to grant a permit for connecting PV resources to their distribution grid. These conditions, listed below, were evaluated in this study[12]:

- Bus voltage deviation due to PV generation of no more than $\pm 5\%$ from nominal.
- PV generation does not cause reverse power flow at the substation feeder breaker.
- Impact on the feeder when PV generation exceeds 13.3% of the feeder's load rating.
- Impact on the feeder when total PV generation exceeds 50% of the existing annual minimum load on that feeder.

Methodology for the study was to import the feeder topology into a commercially available utility analysis program, import the appropriate load for the scenario, and perform a power flow analysis. Then the days and percentage of solar generation were varied to generate results for analysis. Each variation was evaluated with respect to the interconnect conditions.

In total, this study took a two weeks to perform, and included 3 feeders and 4 days of AMI data. A total of about 1000 scenario variations were run and examined, and the results were used to create new criteria for the PV hosting capacity on the feeders. Given practice, performing the study again has been estimated to take a week.

A. Bulk Power Flow Analysis

As an example of data-drive system analysis, we performed a time series power flow analysis for one of the feeders for a full year using the full AMI data set to load the secondary sides of all transformers. The process was:

- 1) Feeder topologies were exported one time in OpenDSS format ('Feeder' in Table II), creating a single work unit.
- 2) A time-stamped load profile was exported for each transformer on the feeder ('Loads' in Table II), creating $8760 \frac{\text{hours}}{\text{year}} \times 4 \frac{\text{meas}}{\text{hr}}$ for 35,040 work units.
- 3) OpenDSS was run for each time stamp, using one feeder topology and loads for each time stamp. Relevant results are exported from OpenDSS and summarized for each run ('Scenario' in Table II), creating 35,040 result sets.
- 4) Results across all runs were summarized ('Summary' in Table II).

The results of the power flow analysis are given in Figure 4. The figure shows the values for 2018 were well within $\pm 5\%$. However, it is important to note the manual analysis *would not have identified* the unusually low minimum voltages in October 2018. Further investigation of these timestamps could offer insights into system operation and/or future problems.

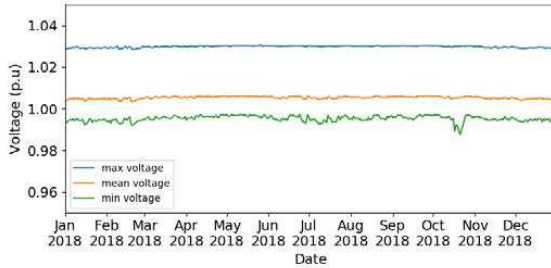


Fig. 4: Daily power flow results for one feeder. Mean, maximum and minimum p.u. voltages were extracted for all timestamps for each day, and are plotted by calendar day.

B. Parallelization

The database queries and power flow analyses described above require little communication between processes, making them amenable for simple coarse-grained parallelization. Once the phases in subsection III-A were broken into work units, the Python *multiprocessing* module was used to create process pools, and each work unit was dispatched by a controller to an individual process in the pool. That process would run the work item to completion, then take on another. For example, inputs to the 'Scenario' work items were the feeder plus a

timestamped load file. The coordinator would dispatch the first load file, say 1/1/2018 00:00:00 to the first process, the next load (1/1/2018 00:15:00) to the second, etc. As each process completes, the coordinator dispatches another work item to the process, continuing until all work items are complete. This simple form of parallelization is readily implemented; more sophisticated parallelization schemes could offer additional opportunities for performance improvement.

TABLE II: Performance – See III-A for description of Phases

Phase	Runtime (s)	Scaled				
	1 Proc.	1	2	5	10	20
Feeder	3	1	0.98	1.01	1.00	1.47
Loads	15772	1	0.51	0.21	0.12	0.12
Scenario	14760	1	0.50	0.20	0.12	0.10
Summary	72	1	0.66	0.54	0.65	0.90
Total	30708	1	0.50	0.21	0.12	0.12
Ideal	1	1	0.50	0.20	0.10	0.05

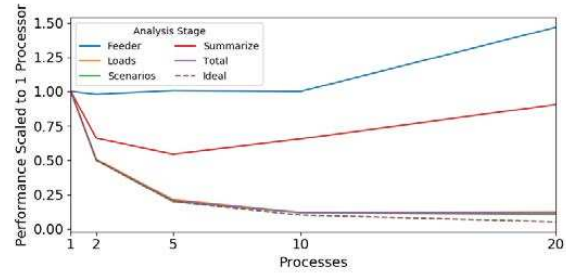


Fig. 5: Performance Summary

Table II shows timing results for the power flow analysis described above. Hardware used was a dual-processor 2.6 GHz system with 8 cores per processor, 32 GB RAM, 2 TB SSD (database), 1.2 TB fast hard drive (OS), and 6 TB slow hard drive (bulk data). Running all work items took more than 30,000 seconds (or 8.3 hours) on a single process. The table also shows the performance for running on 1, 2, 5, 10, and 20 processes, with the values scaled to the single process case. To get an indication of how well each task is suited to parallelization, the theoretical maximum ("ideal") speedup is shown for each process count. Figure 5 shows a graphical representation of the relative speedup vs. the number of processes. These results show that the optimal number of processes to allocate to running the power flow analysis on the particular server used is 10. The actual total runtime for 10 processes is 3,674 seconds – slightly over 1 hour – indicating that this analysis could be completed for all FCU feeders over a weekend, on one computer, utilizing *actual* loads for an entire year.

To tie back to the case study, FCU performed about 1000 scenarios for its PV distribution system impact study, looking at 4 feeders and 3 production days, using an estimated week

of engineering time each time. Given the techniques described here, those same 1000 scenarios could be run in fewer than two minutes. Alternatively, *all feeders* could be analyzed on a regular basis using 2-3 days of run time. By setting common-sense thresholds, the post processing could flag a small number of specific scenarios (feeder/day/etc.) requiring additional analysis.

IV. CONCLUSIONS

In this study, we make the argument that bringing the wealth of knowledge embedded in AMI data *inside* utility operations does not require excessive computing power, staff time, or complex software. An intelligently designed data schema using standard – indeed open source – data management tools supports fast, efficient, queries through billions of AMI records. We also show that this work can be done on standard, single-box, multi-core, server hardware that is well within the budget and IT management capabilities of small-to-medium utilities. Given these investments, questions can then be answered utilizing *actual* data, rather than historical approximations that neglect the heterogeneity inherent in modern utilities, where increasing distributed generation and electric vehicle loads are rapidly changing circuit loadings, impacting customer programs, and potentially stressing equipment.

Ongoing, work using these data have shown the value of tying AMI data into maintenance and operational analysis. Ongoing investigations include:

- Developing and verifying privacy preserving methods to anonymize AMI data, allowing it to be used more widely within the utility, and potentially for externally-visible uses which improve customer and community engagement.
- Automating power flow analyses to cover all feeders in Fort Collins (currently more than 80) and to include scenarios with more aggressive PV and electric vehicle penetrations.
- Automating techniques to systematically clean and verify AMI data, and to verify simulation results.

ACKNOWLEDGEMENT

This work is supported by the NSF (SRN: Integrated Urban Infrastructure Solutions for Environmentally Sustainable, Healthy and Livable Cities, award 1444745).

REFERENCES

- [1] (). Nearly half of all U.S. electricity customers have smart meters - Today in Energy - U.S. Energy Information Administration (EIA), [Online]. Available: <https://www.eia.gov/todayinenergy/detail.php?id=34012> (visited on 08/05/2019).
- [2] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, "Big data analytics in power distribution systems," in *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Feb. 2015, pp. 1–5. DOI: 10.1109/ISGT.2015.7131868.
- [3] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," in *2013 IEEE International Workshop on Intelligent Energy Systems (IWIES)*, Nov. 2013, pp. 40–45. DOI: 10.1109/IWIES.2013.6698559.
- [4] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big Data management in smart grid: Concepts, requirements and implementation," *Journal of Big Data*, vol. 4, no. 1, p. 13, Apr. 28, 2017, ISSN: 2196-1115. DOI: 10.1186/s40537-017-0070-y. [Online]. Available: <https://doi.org/10.1186/s40537-017-0070-y> (visited on 10/13/2019).
- [5] K.-l. Zhou, S.-l. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, Aug. 1, 2013, ISSN: 1364-0321. DOI: 10.1016/j.rser.2013.03.023. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032113001822> (visited on 10/13/2019).
- [6] O. Ardakanian, N. Koochakzadeh, and R. P. Singh, "Computing Electricity Consumption Profiles from Household Smart Meter Data," p. 8,
- [7] F. L. Quilumba, W. Lee, H. Huang, D. Y. Wang, and R. Szabados, "An overview of AMI data preprocessing to enhance the performance of load forecasting," in *2014 IEEE Industry Application Society Annual Meeting*, Oct. 2014, pp. 1–7. DOI: 10.1109/IAS.2014.6978369.
- [8] P. Siano, "Demand response and smart grids—A survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, Feb. 1, 2014, ISSN: 1364-0321. DOI: 10.1016/j.rser.2013.10.022. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032113007211> (visited on 10/13/2019).
- [9] M. Baran and T. E. McDermott, "Distribution system state estimation using AMI data," in *2009 IEEE/PES Power Systems Conference and Exposition*, Mar. 2009, pp. 1–3. DOI: 10.1109/PSCE.2009.4840257.
- [10] J. Peppanen, J. Grimaldo, M. J. Reno, S. Grijalva, and R. G. Harley, "Increasing distribution system model accuracy with extensive deployment of smart meters," in *2014 IEEE PES General Meeting — Conference Exposition*, Jul. 2014, pp. 1–5. DOI: 10.1109/PESGM.2014.6939840.
- [11] (). FERC: Critical Energy/Electric Infrastructure Information (CEII), [Online]. Available: <https://www.ferc.gov/legal/ceii-foia/ceii.asp> (visited on 10/12/2019).
- [12] City of Fort Collins Utility Services. (). Interconnection Standards for Generating Facilities (GF) Connected to the Fort Collins Distribution System, [Online]. Available: https://www.fcgov.com/utilities/img/site_specific/uploads/interconnection_standards.pdf (visited on 12/29/2017).

Appendix B

Mechanistic Emissions Modeling

B.1 Methane Emissions from the Natural Gas Transmission and Storage System in the United States



Methane Emissions from the Natural Gas Transmission and Storage System in the United States

Daniel J. Zimmerle,^{*,†} Laurie L. Williams,[‡] Timothy L. Vaughn,[†] Casey Quinn,[†] R. Subramanian,[§] Gerald P. Duggan,[†] Bryan Willson,[†] Jean D. Opsomer,^{||} Anthony J. Marchese,[†] David M. Martinez,[†] and Allen L. Robinson[§]

[†]Energy Institute and Department of Mechanical Engineering, Colorado State University, Fort Collins, Colorado 80524, United States

[‡]Department of Physics and Engineering, Fort Lewis College, Durango, Colorado 81301, United States

[§]Center for Atmospheric Particle Studies (CAPS) and the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States

^{||}Department of Statistics, Colorado State University, Fort Collins, Colorado 80524, United States

S Supporting Information

ABSTRACT: The recent growth in production and utilization of natural gas offers potential climate benefits, but those benefits depend on lifecycle emissions of methane, the primary component of natural gas and a potent greenhouse gas. This study estimates methane emissions from the transmission and storage (T&S) sector of the United States natural gas industry using new data collected during 2012, including 2,292 onsite measurements, additional emissions data from 677 facilities and activity data from 922 facilities. The largest emission sources were fugitive emissions from certain compressor-related equipment and “super-emitter” facilities. We estimate total methane emissions from the T&S sector at 1,503 [1,220 to 1,950] Gg/yr (95% confidence interval) compared to the 2012 Environmental Protection Agency’s Greenhouse Gas Inventory (GHGI) estimate of 2,071 [1,680 to 2,690] Gg/yr. While the overlap in confidence intervals indicates that the difference is not statistically significant, this is the result of several significant, but offsetting, factors. Factors which reduce the study estimate include a lower estimated facility count, a shift away from engines toward lower-emitting turbine and electric compressor drivers, and reductions in the usage of gas-driven pneumatic devices. Factors that increase the study estimate relative to the GHGI include updated emission rates in certain emission categories and explicit treatment of skewed emissions at both component and facility levels. For T&S stations that are required to report to the EPA’s Greenhouse Gas Reporting Program (GHGRP), this study estimates total emissions to be 260% [215% to 330%] of the reportable emissions for these stations, primarily due to the inclusion of emission sources that are not reported under the GHGRP rules, updated emission factors, and super-emitter emissions.



■ INTRODUCTION

The recent dramatic growth in production and utilization of natural gas creates potential economic, climate, energy security, and public health benefits. However, climate benefits depend critically on lifecycle emissions of methane, the primary component of natural gas and a potent greenhouse gas.¹ Methane emissions from the United States natural gas system have been estimated using both top-down and bottom-up approaches. Currently, the most complete bottom-up inventory of natural gas system emissions is the 2012 U.S. Environmental Protection Agency’s (EPA) Greenhouse Gas Inventory (GHGI).² It estimates an overall system-wide methane loss rate of 6.2 Tg/yr of methane, or 1.3% of methane transported (Supporting Information (SI), Section 16) and that the transmission and storage sector (T&S) is the largest emitting

segment of the natural gas industry,² accounting for about one-third of methane emissions.

Recent top down studies using aircraft-, tower-, ground-based sensing and atmospheric transport models^{3–7} suggest that the GHGI may underestimate methane emissions from the oil and gas industry. However, attribution to particular sites or even sectors can be challenging for top-down atmospheric studies, while bottom-up estimates such as the GHGI are limited by sample size, possible sample bias, and measurement limitations. Additionally, Brandt et al.⁸ argue that bottom-up inventories

Received: April 2, 2015

Revised: July 1, 2015

Accepted: July 2, 2015

Published: July 21, 2015



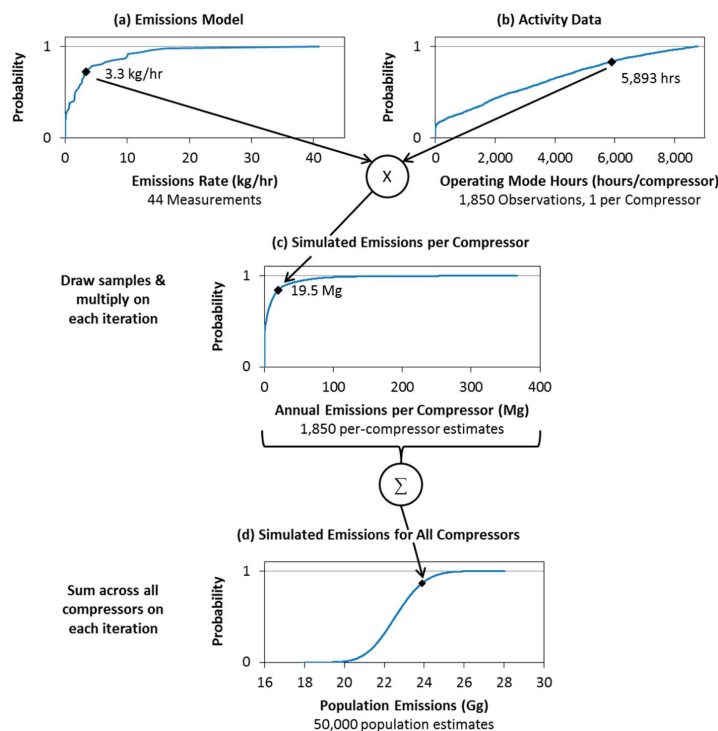


Figure 1. Schematic of an emission submodel (rod packing vent emissions for reciprocating compressors in operating mode for partner facilities) to illustrate the overall modeling approach. This model is based on cumulative distribution functions (a) containing 44 emission rate measurements from the field measurement campaign and (b) showing operating mode hours, from partner data, for the 1,850 reciprocating compressors in this portion of the model. On each Monte Carlo iteration, an emission rate is drawn from (a) for each compressor and multiplied by the operating hours for that compressor, as shown by the arrows, and to produce (c), a distribution of emissions per compressor. Following each of the 50,000 Monte Carlo iterations, population emissions are computed by summing simulated emissions for all compressors to produce a distribution of emissions as shown in (d).

consistently underestimate methane emissions and that a small number of “super-emitters” could be responsible for a large fraction of the emissions. Recent measurement studies have also found significantly skewed distributions for component-level emissions, including detailed surveys of exploration and production well sites,⁹ pneumatic controllers,^{10,11} liquid unloadings,¹² and distribution facilities.¹³ Additional studies, using local atmospheric measurements, have identified super-emitters at both well sites¹⁴ and gathering facilities.¹⁵

Quantifying the contribution of large emission sources at the component or facility level requires substantial data to define both the proportion of components or facilities with high emissions and the magnitude of those emissions. Existing emissions inventories (e.g., EPA GHGI) are often based on relatively small data sets, e.g., the GHGI for T&S is based on emissions data from only 15 compressor stations, less than 1% of the total station population. Recently, large new emissions and activity data sets have become available for T&S, including field measurements¹⁶ and extensive data from the EPA’s Greenhouse Gas Reporting Program (GHGRP),¹⁷ but these data have not yet been incorporated into national estimates like the GHGI.

This study defines the T&S sector as the natural gas system infrastructure between two custody transfer points—the receipt meter between the gathering and processing sector and the transmission system, and the delivery meter between the transmission system and the distribution system (commonly called a “city gate”) or industrial end user (SI, Section 3). The T&S sector contains both interstate and intrastate infrastructure, including compressor stations, pressurized pipeline networks, metering and regulation stations, and supporting equipment. This study estimates emissions from compressor stations associated with the T&S sector and connected to underground storage fields. These facilities can be equipped with gas scrubbers, separators, after-coolers, dehydrators, storage tanks, gas-driven pneumatic devices, valves, and connectors.

The GHGI² estimate for the T&S sector is largely based on activity and emissions data collected in the early 1990s. For example, many of the emission factors are from the 1996 study by the Gas Research Institute,¹⁸ with updates of centrifugal compressor emissions and activity from a 2010 study by ICF.¹⁹ GHGI estimates transmission station and compressor counts based upon ratios of stations and compressors to pipeline miles from 1992 data, the baseline year, which are then scaled using

current pipeline miles.²⁰ There have been substantial changes to the United States natural gas system over the past decade due to the development of shale gas and other factors, raising questions about the representativeness of the data sets underlying the GHGI.

This study develops a new estimate of T&S methane emissions based upon new data for emissions, equipment census, and operations. Results are compared to the GHGI and the GHGRP. For T&S, stations which emit more than 25,000 t of CO₂ equivalent annually are required to report to the GHGRP.

METHOD

This study utilized probabilistic emission and activity models and Monte Carlo methods, as in Ross,²¹ to estimate emissions and associated uncertainty. These methods propagate the variability of input emissions and activity data through to a distribution of methane emission estimates that we term the “study model estimate” (SME) for the T&S sector.

The emission modeling approach is described in Figure 1, using rod packing emissions for reciprocating compressors in operating mode as an example. Each emission category was modeled using empirical cumulative distribution functions (CDFs) constructed with emissions and activity data obtained in a field measurement campaign,¹⁶ data from six partner companies who are major T&S operators, and public data from the GHGRP program. In total, the model is based on new emissions data for 677 facilities and activity data from 922 facilities (SI, Table S6-b), representing more than one-third of all United States T&S facilities. None of these data, to our knowledge, have been used in previous national estimates. Emissions categories are based on those defined by the GHGRP and include compressor isolation and blowdown valves, seal vents, and rod packing vents; pneumatic devices; component leaks; and tanks (SI, Section 4). Facilities are divided into multiple groups for simulation (SI, Section 6). For each emission category in each group of facilities, each Monte Carlo iteration provides a single emission estimate obtained by drawing an emissions rate(s) from the appropriate emission CDF(s) and combining it—typically multiplying—with the appropriate activity data (e.g., number of compressors). This process was repeated 50,000 times, and these iterations were combined to develop a distribution of emissions estimates for each category for each group of facilities. These estimates were then combined using Monte Carlo methods to develop a distribution of estimates for the entire sector.

Emissions Models. Seventeen emissions models (in addition to super-emitter and exhaust models, below) were developed from 2,292 new onsite measurements of individual emission sources. Of these, 1,279 measurements were collected during a six month field measurement campaign at 45 T&S facilities operated by six partner companies,¹⁶ and the remaining 1,013 measurements were made during 2012 by three of the partner companies as part of their greenhouse gas reporting activities, using instruments and protocols aligned with the project measurement protocols described by Subramanian et al.¹⁶ Emission categories are detailed in the SI, Section 4, and the origin of all measurements is documented in CDFMaster.xlsx in the SI zipped file.

A detailed analysis of the field measurement study data can be found in a recent paper published by the authors,¹⁶ but key results are included here. A crucial question is whether onsite measurements accurately represent facility emissions. To

address this, the field study simultaneously measured methane emissions using two established, independent techniques — comprehensive onsite measurements and downwind tracer flux—to verify that the comprehensive onsite surveys captured all major sources. In addition to measuring sources included in the GHGRP, the onsite protocol also required measurements in all compressor operating modes (e.g., “not operating pressurized”) and sources (e.g., dry seal vents) that are not included in the GHGRP, restricted the type of measurement instruments that could be utilized, and used engineering estimation for unmeasured emission sources. Tracer flux methods estimate the facility level emission rate by comparing the downwind concentrations of methane to those of tracer gases released from the facility at a known rate.²²

For 38 facilities with paired onsite and tracer flux measurements, onsite and tracer flux measurements agreed to within the experimental uncertainty. The agreement between these two independent methods creates confidence that the onsite measurements captured most methane emissions and therefore allows emissions at T&S facilities to be modeled in detail using the category-specific activity and emission data. Category-specific modeling supports detailed understanding of the drivers behind emissions, and differences between this study, the GHGI, and the GHGRP. Two facilities with paired onsite and tracer flux measurements were identified by the tracer flux technique as having emissions greater than 200 standard cubic feet per minute (SCFM) and are treated as super-emitters in the model (see below). Standard conditions are defined as 60 °F and 1 atm.

The equipment-level emission data are highly skewed. Figure 2 presents cumulative distribution functions (CDFs) of the emissions data for the 17 emission models developed from field study and partner data (SI, Table S7-b). A long tail is readily evident in most models. In 12 of the 17 models, the maximum measured emission rate is at least double the 97.5% fractile of the distribution. The inset plot shows the fraction of aggregate measured emissions in each model attributable to the largest

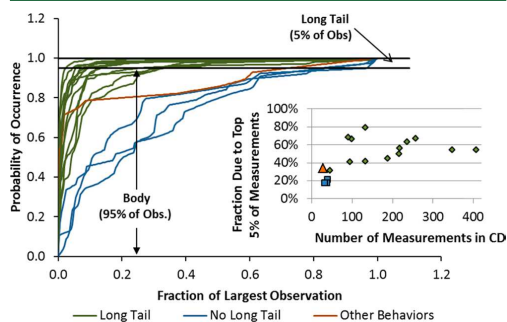


Figure 2. Normalized cumulative distributions of the 17 emission models developed from data collected in this study (SI, Section 7). Inset illustrates the fraction of all measured emissions due to the largest 5% of measurements. Green lines/symbols indicate models with substantial “long tail” behavior, where 5% of measurements are responsible for >30% of total measured emissions. Blue models are substantially less skewed. The orange model exhibits other behavior, namely, a high zero offset present in field measurement data. Inset illustrates that skewed behavior is detected in all models containing more than 90 observations and detected in no model containing less than 40 observations.

5% of measurements. In 13 models, 5% of measurements account for at least 30% of total emissions, and in 8 models, more than 50% of total emissions. The inset illustrates that the highly skewed character is in every model containing more than 90 observations but in no model containing less than 40 observations. This highlights the importance of large data sets, such as the one used here, to accurately characterize the long tail on emissions distributions. Conversely, for the four models which *do not* exhibit a long tail, these models may underestimate emissions. These models include rod packing vents in one operating mode, wet and dry seal vents, and tank vents.

Since the GHGRP specifies measurement and reporting methods all reporters must use, the mean of reported emissions in each emissions category should be approximately equal between large subpopulations of facilities. However, an examination of data reported to the GHGRP indicated that reported emissions from nonpartner facilities were, on average, approximately 1.4 times larger than that those reported by partner facilities (SI, Table S8-a). Differences were particularly pronounced in categories where the GHGRP requires direct measurement, such as reciprocating compressor venting (1.8 times larger), tank vents (4.6 times larger), and centrifugal compressor venting (7.7 times larger). A difference in reported emissions suggests emissions are impacted by differences in equipment, operational methods, or maintenance practices between partner and nonpartner facilities. Therefore, SME emission models used to analyze nonpartner facilities were modified to account for the reported emission differences, using scaling and differential methods described in the SI, Section 8. Similarly, different models for pneumatic device counts were developed for partner and nonpartner facilities to capture significant differences in pneumatic device utilization (SI, Section 9).

Models for engine and turbine exhaust methane emissions were based upon data underlying EPA standard *Compilation of Air Pollutant Emission Factors* (AP-42)²³ and additional measurements provided by the study partners. AP-42 source data were filtered to include only measurements with complete machine type information and direct measurements of methane. These were combined with partner measurements, particularly for four-stroke, rich burn engines, to develop exhaust emission factors (SI, Section 10).

Finally, when insufficient data existed to develop new models for an emission category, emission factors from the GHGRP were utilized (e.g., intermittent bleed pneumatics and wellhead components, see SI, Section 7).

Activity Models. Activity models are based on data from 922 T&S facilities, including detailed site configuration data for the 686 T&S facilities operated by the six partner companies, and data for 236 nonpartner facilities reporting to the GHGRP. The partner company transmission facilities represent $\approx 56\%$ of the interstate transmission facilities reported to the U.S. Federal Energy Regulatory Commission (FERC) under Form 2.²⁴ Study partner activity data included facility counts, compressor and prime mover descriptions, compressor operating hours in three operating modes (operating/not operating pressurized/not operating depressurized), and component counts for pneumatics and storage wells. Compressor and paired prime mover descriptions determine compressor fugitive and exhaust methane emissions, which contribute a major portion of emissions in the T&S segment.²

Differing levels of activity information were available for different facilities, ranging from comprehensive data for many

study partner facilities to essentially no information for nonpartner facilities not reported to the GHGRP. To account for differing information detail, the model divided the T&S sector into five activity data sets, termed *Lanes*, and a hierarchy of models was developed to estimate emissions for each lane. For partner facilities with complete activity data—including compressor type, operating hours, pneumatic device counts, etc.—activity data were used directly in the model, as shown in Figure 1. Where activity data were incomplete, ambiguous, or absent, activity data were estimated from activity data for similar facilities with known data.

Super-Emitter Emissions. In addition to the equipment-level emissions models, this study also explicitly modeled facility-level “super-emitters”. During the field measurement campaign, two facilities were classified as super-emitters due to elevated emissions observed by tracer flux methods which were not measured using onsite measurements or were inaccurately measured with portable acoustic meters. At these facilities, onsite observers indicated that venting was likely due to an anomalous condition, such as gas leaking through a faulty isolation valve. Due to the difficulties in measuring these types of sources with onsite surveys, they are not fully captured in the long tails of equipment-level emissions distributions. To account for these emissions, this study modeled super-emitter emissions at the facility level, utilizing an emissions distribution developed from tracer flux measurements and a frequency model representing the probability of identifying a super-emitter within a small sample population. These models are shown in Figure 3 and in the SI, Section 14. The frequency model was developed by simulating the sampling process used in the field study, i.e., simulating the probability of finding one super-emitter in a random sample of 45 facilities from the available study population of 686 partner facilities. This process produces a probability distribution and confidence interval

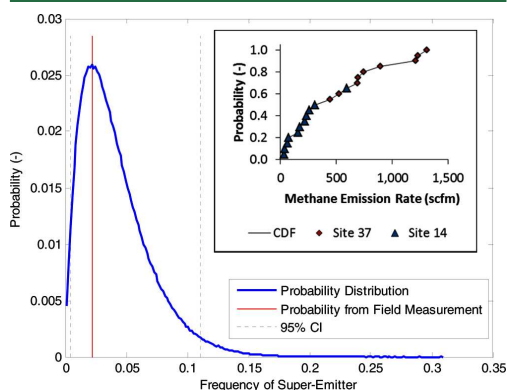


Figure 3. Super-emitter model. Outer plot illustrates the frequency distribution model for super-emitters based upon the probability of finding one super-emitter in 45 randomly sampled facilities from a population of 686 facilities. The mode of the distribution (red line) lies at the probability observed in the field measurement campaign (2.2%). The mean of the distribution is at 4.1%, indicating that approximately 1 in 25 facilities exhibits super-emitter emissions during any given hour of the year. Inset figure illustrates the emission rate model utilized to model super-emitters. The mean of the emission distribution is 496 SCFM. Additional details are provided in the SI, Section 14.

similar to the Wilson score interval²⁵ but applied to a finite population size. The mode of the distribution matches the frequency seen in the field study (2.2%). However, the distribution has a positive skew and a mean of 4.1%; in other words, on average about 1 in 25 facilities is estimated to be a super-emitter at any one time. On each iteration of the Monte Carlo model, the number of facilities simulated as super-emitters is drawn from the frequency distribution in Figure 3. An emissions rate is then drawn for each super-emitter facility from the emissions model shown in the inset of Figure 3 and multiplied by 8,784, the number of hours in 2012. While no one facility may emit at a super-emitter rate for the entire year, the SME assumes that the modeled frequency is representative of the entire T&S sector during any given hour of the year. This method directly represents the uncertainty inherent in observing an infrequent event in a small sample drawn from a larger population.

Comparing to GHGI. The SME includes emission categories that contribute $\approx 87\%$ of the total emissions in the T&S GHGI. However, sufficient data were not available to estimate emissions from liquefied natural gas (LNG) operations, metering and regulation stations, storage wellheads, pipeline venting, pipeline leaks, and other minor categories (SI, Section 15). In total, the GHGI estimates that these excluded categories emit 266 Gg/yr of methane. These emissions were added, without adjustment, to the SME to estimate the total emissions for the entire T&S sector.

Uncertainty Estimation. Uncertainty was defined empirically as the 2.5% and 97.5% fractiles of each computed distribution to produce a 95% confidence interval (CI). The reported CI includes the impact of variability in the emission and activity models, the frequency uncertainty described for the super-emitter frequency model, and uncertainty in facility count. An additional sensitivity analysis was also performed for transmission station count. The GHGI provides a CI for methane emissions from entire the natural gas system of $[+30\%/-19\%]$ ² but does not provide a CI for individual sectors, such as T&S. In this study, we apply the overall GHGI CI to the T&S sector on the assumption that uncertainty in T&S is similar to that of entire natural gas system. Since no CI is provided for subsets of the T&S sector, comparisons for individual emissions categories use the mean value reported in the GHGI.

■ RESULTS AND DISCUSSION

Study Model Estimate. Figure 4 summarizes the predicted methane emissions from the Monte Carlo simulation and comparison to the GHGI and GHGRP. The mean SME is 1,237 Gg/yr $[+36\%/-23\%]$ for the categories explicitly included in our model. Adding unmodeled categories, total T&S emissions are estimated as 1,503 Gg/yr $[+30\%/-19\%]$, corresponding to a methane loss rate of 0.35% [0.28% to 0.45%] of the methane transported by the T&S sector in 2012 (based upon 434 Tg of methane contained in a natural gas throughput of 23.8×10^{12} scf, SI, Section 16).²⁶

On a per station basis, the mean emission rate from compressor stations associated with underground storage facilities (847 Mg/station/yr $[+53\%/-35\%]$) is higher than transmission stations (670 Mg/station/yr $[+52\%/-34\%]$). Underlying drivers include a higher proportion of engine-driven compressors at storage facilities and a larger number of gas-driven pneumatic devices (see below). Both of these equipment differences result, in part, from operational

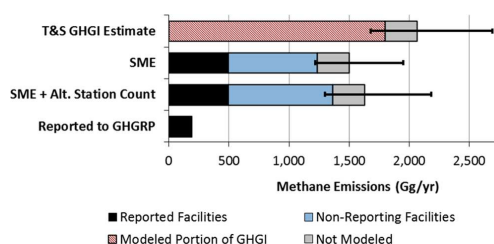


Figure 4. Estimated emissions from the SME for both transmission station count models compared with the 2012 GHGI, including 266 Gg/yr of emissions not modeled in this study ("Not Modeled" bars). Overlapping CI for the GHGI and SME with base station count indicates the two estimates are not statistically different. The GHGI falls within the CI using the alternative station count model, which models 16% more transmission stations and total emissions $\approx 10\%$ higher than the baseline SME. The bottom bar represents the emissions reported to the GHGRP for T&S facilities, including wellhead components. The SME estimate for the same facilities (495 $[+26\%/-17\%]$ Gg/yr, the black portions of the SME bars) represents 260% [215% to 330%] of methane emissions reported the GHGRP. In total, emissions reported to the GHGRP (191 Gg/yr) are 15% [11% to 20%] of the SME emissions estimate (1237 Gg/yr) and 11% of the GHGI (1805 Gg/yr) for the modeled emission categories.

requirements of underground storage facilities, including the high pressures required for injecting gas into the storage field and additional equipment for processing gas extracted from the field. However, despite lower emissions per station, 75% of SME emissions for the T&S sector are from transmission stations, due to the much larger station count.

Figure 5 breaks down the SME by station and emission type for major emission categories. Figure 6 illustrates the major components of the SME results. Fugitive emissions account for 75% of the total SME, and the balance is from pneumatics, engine exhaust, and station venting. Fugitive emissions are unintended emissions from components such as connectors, valves, meters, compressor equipment (seals, isolation, and blowdown valves, etc.), and similar. The SME breaks fugitive emissions into two parts: First, the SME uses equipment-level emissions, which are characterized using empirical distributions of both emissions and activity data that explicitly accounts for the influence of high emitters on equipment-level emissions. Second, super-emitters are characterized by a separate facility-level model which is independent of all equipment categories.

Approximately 50% of the fugitive emissions are from major compressor equipment, including seal vents, unit isolation and blowdown valves, and rod packing vents (SI, Table S17-a). The next largest category of fugitives is super-emitters, which contribute 39% [24% to 48%] of transmission fugitives and 36% [10% to 54%] of storage station fugitives. Finally, leaks from smaller compressor and noncompressor components (open-ended lines, connectors, etc.) contribute 12–13% of fugitive emissions.

Comparisons between SME and GHGI. Figures 4 and 5 compare the SME to the GHGI estimate for the T&S sector. Considering *only* the source categories explicitly included in our model, the overlapping CI of the SME (1,237 [950 to 1,680] Gg/yr) and the GHGI (1,805 [1,460 to 2,350] Gg/yr) indicate that the difference is not statistically significant. However, this statistical similarity obscures the many significant differences in individual emission categories, often in both activity and

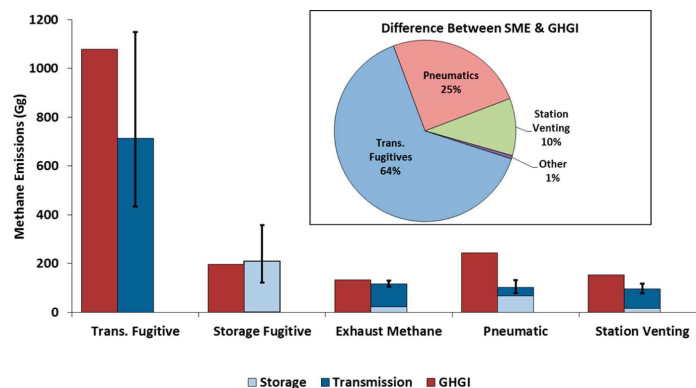


Figure 5. Breakdown of SME into source categories and comparison to the GHGI. Vertical bar chart illustrates the SME divided by major categories, colored by station type, and compared to the GHGI. Black lines represent 95% CI for the SME. The GHGI lies within the CI for three of the five major categories (transmission and storage stations and compressor exhaust) and outside the CI for pneumatic devices and station venting. Inset pie chart summarizes the difference between the SME and GHGI. Two-thirds of the difference originates in transmission fugitives, 25% in pneumatics, and 10% in station venting.

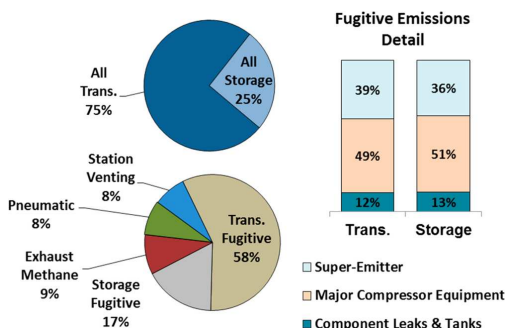


Figure 6. Breakdown of SME emissions by source category. Although emissions per station are higher for storage, 75% of all emissions originate in transmission facilities, due to the much higher station count. Fugitive emissions account for 75% of all emissions. The bar graph further decomposes fugitive emissions. Major compressor equipment, such as isolation and blowdown valve vents and compressor seals, account for approximately 50% of all fugitive emissions, followed by 36–39% modeled at the station level as super-emitters. The remainder is due to other component leaks and tank vents.

emission estimates (SI, Table S17-b). Approximately two-thirds of the difference is in transmission fugitives, followed by 25% for pneumatic devices and 10% in station venting. These differences are driven by three primary factors—the reduced estimate of transmission station count, a significant change in compressor technology mix, and differences in the type and number of gas-driven pneumatic devices. We discuss each of these drivers in turn and then touch on other emission categories.

Station Count. The SME estimates a total of 1,758 [+27%/–19%] stations versus 2,143 in the GHGI (which is based on extrapolation of 1992 data). Controlling for other factors, emissions largely scale with station count, and the reduced station count therefore reduces the overall emissions estimate. Since no complete national data exists for T&S stations, station

count was estimated from available data (SI, Section 11). The count of underground storage compressor stations (382 [+9%/–9%]) was estimated from an EIA inventory of underground storage fields²⁷ and the ratio of storage stations to storage fields for the 72 storage stations operated by the study partners. The GHGI estimates 344 storage stations.

Two approaches were used to estimate the national count of transmission compressor stations. First, a bootstrap method was utilized, based on the transmission stations operated by the study partners and nonpartner facilities reported to the GHGRP. This approach estimates there are 1,375 [+32%/–22%] transmission stations. While the transmission station count is smaller than the GHGI (1,799 stations), the total compressor power of this estimate is very similar to the GHGI estimate due to changes in compressor technology, discussed below. However, since study partner assets are concentrated in the interstate pipeline network, this estimate may not fully account for smaller stations located on intrastate pipeline networks. Therefore, an alternative transmission station count estimate was developed utilizing FERC Form 2 data and scaling by pipeline miles, similar to methods utilized for the GHGI estimate. This method results in a mean estimate of 1,595 [+45%/–25%] transmission stations.

A sensitivity analysis was performed to compare the effect of these two station count estimates on total emissions and is included in Figure 4 as “SME + Alt. Station Count.” The alternative estimate of station count is 16% higher than the mean SME estimate. These additional stations are modeled in the SME as smaller facilities similar to nonreporting transmission stations, resulting in a 13% increase in compressor count and a 10% increase in methane emissions to 1,633 Gg/yr [+34%/–21%]. The alternate station emissions estimate corresponds to 76% [57% to 106%] of the mean GHGI estimate for modeled emissions categories and a methane loss rate of 0.38% [0.30% to 0.50%] of methane transported in the T&S sector in 2012. The CI of the SME with the alternative station counts overlaps with both the baseline SME estimate and the GHGI, indicating that differences are not statistically significant.

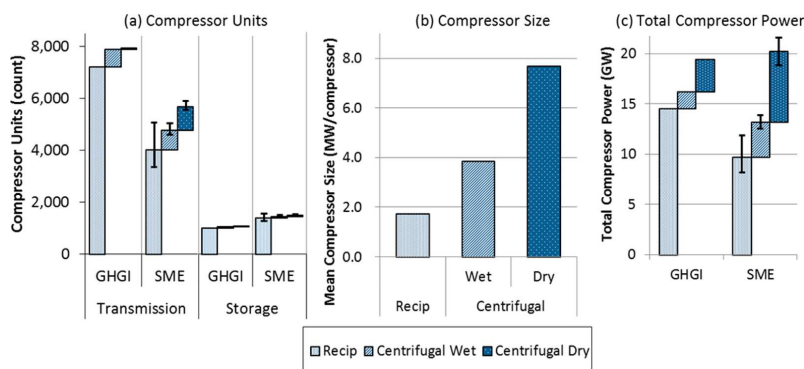


Figure 7. Analysis of compressor technology. (a) The SME, based upon analysis of partner and GHGRP data, indicates substantially more use of centrifugal compressors in transmission (29% [+27%/-19%] of compressors) than the GHGI (9% of compressors). The data also suggest that over 50% of all centrifugal compressors utilize dry seals where the GHGI estimates only 10% dry seals. In contrast, for storage stations, the mix of compressors is similar: GHGI (9% centrifugal) and SME (8% centrifugal). (b) Based upon analysis of partner data, centrifugal compressors, particularly those with dry seals, tend to be significantly larger than reciprocating compressors. (c) The GHGI and the SME (using the base station count estimate) predict similar total compressor power of ≈ 20 GW. However, the SME indicates that substantially more power is implemented using fewer, larger, centrifugal compressors.

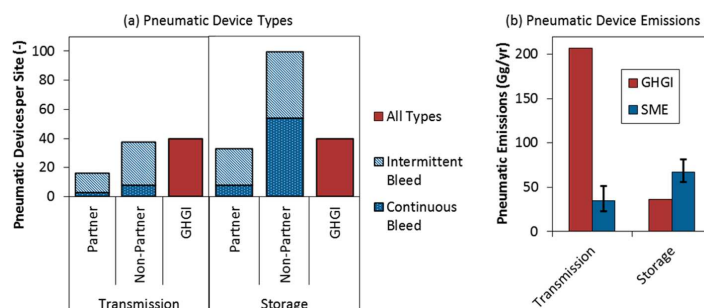


Figure 8. Publicly reported data about partner and nonpartner facilities shows pneumatic devices are unevenly deployed across T&S. Partner facilities utilize substantially fewer gas-driven pneumatic devices. Transmission stations utilize about 33% of the pneumatics per station as storage and a higher proportion of lower-emitting intermittent devices. For transmission stations, SME emissions (35 Gg/yr [+49%/-34%]) are 17% of the GHGI estimate (207 Gg/yr) due to the change in device count, device mix, and a 24% lower station count. For storage, SME emissions (67 Gg/yr [+21%/-18%]) are 190% of the GHGI estimate (36 Gg/yr) due to an increase in devices per station and an 11% larger station count estimate.

Compressor Technology Mix. Partner and GHGRP data indicate a substantial shift in T&S facility configurations relative to the 1992 baseline utilized by the GHGI, with the biggest change in compressor technology for transmission stations. This shift, illustrated in Figure 7, has a substantial impact on methane emissions. Study activity data indicate that 29% [+27%/-19%] of transmission compressors are centrifugal compressors, more than three times the GHGI estimate. Further, study data indicate that over 50% of all centrifugal compressors utilize dry seals, compared with only $\approx 10\%$ in the GHGI.

The shift in compressor type also influences how compressor capacity is implemented. Detailed descriptions from 461 transmission stations (3,006 compressors) indicate that 96% of centrifugal compressors are powered by combustion turbines or electric drives while 98% of reciprocating compressors are powered by internal combustion (IC) engines. The GHGI does not separately identify electrically driven compressors, which our data indicate represent 9% of transmission and 15% of

storage compression capacity. Turbine- and electrically driven units are larger than engine-driven units. Partner data indicate a mean of 5.9 MW/unit for turbines and 9 MW/unit for electric drives versus 1.7 MW/unit for engines. Using average compressor sizes derived from study partner data, the SME predicts a sector total compression capacity of 20.3 GW [+25%/-18%], using the base station count, versus 19.5 GW for the GHGI and (coincidentally) 20.3 GW estimated by scaling FERC Form 2 data, a difference of less than 5% (SI, Section 13). We also predict similar overall compressor usage, (74×10^9 HP · h [+16%/-13%]) compared to the GHGI (69×10^9 HP · h), with similar transmission utilization (102% [90% to 118%] of GHGI) and higher underground storage utilization (156% [124% to 194%] of GHGI). Therefore, our new data reveal a similar aggregate capacity and capacity utilization as the GHGI but suggests a substantial shift in the implementation of that capacity toward fewer, larger, centrifugal compressor units. This shift has two impacts that result in lower emissions: First, methane emissions from turbine exhaust are 2 orders of

magnitude lower on a per-power basis than engine exhaust. Second, centrifugal compressors with dry seals exhibit approximately 33% less fugitive emissions than either reciprocating compressors or centrifugal compressors with wet seals.

Pneumatic Device Emissions. The SME estimates substantially lower methane emissions from pneumatic devices than the GHGI (42% [32% to 55%] of the GHGI estimate), which accounts for 25% of the total difference between the mean SME and GHGI. Gas-driven pneumatic devices are mechanical actuators operated by compressed natural gas typically drawn from pressurized pipelines. Emissions data from this study indicate an emission rate of 3.5 Mg/device/yr [+13%/−12%] for continuous bleed devices (combining both high and low-bleed types) and uses the GHGRP emission factor of 0.40 Mg/device/yr to model intermittent-bleed devices.

Activity models for the SME are based on pneumatic device counts—a reporting requirement of the GHGRP—from 498 partner and nonpartner facilities reporting to the GHGRP. These data indicate a substantially lower utilization of gas-driven pneumatics in transmission (24 devices/station [+49%/−33%]) compared to storage (85 devices/station [+17%/−15%]). In contrast, the GHGI assumes 39 devices/station at all facilities and does not distinguish between intermittent- and continuous-bleed devices. As indicated in Figure 8, usage of gas pneumatics is uneven across the sector. Partner facilities utilize fewer continuous bleed devices in transmission (2.4 devices/station) compared to nonpartner facilities (7.4 devices/station). The difference in storage facilities was even more pronounced, with nonpartner facilities utilizing 7.3 times as many continuous bleed devices (53.7 versus 7.4 devices/facility). These differences are modeled in the SME (SI, Section 9). Data also indicate a much higher proportion of intermittent-bleed devices in transmission (81% of devices) than in underground storage (51% of devices). These complex interactions lead to three results. First, a lower aggregate emission factor for transmission stations (1.0 Mg/device/yr) than the GHGI; second, lower emissions for transmission (17% [11% to 25%] of the GHGI estimate) due to the combination of fewer stations and fewer devices per station coupled with a larger proportion of intermittent devices; and third, an increase in pneumatic device emissions for storage sites, where the larger device count per station (239% [203% to 279%] of GHGI) more than offsets the decrease in emission factor from 2.7 to 2.1 Mg/device/yr.

Station and Compressor Fugitive Emissions. The treatment of fugitive emissions in the SME differs from that in the GHGI in several respects. First, as described earlier, super-emitters are explicitly modeled as a separate facility-level emissions category. Second, driven by available data and observations from the field measurement campaign, the SME includes all component leaks in station fugitives, whereas the GHGI splits component leaks into compressor- and noncompressor component types. In our model, compressor fugitive emissions are restricted to emissions from major compressor equipment (unit isolation valves, blowdown valves, shaft seals, and rod packing vents).

The GHGI divides centrifugal compressors into two categories by seal type and utilizes emission factors updated in a 2010 ICF study.¹⁹ The underlying emission factors from the ICF study indicate an emission rate from wet seals (48 SCFM/compressor) that is eight times that of dry seals (6 SCFM/compressor), while emission rates measured in this study indicate an emission rate of wet seals (16 SCFM/

compressor) approximately three times that of dry seals (5 SCFM/compressor). While this study indicated similar emission rates for dry seals, emission rates for wet seals were significantly lower than utilized for the GHGI (SI, CDFMaster.xls). These results indicate that the emission advantages of dry seals may be overestimated in current literature and may warrant additional study. Annualized emissions are also influenced by the time compressors spend in each operating mode, since seal emissions vary significantly with operating mode. Including these effects, the SME indicates a similar ratio of emissions between wet and dry seal types (1.6 for GHGI versus 1.7 for SME); however, the SME indicates that 22% of emissions for dry-seal compressors are due to seal vents, versus 8% for GHGI.

Compressor Exhaust. Exhaust from both the engines and turbines used to drive compressors contains unburned methane. Methane emission rates in engine exhaust are 2 orders of magnitude higher than in turbine exhaust and therefore dominate exhaust emissions. For engines, the GHGI utilizes an emission factor from the EPA/GRI study of 4.6 g/hp-h. However, total engine emissions are adjusted downward by 124 Gg/yr due to “voluntary reductions,” producing a net emission factor of 2.4 g/hp-h (SI, Section 15). The SME derived a completely independent emission factor of 3.7 g/hp-h that is 1.6 times the GHGI emission factor (SI, Table S17-a). Therefore, the voluntary reduction included in the GHGI appears to overestimate actual reductions in exhaust methane emissions.

Station Venting. Station venting includes intentional releases of gas to depressurize equipment, which are also known as “blowdowns.” The GHGI estimates station venting using emission factors from the 1996 EPA/GRI study¹⁸ based upon engineering estimates of both the frequency and volume of blowdowns for 15 compressor stations. The SME utilized logged blowdown events from 617 transmission facilities and driven by GHGRP rules, includes all blowdowns larger than 50 ft³ occurring within station boundaries (SI, Section 4). Based upon these data, the SME estimates lower transmission station venting emissions, 54 Mg/station/yr versus 72 Mg/station/yr, for the GHGI. Since no data were available for underground storage stations, storage stations were modeled using transmission station data weighted by the number of compressors at the facility. This model indicates a smaller emission factor for underground storage stations of 43 Mg/station/yr. As with facility counts, more comprehensive reporting of blowdowns would reduce the uncertainty in these estimates.

Comparison to the GHGRP. Approximately 40% [37% to 43%] of the SME emissions are attributed to the 498 partner and nonpartner facilities that reported to the GHGRP in 2012. These facilities represent 28% of the mean SME facility count for the T&S sector. For this subpopulation of facilities, the SME emissions are 495 Gg/yr [+26%/−17%] versus 191 Gg/yr reported to the GHGRP using GHGRP rules and emission factors (black bars in Figure 4). Three-quarters of the difference between the SME and GHGRP is driven by three emission areas: (a) facility-level super-emitter emissions not included in the GHGRP (33%), (b) methane in engine exhaust estimated using emissions factors based up AP-42 rather than the GHGRP's Subpart C emissions factors (27%), and (c) rod packing vent emissions in not operating, pressurized mode, an emission source omitted from the GHGRP (15%). The remaining difference is due to other sources omitted from the GHGRP, updated emission factors, and differences between

GHGRP measurement methods and those allowed in the study protocol (SI, Table S18-b).

Due to these factors and the reporting threshold for T&S facilities (25,000 mtCO₂e in 2012), the 191 Gg/yr reported under GHGRP represents 15% [11% to 20%] of the 1,237 Gg/yr methane emissions in the SME-modeled portion of the T&S sector. This under-reporting is due to the requirements of the GHGRP rules, not noncompliance by industry.

The GHGRP is a significant new source of greenhouse gas emissions data, covering approximately 25% of T&S facilities, and providing new and valuable activity data, particularly for pneumatic devices, component leaks and compressor equipment, and new emission data in certain reporting categories. However, emissions reported to the GHGRP significantly underestimate emissions from reporting facilities, and need to be interpreted carefully.

Impact of Long Tails in Emissions Distributions. The empirical emissions distributions utilized in the SME allow the impact of long tail emissions to be assessed for fugitive emissions at the component level and for super-emitter emissions at the facility level:

- **Impact of "Long Tail" on Fugitive Emissions:** As illustrated in Figure 2, the "body" of most fugitive emissions distributions represents at least 95% of all measurements (SI, Section 7). These emissions occur in many components and therefore scale with activity levels, such as equipment counts and operating hours, and are characterized by the count and choice of equipment. In contrast, the long tails in Figure 2 indicate emissions that are significantly higher than average are likely in most emission categories and are likely due to equipment or control malfunctions. These emissions, and by extension a substantial proportion of total emissions, would likely be effectively reduced by reducing the frequency at which malfunctions occur and/or reducing the emission rate should a malfunction occur. For 11 of the 17 emission models in Figure 2, the largest 5% of measurements account for 40–75% of measured emissions. Assuming a conservative 40% of emissions in each of these categories could be identified and eliminated, total emissions would be reduced by 14–15%.
- **Impact of Super-Emitters on T&S Emissions:** The paired onsite and tracer measurements made in this study¹⁶ indicate that most emissions may be characterized by comprehensive onsite measurements, but some large facility-level super-emitters are not readily measured by onsite methods. Qualitative observations made during this study indicate that these super-emitter emissions do not scale with equipment counts except at the coarsest level (e.g., facilities) and are better modeled as a frequency of occurrence within a population of facilities. Facility-level super-emitters account for 23% [9 to 39%] of the SME for modeled categories. The model presented here provides a systematic treatment of these emitters. However, the wide uncertainty in the estimate indicates that additional data on both frequency and magnitude would improve national emission estimates.

Together these two forms of skewed emissions distributions account for nearly 40% of the SME emissions. Attention to these emissions categories could have significant impact on overall emissions from the T&S sector.

Considerations for Inventory Development. Although our analysis indicates that the GHGI agrees, within statistical uncertainty, with our estimate of overall emissions from the T&S sector, there are significant differences in a number of key drivers of these emissions. These detailed differences matter if the GHGI is used to assess, for example, the targets and costs of potential emission reduction strategies.²⁸ Our extensive new data revealed errors in the GHGI activity data (both counts and technology mix) that have an impact on overall emissions similar in magnitude to changes in emission rates, including facility-level super-emitters (SI, Table S17-b). Our analysis suggests that a limitation of the GHGI for the T&S sector may not be gaps in understanding emissions but instead what many would consider to be more fundamental, "easy to get" data: counts of facilities and major equipment at those facilities. From an inventory development perspective, one of the most valuable aspects of the GHGRP as it is currently constructed is that it provides this information for the estimated 28% of sites that report to GHGRP. Comprehensive reporting of activity data, including intrastate systems, even if unaccompanied by emissions measurement, would substantially improve inventory completeness and reduce uncertainty.

■ ASSOCIATED CONTENT

■ Supporting Information

Additional descriptions of the model, emissions models (CDFMaster.xlsx), principal activity models (Activity Models.xlsx), and data input data from study partners and the GHGRP (Input Data Set.xlsx). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.5b01669.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dan.zimmerle@colostate.edu. Phone: 970-581-9945. Address: Energy Institute, Colorado State University, 430 N. College Avenue, Fort Collins, Colorado 80524, United States (Daniel Zimmerle).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the sponsors of this work for financial support, technical advice, and access to sites for sampling. The sponsors were the Environmental Defense Fund (EDF), Dominion, Dow Chemical, Enable Gas Transmission, LLC (formerly CenterPoint Energy Gas Transmission Company, LLC), Interstate Natural Gas Association of America (INGAA), Kinder Morgan, Columbia Pipeline Group, TransCanada, and The Williams Companies, Inc. Funding for EDF's methane research series, including this work, is provided by Fiona and Stan Druckenmiller, Heising-Simons Foundation, Bill and Susan Oberndorf, Betsy and Sam Reeves, Robertson Foundation, Alfred P. Sloan Foundation, TomKat Charitable Trust, and the Walton Family Foundation. We also thank the project science advisory panel for valuable feedback: J. Kuo, D. Picard, R. Talbot, and M. Whelan. The views and opinions expressed are those of the authors and do not necessarily reflect those of the sponsors or the project science advisory panel.

■ REFERENCES

- (1) Alvarez, R. A.; Pacala, S. W.; Winebrake, J. J.; Chameides, W. L.; Hamburg, S. P. Greater focus needed on methane leakage from natural gas infrastructure. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (17), 6435–6440.
- (2) U.S. Environmental Protection Agency. *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2012*; EPA 430-R-14-003; Washington, DC, 2014.
- (3) Pétron, G.; Frost, G.; Miller, B. R.; Hirsch, A. I.; Montzka, S. A.; Karion, A.; Trainer, M.; Sweeney, C.; Andrews, A. E.; Miller, L.; et al. Hydrocarbon emissions characterization in the Colorado Front Range: A pilot study. *J. Geophys. Res.* **2012**, *117* (D4), D04304.
- (4) Miller, S. M.; Wofsy, S. C.; Michalak, A. M.; Kort, E. A.; Andrews, A. E.; Biraud, S. C.; Dlugokencky, E. J.; Eluszkiewicz, J.; Fischer, M. L.; Janssens-Maenhout, G.; et al. Anthropogenic emissions of methane in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (50), 20018–20022.
- (5) Pétron, G.; Karion, A.; Sweeney, C.; Miller, B. R.; Montzka, S. A.; Frost, G. J.; Trainer, M.; Tans, P.; Andrews, A.; Kofler, J.; et al. A new look at methane and nonmethane hydrocarbon emissions from oil and natural gas operations in the Colorado Denver-Julesburg Basin. *J. Geophys. Res. Atmospheres* **2014**, *119* (11), 6836.
- (6) Peischl, J.; Ryerson, T. B.; Brioude, J.; Aikin, K. C.; Andrews, A. E.; Atlas, E.; Blake, D.; Daube, B. C.; de Gouw, J. A.; Dlugokencky, E.; et al. Quantifying sources of methane using light alkanes in the Los Angeles basin, California. *J. Geophys. Res. Atmospheres* **2013**, *118* (10), 4974–4990.
- (7) Karion, A.; Sweeney, C.; Pétron, G.; Frost, G.; Michael Hardesty, R.; Kofler, J.; Miller, B. R.; Newberger, T.; Wolter, S.; Banta, R.; et al. Methane emissions estimate from airborne measurements over a western United States natural gas field. *Geophys. Res. Lett.* **2013**, *40* (16), 4393–4397.
- (8) Brandt, A. R.; Heath, G. A.; Kort, E. A.; O'Sullivan, F.; Pétron, G.; Jordaan, S. M.; Tans, P.; Wilcox, J.; Gopstein, A. M.; Arent, D.; et al. Methane Leaks from North American Natural Gas Systems. *Science* **2014**, *343* (6172), 733–735.
- (9) Allen, D. T.; Torres, V. M.; Thomas, J.; Sullivan, D. W.; Harrison, M.; Hendler, A.; Herndon, S. C.; Kolb, C. E.; Fraser, M. P.; Hill, A. D.; et al. Measurements of methane emissions at natural gas production sites in the United States. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (44), 17768–17773.
- (10) Allen, D. T.; Pacsi, A. P.; Sullivan, D. W.; Zavala-Araiza, D.; Harrison, M.; Keen, K.; Fraser, M. P.; Daniel Hill, A.; Sawyer, R. F.; Seinfeld, J. H. Methane Emissions from Process Equipment at Natural Gas Production Sites in the United States: Pneumatic Controllers. *Environ. Sci. Technol.* **2015**, *49* (1), 633–640.
- (11) Prasino Group. *Final Report for Determining Bleed Rates for Pneumatic in Devices British Columbia*; Calgary, Canada, 2014.
- (12) Allen, D. T.; Sullivan, D. W.; Zavala-Araiza, D.; Pacsi, A. P.; Harrison, M.; Keen, K.; Fraser, M. P.; Daniel Hill, A.; Lamb, B. K.; Sawyer, R. F.; et al. Methane Emissions from Process Equipment at Natural Gas Production Sites in the United States: Liquid Unloadings. *Environ. Sci. Technol.* **2015**, *49* (1), 641–648.
- (13) Lamb, B. K.; Edburg, S. L.; Ferrara, T. W.; Howard, T.; Harrison, M. R.; Kolb, C. E.; Townsend-Small, A.; Dyck, W.; Possolo, A.; Whetstone, J. R. Direct Measurements Show Decreasing Methane Emissions from Natural Gas Local Distribution Systems in the United States. *Environ. Sci. Technol.* **2015**, *49* (8), 5161–5169.
- (14) Brantley, H. L.; Thoma, E. D.; Squier, W. C.; Guven, B. B.; Lyon, D. Assessment of Methane Emissions from Oil and Gas Production Pads using Mobile Measurements. *Environ. Sci. Technol.* **2014**, *48* (24), 14508–14515.
- (15) Mitchell, A. L.; Tkacik, D. S.; Roscioli, J. R.; Herndon, S. C.; Yacovitch, T. I.; Martinez, D. M.; Vaughn, T. L.; Williams, L. L.; Sullivan, M. R.; Floerchinger, C.; et al. Measurements of Methane Emissions from Natural Gas Gathering Facilities and Processing Plants: Measurement Results. *Environ. Sci. Technol.* **2015**, *49* (5), 3219–3227.
- (16) Subramanian, R.; Williams, L. L.; Vaughn, T. L.; Zimmerle, D.; Roscioli, J. R.; Herndon, S. C.; Yacovitch, T. I.; Floerchinger, C.; Tkacik, D. S.; Mitchell, A. L.; et al. Methane Emissions from Natural Gas Compressor Stations in the Transmission and Storage Sector: Measurements and Comparisons with the EPA Greenhouse Gas Reporting Program Protocol. *Environ. Sci. Technol.* **2015**, *49* (5), 3252–3261.
- (17) Greenhouse Gas Reporting Program Home. U.S. Environmental Protection Agency. <http://www.epa.gov/ghgreporting/> (accessed June 24, 2015).
- (18) Harrison, M. R.; Shires, T. M.; Wessels, J. K.; Cowgill, R. M. *Methane Emissions from the Natural Gas Industry*; GRI-94/0257 and EPA-600/R96–080; Gas Research Institute and U.S. Environmental Protection Agency: Washington, DC, 1996.
- (19) ICF International. *Emissions from Centrifugal Compressors*; Document provided via personal correspondence with U.S. EPA personnel, 2010.
- (20) U.S. EPA Office of Air Quality Planning and Standards. *Oil and Natural Gas Sector Leaks*; Report for Oil and Natural Gas Sector Leaks Review Panel; Washington, DC, 2014.
- (21) Ross, S. M. *Simulation*; Academic Press, 2013.
- (22) Roscioli, J. R.; Yacovitch, T. I.; Floerchinger, C.; Mitchell, A. L.; Tkacik, D. S.; Subramanian, R.; Martinez, D. M.; Vaughn, T. L.; Williams, L.; Zimmerle, D.; et al. Measurements of methane emissions from natural gas gathering facilities and processing plants: measurement methods. *Atmos. Meas. Tech.* **2015**, *8* (5), 2017–2035.
- (23) Emissions Factors & AP 42, Compilation of Air Pollutant Emission Factors. Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency. <http://www.epa.gov/ttn/chief/ap42/index.html> (accessed July 2105).
- (24) Form 2/2A - Major and Non-major Natural Gas Pipeline Annual Report. Federal Energy Regulatory Commission (FERC). <http://www.ferc.gov/docs-filing/forms/form-2/data.asp> (accessed July 2015).
- (25) Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **1927**, *22* (158), 209–212.
- (26) Natural Gas, Natural Gas Annual with Data for 2012, Table 1. U.S. Energy Information Agency. <http://www.eia.gov/naturalgas/annual/archive/2012/> (accessed July 2015).
- (27) Natural Gas Annual Respondent Query System: 191 Field Level Storage Data (Annual), U.S. Energy Information Agency. <http://www.eia.gov/> (accessed July 2015).
- (28) ICF International. *ICF Methane Cost Curve Report*; Fairfax, VA, 2014.

B.2 A Methane Emission Estimation Tool (MEET) for predictions of emissions from midstream gas compressor stations with fine scale temporal and spatial resolution.

Supplementary Information

B.2.1 Compressor Station Equipment

Traditionally, the midstream natural gas supply chain has been divided into two sectors: G&P and T&S. The compressor stations in these sectors compress gas for transport between facilities in the natural gas supply chain. The MEET model includes models and data for gathering, transmission, and storage stations. Compressors at gas processing plants are often similar to those in transmission or gathering stations [23]. The model does not include the other equipment at processing plants, such as gas fractionation systems.

A compressor station is a physically proximate group of equipment, typically surrounded by a fence or similar outer marker, which receives gas from one or more pipelines from well pads or other stations, compresses and often upgrades the gas, and injects the gas into another pipeline. In upstream and storage applications, stations often include equipment that upgrades gas quality, including dehydrators and acid gas removal units (AGRU), and often host equipment to support connected pipelines, such as pig launchers and receivers, or metering equipment.

The equipment model utilizes a hierarchy of station-unit-component, which matches well with available emission and activity data. Equipment groups utilized in this work are shown in Figure B.1. A station contains major equipment units (dark outlines in figure). Each unit contains various types and numbers of components (lighter or dotted lines). Figure B.2 shows an aerial view of a compressor station with key equipment marked.

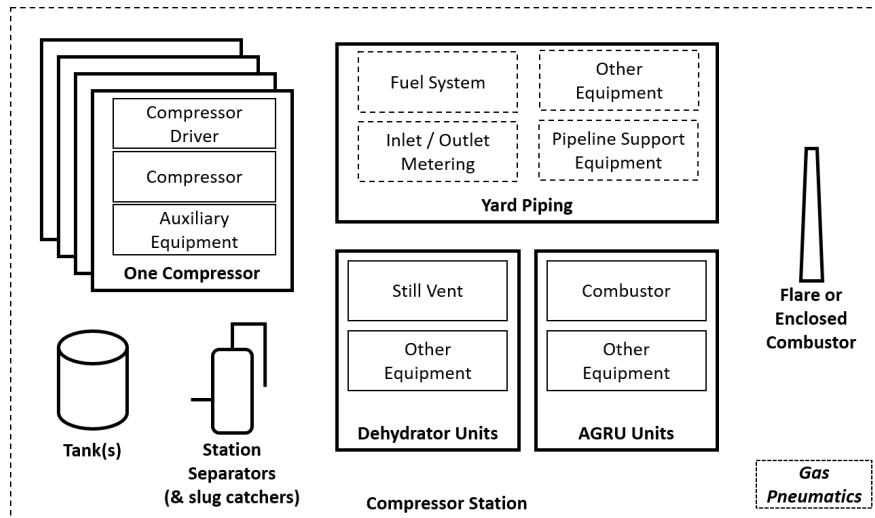


Figure B.1: Compressor station equipment included in the model. Major equipment units are shown in dark lines, while constituent components of each unit are shown in lighter or dotted lines. Not all equipment is present at every station.

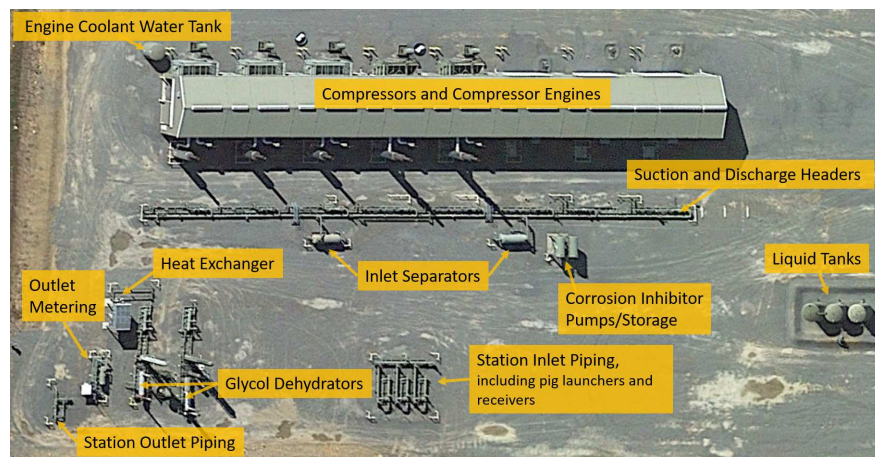


Figure B.2: Aerial view of a large compressor station that includes both compression and dehydration, pipeline pig launchers and receivers for incoming and outgoing pipelines, and metering for delivered gas. This station is connected to well pads upstream via gathering pipelines and connects to a nearby transmission line to deliver gas.

Common emissions inventory practice, such as the U.S. EPA Greenhouse Gas Inventory (GHGI), tracks gas-driven pneumatic controllers (PCs) at the station level, rather than the major equipment units on which they are deployed. Therefore, PCs may need to be aggregated to the station level if counts per major equipment unit are not available.

In keeping with common practice, we distinguish between station-level separators and unit separators. Station separators are typically larger separators installed at the inlet of the station to capture liquids entrained in the incoming gas. Unit separators are included on the skids of individual equipment units, such as inlet and interstage separators on compressor skids, dropout tanks on dehydrators, etc. Station separators are modeled as separate major equipment units. Component counts for unit separators must be included in the component counts for major equipment units.

Figure B.3 shows the principal components of a single compressor, including support components which are outside the compressor skid itself. Every compressor has three principal valves: suction and discharge isolation valves and a blowdown valve. (Some compressors also have a bypass valve which connects the suction to discharge to bypass the compressor; these are not modeled in this revision of MEET.) These three valves are in different states for each of the compressor's operating modes, as shown in Table B.2.

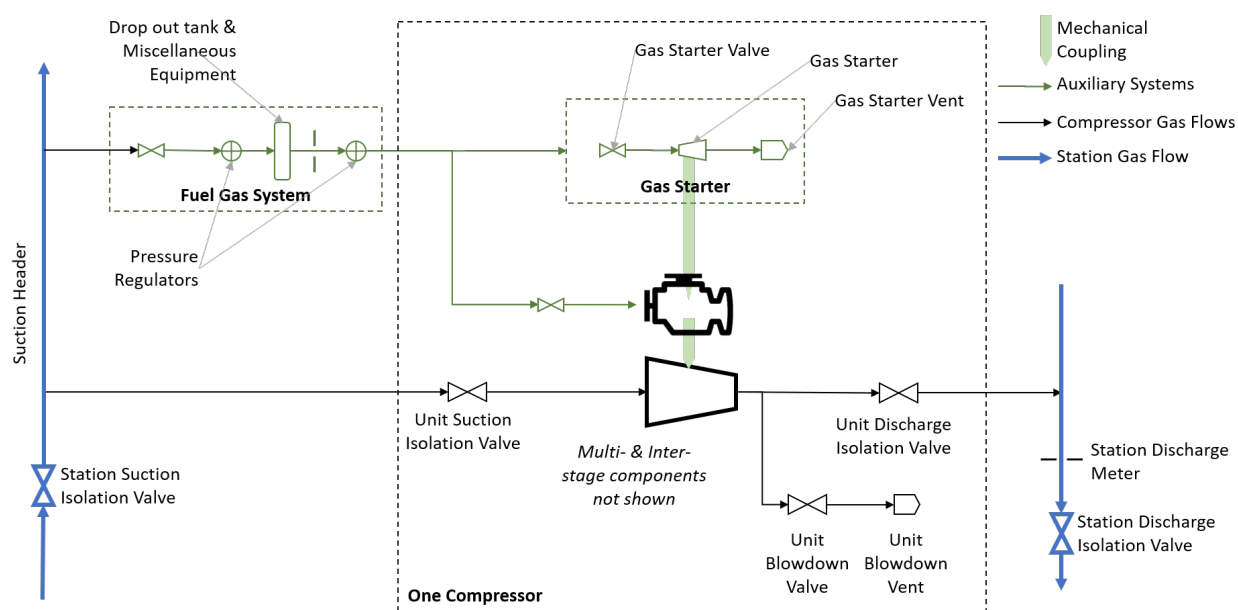


Figure B.3: Simplified single compressor schematic. The compressor is shown with a gas-powered starter; many compressors use air or electric starters, which do not have direct emissions. Typical compressor drivers – engines and turbines – also have exhaust emissions, not shown. The fuel gas system provides combustion fuel throughout the station and may also provide gas for gas-driven PCs. Some stations have multiple fuel gas systems.

Figure B.4 illustrates a single compressor in more detail, without the driver, starter, or fuel gas system. In addition to the major valves in Figure B.3, additional vents and associated valves are shown. A key emission source is rod packing or centrifugal compressor seal vents, which are typically routed to the atmosphere through vent lines on the compressor, or into a combined vent stack for the compressor station or compressor house. For this model, we include dog house (distance piece) vents in rod packing, as these emissions originate with gas movement through the rod packing seals [33, 71, 62]. Compressors may also have additional vents which are also routed to the atmosphere, but are used infrequently, (e.g. pocket vents, which vent gas when the pocket size is being adjusted, a rare, and typically manual, operation).

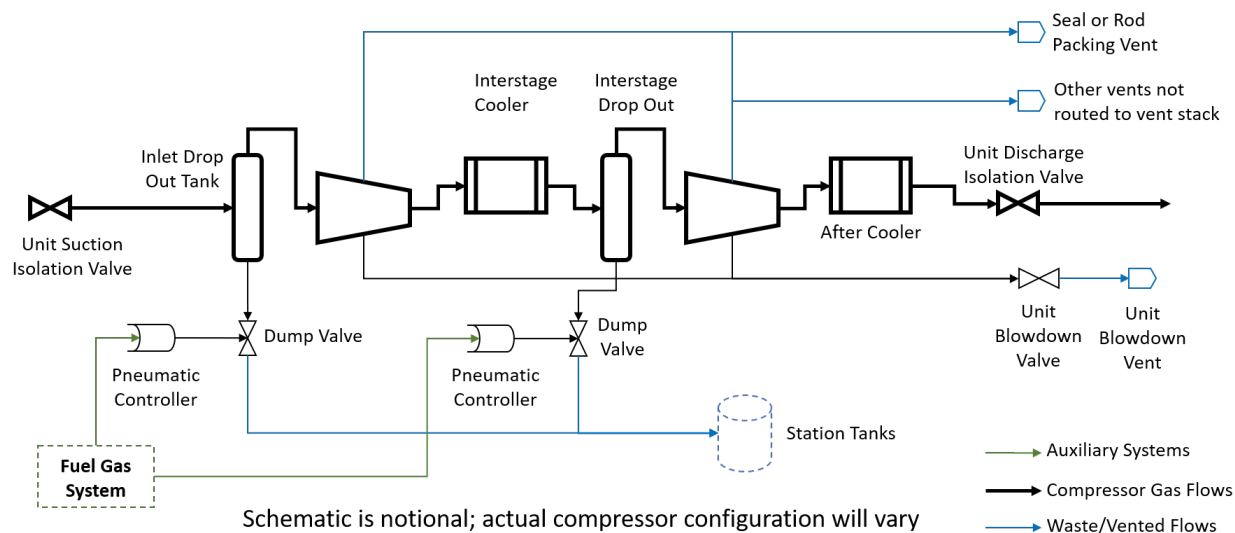


Figure B.4: Compressor schematic for a multi-stage compressor. Interstage separators are often located on the compressor skid for smaller compressors (<2000 HP) typical of gathering stations, and as separate yard equipment for larger compressors typical of transmission stations.

Modeled Emission Sources

Eight component categories are included in the model for each unit of major equipment, based upon data from two recent studies of compressor stations [61, 33]. Emissions from these components are always fugitive emissions (Section B.2.5):

- Threaded connectors
- Open Ended Lines (OELs)
- Valves
- Meters
- Pressure Reducer Valves (PRVs)
- Flanged connectors
- Regulators
- Gauges

A final component type, gas-driven PCs, are modeled similarly, but emissions are classified differently. PCs vent gas when operating; the rate depends upon the type of PC. In addition, these devices may malfunction and either vent excessive gas or leak gas when emissions should be zero. Current emission factors, based upon data collected in field studies, co-mingle these emissions. Therefore, all PC emissions from continuous bleed controllers are currently classified as *venting*, but may include some fugitive emissions.

For intermittent controllers, the compressor station model includes emissions behavior taken from Luck et al. [56, 55], that includes both vented and fugitive emissions from periods where the controller is malfunctioning. Actuation timing was also extracted from these data. Therefore, simulation results used for this study include highly variable emissions characteristics typical of

controllers on gathering compressor stations, and those emissions are dynamically assigned to either vented or fugitive classifications; when an intermittent pneumatic controller is behaving normally, all emissions are categorized as “vented”; when the controller is behaving abnormally all emissions are categorized as “fugitive”

Finally, special consideration is required for PC attached to compressor units. For this study, we assume all PC assigned to compressors are attached to operating components of the compressor, including pressure regulators, separator dumps, and other control and safety valves. PCs on fuel systems and other supporting equipment are included in the yard piping category. For the PCs on compressors, the operational state when the compressor is pressurized but not operating (NOP mode) is ambiguous, as some controllers may still be active (e.g. continuous bleed pressure regulators), while others (e.g. dump valve controllers) are likely to be inactive. Therefore, to simplify the operational behavior, the current model assumes:

- All gas PCs are active whenever the compressor is pressurized – that is, in NOP and OP modes.
- All gas pneumatics controllers are depressurized and inactive when the compressor depressurized – NOD mode.

Working from Figure B.3 and B.4, all compressors are assumed to have the following additional ‘major’ components:

- One ‘single unit’ vent, as defined in Zimmerle et al. [71], and multiple vents are assumed to be plumbed into this vent. Emissions from this vent are fugitive emissions (Section B.2.5).
- One rod packing or seal vent that includes emissions from all compressor cylinders (Section B.2.7)
- One blowdown valve, which has the potential for both fugitive and vented emissions as defined in Table B.2. (Section B.2.6)

- One starter valve, which has the potential for both fugitive and vented emissions as defined in Table B.2. (Section B.2.6)
- Two isolation valves (suction and discharge). Emissions are treated as leaks, Section B.2.5, but only occur in specific operating modes, defined in Table B.2.
- One pocket vent per cylinder; emissions from pocket vents are always considered fugitive emissions, as venting only occurs when the pocket size is being changed. (Section B.2.5)

Sources with No Models or Simplified Models

In addition to the emissions included in the model, several emission types are *not* included or are modeled in a simplified fashion. First, considering vented emissions, two vent types have not been discussed:

- Unlike production sites in the MEET model, tank flash emissions are not modeled, due to a lack of data for the composition of fluids in compressor station tanks. Fugitive emissions from other tank components are included in yard piping, as in Zimmerle et al. [71]
- Full or partial station blowdowns are not modeled. These events are rare, but potentially large. However, little data for timing or frequency exists; see Section B.2.6.

Second, there are several additional combustion emission sources on compressor stations:

- Combustion exhaust emissions from AGRU steam heaters or dehydrator reboilers are not modeled; these are not included in GHGRP models and little empirical data is available for use in models.
- The model assumes that AGRU vents are routed to on-site flares, and those emissions are included in flare emissions.

Table B.1: Valve states for compressor operating modes

Valve	OP	NOP	NOD	START	BLOWDOWN
Unit isolation	Open	Assume Closed ¹	Closed	Open	Closed
Blowdown	Closed	Closed	Assume Open ²	Closed	Open
Starter	Closed	Closed	Closed	Open	Closed

¹ Both open or closed positions are valid; model assumes closed.

² Some operators close the blowdown valve after blowing down, others do not. The model assumes open, i.e. not leaking.

- Since little data exists for operating states of dehydrators, dehydrator reboilers vents are modeled by drawing random samples from GHGRP reports. This same method is used for flare exhaust emissions.

Major Valves States by Operating Modes

Compressor operating modes are defined in the main paper. In each state, the major valves in Figure B.4 change position, as shown in Table B.1. When the valve is closed, it may be leaking – i.e. the valve is not sealing properly. When a valve is open it is not leaking through the valve seal but may still leak through valve stem packing. Valve stem emissions are always considered leaks, and are included in component emissions, category ‘*valve*,’ while valve seal leaks are modeled with separate emission factors for specific valves.

Emitter States in Each Compressor Operating Mode

The model tracks the state of the compressor and suspends/restarts emissions from mode-dependent leaks when the compressor is in a different mode. For example, venting from rod packing only occurs in OP and NOP modes, and is rare in NOD mode. In all cases, both the frequency and emission rate are set by the emission factors selected for the model. Emissions from the gas starter are vented during the starting operation. Emissions at other times are leaks. Similarly, emissions through the blowdown valve are classified as vented emissions when the compressor is

actively in the process of blowing down; emissions through this valve are considered leaks at any other time, as no gas should be emitted from this valve when the compressor is not blowing down.

Table B.2 shows all emission types included in the current compressor station model, with conditions noted in the table legend. All component leaks are modeled by type of component, and assumed to be quasi-steady state; their emissions do not vary over time, except in cases where the operating mode of the compressor changes the leak between active and inactive.

Table B.2: Emitter states by compressor mode

Emitter	OP	NOP	NOD	START	BLOWDOWN
Components					
Connector (flanged)	F	F		F	F
Connector (threaded)	F	F		F	F
Gauges	F	F		F	F
Meter	F	F		F	F
OEL	F	F		F	F
PRV	F	F		F	F
Regulator	F	F		F	F
Valve	F	F		F	F
Pocket vent	F	F		F	F
Gas pneumatics⁴	V⁴	V⁴		V	V
Major Valves					
Blowdown valve	F	F		F	V ¹
Isolation valve		F	F		F
Starter valve	F	F	F	V ¹	F
Centrifugal seals and rod packing					
Rod packing	V ²	F ²	F ^{2,3}		
Dry seal	V				
Wet seal	V				
Combined vents and stacks					
Single unit vent	F	F		F	
Combustion exhaust	C				

Key: *F* indicates a fugitive emission, *V* indicates a vented emission, *C* indicates combustion exhaust emissions, and blank indicates the emitter is inactive.

¹ Vented emissions are episodic, see Section B.2.6.

² Rod packing emission rates are available for each compressor operating mode.

³ Rod packing vents occasionally emit gas which enters the compressor through leaking valves in NOD mode.

⁴ The intermittent PCs emission model may produce either fugitive and vented emissions.

B.2.2 Operating Mode Timing

Two possibilities exist to model the time series of compressor operating states. In rare cases, start and stop times, and blowdown events, may be acquired directly from SCADA records. In the more common case when these data are not available, the primary activity data driving emissions is an estimate of the run time of the compressor over the simulation period, the uptime ratio, defined as the fraction of time the compressor is operating. Transmission, storage and gas processing stations that exceed reporting thresholds report hours in OP/NOP/NOD modes to the GHGRP. For other sectors, public annual data is not available, but uptime can often be estimated from SCADA or maintenance records.

To model runtimes, we use two steps: probability distributions for the time in each state (Figure 2, main paper) and a transition matrix to define the probability that the unit will transition between the current mode and all other possible modes. When a compressor enters a mode, the time in that mode is randomly drawn from an appropriate distribution. When the time in-mode expires, the model uses the transition matrix to move to the next state.

The transition matrix defines the probability of transitioning from one state to another whenever there is a choice in possible destination states, as shown in in Table B.3. Probabilities in each row sum to 1. Diagonal values are zero, because we are drawing duration from duration probability distributions to determine how long we are in each state. We also assume that transitions from NOD to NOP are very rare ($\Pi \approx 0$), or short; if a depressurized compressor is pressurized, it will also be started.

Table B.3: Transition matrix for compressor states

Current State	Future State [†]			Constraint
	OP	NOP	NOD	
OP	0	$\Pi_{OP,NOP}$	$\Pi_{OP,NOD}$	$\Pi_{OP,NOP} + \Pi_{OP,NOD} = 1$
NOP	$\Pi_{NOP,OP}$	0	$\Pi_{NOP,NOD}$	$\Pi_{NOP,OP} + \Pi_{NOP,NOD} = 1$
NOD	$\Pi_{NOD,OP}$	$\approx 0^2$	0	$\Pi_{NOD,OP} = 1$

[†] Short-term episodic states in Figure 2 of the main paper are not included in the table because they are episodic transitions between the OP, NOP, and NOD states.

² Probability is zero because the model assumes that pressurizing the compressor will always be followed by an attempt to start the compressor.

Duration probability distributions determine the distribution of time the compressor spends in each state. For each time, $t_x, x = OP, NOP, NOD$, there is an assigned probability distribution, $\mathcal{P}_x$, that defines the duration in each state, derived from data analysis and/or engineering judgment. Reasonable defaults are provided in the sample input files. In many cases, only the uptime fraction of the compressor is known or can be estimated. Uptime is defined as:

$$u = \frac{t_{OP}}{t_{OP} + t_{NOP} + t_{NOD}} \quad (\text{B.1})$$

and $t_{OP} + t_{NOP} + t_{NOD}$ is typically the total time in the recording period, such as annual hours. From this type of record, often only $t_{off} = t_{NOP} + t_{NOD}$ is known, and some estimate must be made for the fraction of time a unit spends in NOP versus NOD mode. In this case, transition probabilities and the mean values of duration distributions ($\mathcal{P}_x$) must jointly produce the correct uptime for the compressor. We are given the uptime fraction, u , and want to find the scale factor, f , to apply to $\mathcal{P}_x$ to produce the uptime factor during the simulation, i.e. $t_{off} = ft_{OP}$. u is defined as:

$$u = \frac{\mu_{OP}}{\mu_{OP} + f\mu_{OP}} \quad (\text{B.2})$$

Rearranging:

$$f = \frac{1}{u} - 1 \quad (\text{B.3})$$

Because there are multiple downtime modes, we need to scale downtimes to consider the means and the fraction in each mode. First, we assume mean times in NOD and NOP mode; only the ratio is important because absolute values will be calculated later. The mean time in off mode is then:

$$\begin{aligned} \hat{t}_{off} = & \Pi_{OP,NOP}\Pi_{NOP,OP}\mu_{NOP} + \Pi_{OP,NOD}\Pi_{NOD,OP}\mu_{NOD} \\ & + \Pi_{OP,NOP}\Pi_{NOP,NOD}\mu_{NOD} + \Pi_{OP,NOD}\Pi_{NOD,NOP}\mu_{NOP} \end{aligned} \quad (\text{B.4})$$

where the caret indicates an estimated value. The total on time is t_{OP} . To produce the desired uptime, the mean of the duration distributions, $\mathcal{P}_{NOP}$ and $\mathcal{P}_{NOD}$, must be scaled by the scaling factor:

$$g = \frac{t_{OP}}{\hat{t}_{off}} \frac{1}{u - 1} \quad (\text{B.5})$$

Example code is:

```

totalOff= transMatrix(OP,NOP)*transMatrix(NOP,OP)*nopTime.Mean ...
+ transMatrix(OP,NOD)*transMatrix(NOD,OP)*nodTime.Mean ...
+ transMatrix(OP,NOP)*transMatrix(NOP,NOD)*nodTime.Mean ...
+ transMatrix(OP,NOD)*transMatrix(NOD,NOP)*nopTime.Mean;
totalOn=opTime.Mean;
g=totalOn/totalOff*(1/uptime-1);
nodTime.Mean=nodTime.Mean*g;
nopTime.Mean=nopTime.Mean*g;

```

B.2.3 Gas Composition

Emitted gas composition varies between equipment upstream and downstream of gas upgrading equipment (dehydrators and AGRUs), and gas emitted from tanks likely has a different gas composition from other locations on the station. However, in keeping with common industry practice, the current model assumes one gas composition for all emission locations at the station, with the exception of combustion sources (engine exhaust, flares and combustors), which use other models (e.g. GHGRP or AP-42 models) to calculate emissions.

B.2.4 Sources of Emissions and Activity Data

Reasonable default data is provided in the input files for compressor stations in gathering, transmission, and storage. Transmission or gathering compressor models are likely appropriate for most gas processing plant compressors, depending upon size. Each primary data source is discussed in a subsection below. In general, if the user provides user-specific data, that data is used. If the user does not provide data for an element of the model, the provided input data provides reasonable defaults for these values, which are varied via probability distributions between Monte Carlo iterations. Three default cases are currently in the input data set:

- Gathering Emission Factor (GEF) Study.[71]
- Transmission and Storage (T&S) Study.[33]
- EPA Greenhouse Gas Reporting Program (GHGRP) [129]

Note that there is no model for gathering or transmission pipelines, and default data for pipelines is not included.

Gathering Emission Factor (GEF) Study

The Gathering Emission Factor (GEF) study [62, 71] provides data on the compressor driver types, sizes, cylinder counts, and fuel consumption from a representative sample of 435 compressors and related equipment on 180 gathering compressor stations. Data is reported by major equipment unit: AGRUs, compressors, dehydrators, separators, yard piping, and tanks. Yard piping includes all components not on the skids of other major equipment types.

Activity data includes counts of components using the GHGRP classifications from the transmission and storage sector: gas PCs, threaded and flanged connectors, open-ended lines (OELs), valves, meters, pressure relief valves (PRVs), regulators, and gauges. For each type of major equipment, data includes a distribution of component counts in each component category.

Emissions data are divided into compressor and non-compressor components, e.g. there are separate emission measurements for PRVs on compressor skids and PRVs on non-compressor equipment, such as dehydrators, separators, yard piping, etc. Data includes the number of components screened and measured for each component category, the number of leaks found, and an emission rate distribution for measurements. Since the GEF study includes both the (counted and estimated) number of components screened, a leak probability can be calculated for each component emission factor – i.e. compressor and non-compressor components of each type.

As is typical of gathering stations, 95% of compressors screened were driven by four-stroke reciprocating engine drivers and 90% of compressors were reciprocating. Therefore component counts are strong for those type of compressors and drivers, and relatively weaker for centrifugal or screw compressors, and for turbine or two-stroke compressor drivers. The GEF study has no data on compressor uptime, and this data is not reported for gathering to GHGRP. If the user does not specify operating hours 95% uptime is assumed, with off time equally split between NOP and NOD. Default values are found in:

```
input/CompressorModelData/GatheringCompressorDefaultHours.json
```

Combustion emissions are calculated using EPA AP-42 or EPA Subpart C emission factors.

Transmission and Storage Study

The Transmission and Storage (T&S) measured 36 transmission and 9 storage stations [28, 33]. Activity data does not include a count of components per unit of major equipment. Since component counts are needed to determine leak probabilities, GEF data is used to estimate both component counts. Emissions data is reported for compressor and non-compressor components, for the same components as identified for the GEF study above.

T&S includes OP/NOP and NOD hours for compressors for stations which report to the GHGRP augmented by some partner data collected by the study. Utilization varies substantially between compressor and driver types. Additional up-time data for T&S is updated annually by review of the GHGRP reports for reporting stations.

The study included more centrifugal compressors, compressors with turbine drivers, and compressors with two-stroke reciprocating engine drivers. Therefore, it is recommended that T&S data be used if the compressor being modeled falls into any of these categories.

As with GEF, combustion emissions are calculated using EPA AP-42 or EPA Subpart C emission factors.

EPA Greenhouse Gas Reporting Program (GHGRP)

The third source of data for compressor models is the EPA Greenhouse Gas Reporting Program (GHGRP). While these data are included in the current revision of the model, they are not recommended for use on compressor stations; the GEF and T&S data are more specific and better support dynamic compressor models. However, GHGRP reports provide some additional information that may be of interest to users. For the production and gathering sectors, the only information reported under GHGRP Subpart W is: number of compressors per basin (by sector), total vented carbon dioxide per basin, and total vented methane per basin. The vented CO₂ and CH₄ are reported at an emission rate defined in the *Code of Federal Regulations (CFR)* §98.233(o)(10) and §98.233(p)(10). The limited GHGRP information for compressors operating in these sectors lacks data to develop dynamic or time-variant models. Therefore, compressors in the production sector, when modeled using GHGRP data, are treated as constant emitters, i.e. uptime of 100%. In general, users should utilize GEF or T&S data choices for compressors, rather than GHGRP settings. However, default values from the GHGRP may be more appropriate for well-pad compressors.

As noted previously, GHGRP reports the total number of hours per year in OP, NOP, and NOD modes for compressors at T&S stations which report to the GHGRP. These data can be used to update uptime models, if desired. GHGRP also reports driver rated output (in horsepower) in Subpart W, and rated fuel consumption (in mmBTU/hr) in Subpart C, which can be used to populate engine size and determine combustion emissions.

For compressors operating in sectors other than production and gathering, GHGRP mandates measurements of blowdown valves on compressors in OP and NOP modes, and isolation valves on compressors in NOD mode. Rod packing emissions are only measured in OP mode. The compressors are not measured in all modes each reporting year, but are measured in every mode

at least once in three years. Wet seal oil degassing vents must also be measured and reported in OP mode. Notably absent are dry seals; no data exists for dry seal emissions in the data reported under Subpart W. In general, measurements reported in GHGRP could be used to update emission factors, but should be treated with caution, as some approved methods of emissions estimation for valves (e.g. portable acoustic meters) may have large errors [33].

GHGRP reports data separately for reciprocating and centrifugal compressors. However, no data is reported about the driver types. Generally, models assume all centrifugal compressors are driven by turbines, and all reciprocating compressors are driven by reciprocating engines – an assumption which is true in the vast majority of compressors. The specific type of reciprocating engine (two-stroke or four-stroke, lean burn or rich burn) is not reported, and can have substantial impact on combustion emissions. Therefore, if a user does not specify the specific engine type, combustion emissions can only be calculated using Subpart C, which lists a set of emission factors that is not dependent on engine type. Therefore, analysts using the model should acquire other data and properly populate the compressor data to match actual circumstances.

User Specified

If users would prefer to draw from a distribution(s) that more closely resembles their own equipment or equipment in a specific region, this can be accomplished by choosing ‘User specified’ for the compressor data source. A template Excel workbook showing how the data must be formatted for use by the model is provided in:

DES1/input/CompressorModelData/UserCompressorDataDistributions.xlsx

In order to use this template, it must first be copied and renamed; the new file must remain in the same directory as the template. Enter the name of the renamed file into the field ‘Filename for User Specified Data Source’ on the input spreadsheet. User specified data set requires the

use of user-specified leak data. Because of the unique characteristics of compressors, they inherit their leak model from the compressor model that was specified, rather than the leak model that is defined at the facility level. Therefore, if a user does not have adequate leak survey data for compressors, compressors generated using user specified data may not have representative leaks. Note that for the purposes of the model, certain vents (e.g. reciprocating compressor rod packing vents and centrifugal compressor wet seal vents) utilize the leak model because functionally these vents behave as leaks. Therefore, rod packing and wet seal vents may be absent from compressors initialized using user specified data if there is insufficient data to generate them. This model will be replaced in future revisions.

B.2.5 Leak Emissions

This subsection discusses the modeling of fugitive emissions, i.e. leaks.

Leak Model

One leak model is used for all facility types in the MEET, and requires three sets of information:

1. Distributions of component counts for each component category, per unit of major equipment type; typically from field data.
2. The fraction of each component category that is leaking; typically from field data.
3. Distribution of emission rates for *leaking components*, commonly called a ‘leaker factor’ or ‘leaker distribution.’ Note that leaker data *is not* the same as *average* or *population* emission factors.

Empirical distributions are used for both component counts and emission rate measurements for leaking components.

Component Counts: For a given unit of major equipment (e.g. a compressor), the model randomly assigns component counts for each component type. For one component category (e.g. threaded connectors), each unit of major equipment, k , in each Monte Carlo iteration, m , is assigned $n_{c,km}$ components drawn from an empirical distribution of component counts for that type of major equipment.

Component count distributions are derived from published field studies that counted components for facilities or units of major equipment or from similar sources of data. If component counts are not available for a sector, counts are assumed from studies of other sectors (see Sections B.2.5 through B.2.5).

Leak Initialization: At the start of the simulation run, the model assumes that the fraction of components leaking in each component category equals the fraction seen leaking in field data used by the model. A typical field study screened n_f components in one component category, and found n_l leaking (or emitting). Screening may have been performed on one type of major equipment (e.g. ‘compressors’) or on a classification of major equipment that includes multiple major equipment types (e.g. ‘non-compressor’ equipment).

Therefore, field data estimates the leak probability as:

$$\mathcal{P}_f = \frac{n_l}{n_f} \quad (\text{B.6})$$

Leaks are created by drawing a vector of uniform random numbers, $\mathbf{r} = \{r_i, i = 1 \dots n_{c,km}\}$ and comparing it to the leak probability. For each unit, on each Monte Carlo simulation, the number of simulated leaks is:

$$n_{s,km} = \sum_{i=1}^{n_{c,km}} \langle r_i < \mathcal{P}_f \rangle \quad (\text{B.7})$$

where the expression in angle brackets is zero if false, and one if true, and $\mathcal{P}_f$ is the appropriate leak probability for the major equipment type of k . Over M Monte Carlo simulations and K units of major equipment, the algorithm assures that the probability of a component leaking converges to that seen in the field data. For each component category, the number of leaking components is:

$$N_s = \sum_{m=1}^M \sum_{k=1}^K n_{s,km} \quad (\text{B.8})$$

and the total number of simulated components is

$$N_c = \sum_{m=1}^M \sum_{k=1}^K n_{c,km} \quad (\text{B.9})$$

and the simulated leak probability is

$$\mathcal{P}_s = \frac{N_s}{N_c} \quad (\text{B.10})$$

Over sufficiently large value of $M \cdot N$, the algorithm converges to the leak probability target, that is, $\mathcal{P}_s \rightarrow \mathcal{P}_f$ as $M \cdot N \rightarrow \infty$. In practice, this assignment converges to acceptable error differences when $M \cdot N$ exceeds 10,000, although practical results are produced by as little as $M \cdot N > 1,000$.

Each leaking component is assigned a leak rate by drawing randomly from an empirical distribution of leak rates. There is no variation in the leak rate with simulation time; a leaking component's emission factor will remain the same until the leak has been repaired.

Component count distributions (i.e. activity) and emission distributions must correspond to the same definition of activity. For example, multiple PRVs on a compressor may be vented separately or may be plumbed into a single vent line. Studies may vary on whether all vents on a single compressor are treated as one leaking component (i.e. the 'component' is the combined vent line) or as individual PRVs. Therefore, the emission rate distribution must correspond to the activity definition – either 'emissions per compressor' with one 'component' per compressor, or as 'emis-

sions per PRV on a compressor' with (potentially) multiple PRV components instantiated on each compressor.

Leak Generation and Repair: Additional leaks will occur after the start of the simulation. These leaks are generated when the simulated units for the Monte Carlo run are instantiated. At the start of the simulation, these leaks are in a non-leaking state, and are set to switch states to 'leaking' at some later time.

To simulate these leaks, an assumption must be made on the frequency at which leaks appear on the simulated equipment. Most public data originates from field campaigns that screened for leaks at many field locations, and in the future, these may be augmented by company LDAR records. Operators and facilities may differ in how often leak surveys are performed. Therefore, the leak generation rate seen in the field campaign is a function of:

1. The failure rate of each component category screened.
2. Frequency of surveys and/or the time since the last survey when the field measurements were conducted.
3. The fraction of leaks detected in a survey which are fixed before the next survey.

In general, these qualifying data are not available for published field campaigns. Therefore, the model makes several assumptions to produce a realistic leak generation rate.

First, the model assumes the field study facilities were subject to periodic LDAR surveys, and that all detected leaks were fixed shortly after they were detected, and were definitely fixed prior to the next survey. Leaks detected during any survey therefore reflects the number of leaks which have appeared during the survey period.

Since most field campaigns screen and measure facilities at an arbitrary time during the year, on average the field campaign encountered the facility halfway between periodic screens. Therefore,

if regular leak surveys have been performed on the study locations, the field campaign sees the leaks which started in $\frac{1}{2}$ of the survey interval. If there are no regular surveys, the field campaign sees the long-term ‘stable’ leak frequency.

Unfortunately, data about leak survey frequency for studied facilities is not typically published with field studies. Therefore, the model assumes that the leak frequency encountered in the field study is equivalent to the new leak generation rate on an annual basis – i.e. additional leaks are generated at the rate of $\mathcal{P}_f$ per simulation year. Future model revisions may allow survey frequency to be adjusted, for example, if operator LDAR data is available.

Second, in the default case, where leak repair actions are not independently simulated, the repair action identified above must also be simulated. As a default, the model assumes that leaks are repaired at the same rate as they are generated, i.e. at a frequency of $\mathcal{P}_f$ per simulation year. Since repair actions are independent of leak-start actions, the number of leaks will rise and fall, but will, in general, average out to $\mathcal{P}_s \rightarrow \mathcal{P}_f$ over a multi-year, multi-iteration, Monte Carlo simulation.

Leak Initiation Process Summary: Leaks are modeled using the following steps:

1. Create all compressor stations in the simulation
2. Populate all major equipment units on each station
3. Create initial leak population; all leak emitters are actively emitting (in the appropriate compressor operating state(s)) when created
 - (a) Estimate component count for each component category from activity distributions
 - (b) Draw number of emitters for the component count
 - (c) For each emitter, draw emission rate from appropriate emission distribution
 - (d) Randomly assign the emitters to major equipment units

4. Create additional leak population, i.e. leaks which will start later, as described in the leak model.

Individual components are not tracked. For example, if a compressor has 10 valves, each valve is not instantiated and tracked through the simulation. However, if 2 of the valves are initially leaking, 2 valve leaks are generated and tracked with separate IDs. During a simulation, one valve leak could be repaired, the other valve leak left unrepaired, and, a probability exists that a third valve could start leaking at some later point during the simulation.

With this implementation, it is possible for a single component to leak, be repaired, and begin leaking again, potentially multiple times depending on the length of the simulation and the component's likelihood of developing a leak. In this case, the component will randomly draw a new emission rate every time it begins leaking, and this emission rate will remain constant until the leak has been repaired.

All leaks are assumed to be steady state once started. For leaks on compressors, each component category is assigned to specific operating modes and are emitting in some modes and not emitting in other modes, as shown in Table B.2.

The following subsections describe the data available from the primary studies used in the model. Different studies publish different subsets of these data. For compressor stations, the GEF study provides all three components, while the T&S provides only the leaker emission rate distribution. The data source for compressor station leaks is specified by the '*Leak Emission Data Source*' parameter on the `Facility_List` sheet of the data input spreadsheet. Since compressor leaks tend have different leak probabilities and emission factors compared to non-compressor leak behavior is defined by the 'Data Source' parameter on the Compressor sheet. This may or may not be the same as the data source used for all other leaks at the facility.

Gathering Emission Factor (GEF) Study

The study defines six categories major equipment:

- Acid Gas Removal Units (AGRUs)
- Compressors
- Dehydrators
- Separators
- Yard piping
- Tanks

The study provides a distribution of counts for each major equipment category, based upon component counts collected during the field campaign (AGRUs had very small sample size). Component counts are drawn from this distribution for each unit of major equipment instantiated in the model.

The GEF data also reported the number of screened components and the number of detected emissions for each component category. For common component categories, emission distributions are divided into compressor and non-compressor distributions; in general, the frequency of leaks and the size of leaks is higher for compressor-mounted components.

Two component categories had no reported leaks in the GEF study: gauges and meters. Therefore, while gauges and meters are assigned to units of major equipment, their failed fraction is zero (number of detected emissions was zero), and therefore cannot develop leaks.

Transmission and Storage Study

The study provides leak measurements organized into leaker distributions, with both compressor and non-compressor component classifications. Neither component counts nor number of screened components are provided; the model uses GEF data for these distributions.

The model uses the GEF component categories. For component categories that are missing in T&S, emission rates are taken from the GEF emissions data. T&S did not distinguish between threaded and flanged connectors; therefore, both connector types use the same emission distribution. Similarly, meters and PRVs were combined in T&S, and use the same emission distribution. However, since the leak frequency is set by GEF study data, and GEF found no meter leaks, no leaks will be generated on meters when using the default GEF or T&S leak models. Emissions from leaks on regulators were not measured or reported in T&S; therefore, regulator leaks use GEF leaker emission factors.

EPA Greenhouse Gas Reporting Program (GHGRP)

In general, GHGRP emission and activity factors are not encouraged, as they have been superseded by more recent studies. However, they are included in the current model for comparison. Table W-1B to Subpart W of Part 98 defines default average component counts for major equipment applied to onshore natural gas production and gathering sectors. These data originated in a 1996 study that counted components on well pads and a few units of midstream equipment. The GHGRP program then applied these counts to the gathering and boosting sector. Table W-1B defines counts for the major equipment categories:

- Wellheads
- Separators
- Meters/piping

- Compressors
- In-line heaters
- Dehydrators

Four component categories have counts in Table W-1B:

- Valves
- Connectors
- Open-ended lines
- Pressure relief valves

Only these component categories will generate leaks when using GHGRP data. Only mean values are provided; there is no distribution of counts or emission rates. GHGRP reports include leak counts. However, these data need to be screened to select only reports which include reporter leak survey data, and exclude those reports where the leak count was determined using the defined leak fractions. This varies by sector (T&S requires surveys, gathering does not).

User Specified

The final data source for leaks is user defined. This option is provided for users that wish to use their own leak survey data to determine failed fraction and leaker emission factors, rather than relying on one of the previously described models. Note that at present, this method does not allow the user to specify component counts for major equipment. If a user specified leak model is used, Table W-1B is used to assign component counts. This limits the model to the major equipment and component categories included in Table W-1B.

A Microsoft Excel template for providing user defined leak data is provided. The filename for the template is:

`DES1/inputs/LeakModelData/UserInput_ProductionLeaks.xlsx`

Instructions for using the template can be found on the first sheet, titled ‘READ ME’, of this workbook. If users are planning to provide their own leak data, it is recommended that this file be copied and renamed before editing. The newly created file must remain in the same directory. To use this data, select ‘User specified’ as the Leak Emission Data Source for the appropriate facility(ies), and enter the name of the newly created spreadsheet under ‘Filename for User Specified Leaks’ on the `Facility_List` sheet of the input data workbook.

B.2.6 Start and Blowdown Episodic Venting

Emissions from gas starters and unit or station blowdowns occur episodically during transitions between operating modes (Figure 2, main paper), and are typically high-rate, short-duration, emissions. Timing of these episodic events is included in the compressor transition matrix (Section B.2.2). An emission quantity – either rate or total emissions – and duration of the release is required for a time-domain simulation.

Gas starters: A gas starter uses an expansion turbine to crank an engine. The turbine operates on the pressure of the gas supply and does not burn the gas, and gas is directly vented to atmosphere through an open-ended line. None of the studies measured emissions from gas starters when actively starting a compressor. Therefore, engineering estimates based upon gas starter specification sheets have been used for the default data set. The default distributions are:

Table B.4: Gas starter default distributions

Gas Starter Vents	Distribution Type	Lower	Mean	Upper
Duration (s)	Triangular	5	30	50
Total gas release during start (scf)	Triangular	500	3,800	5,000
Mean emission rate (scfh)	N/A	600	7,600	60,000

The emission rate is assumed constant for the duration, and is defined as total emissions divided by duration. Gas composition is the composition assigned to the station. A random duration and total gas release are drawn for each compressor at instantiation. The same compressor will emit the same total volume of gas every time it transitions into OP mode.

In practice, several tries may be required to start an engine (dotted lines in Figure 2, main paper). However, since no data is available for ‘failed starts,’ and these are not included in the current model.

Unit and station blowdowns: For compressor stations, compressor unit blowdowns occur every time the compressor transitions into NOD mode, which occurs at intervals ranging from hours to weeks. Partial or complete station blowdowns occur infrequently for maintenance activities.

For processing, transmission, storage, and gathering sectors, the total number of blowdowns and total emissions from blowdowns are reported annually in Table I.1 and Table I.2 of the GHGRP [123]. These data provide estimates for CO₂ and CH₄ per blowdown for these sectors. However, the data do not clearly define the reason for the blowdown (e.g. compressor unit, station, or some other subset of equipment at the station). As a result, blowdown frequency and size are not well documented for station piping or other pressurized major equipment units (separators, dehydrators, AGRUs).

Based upon the available data, unit blowdowns are included in the current model. These blowdowns are relatively frequent and are coordinated with other compressor operations. Station and

other equipment blowdowns are not included; these blowdowns are less frequent and solid data for timing and size of these emissions is not readily available.

In practice, blowdown emissions are typically estimated using starting and ending pressures and the volume of piping and vessels blown down. In order to be used as part of a time domain simulation, both the size and duration of the blowdown must be modeled. Since blowdown durations are not recorded in any of the available data sources, example data uses engineering estimates and the study team’s observations of blowdowns while on site. If no data is provided, the model defaults to data from the GHGRP table “EF_W_BLOWDOWN_STACKS_UNITS”.

The default distributions are listed below. Generally, an estimate of the total gas release (scf) should be developed by the analyst from the size of the blowdown volume and typical operating pressure of the unit.

Table B.5: Compressor blowdown default distributions

Compressor Unit Blowdowns	Distribution Type	Lower	Mean	Upper
Duration (s)	Triangular	30	450	1,800
Total gas release during start (scf)	Triangular	1,500	2,200	2,900

As with gas starters, a random duration is drawn for each compressor, and the total gas release is taken from user input or drawn from the default distribution when the compressor instantiated. The same compressor will emit the same total volume of gas every time it transitions into NOD mode. The blowdown rate is assumed constant in the current model; the rate is the total emission divided by the duration.

Any of these underlying assumptions can be modified by editing the file:

```
input/CompressorModelData/Start_and_blowdown_parameters.json
```

B.2.7 Non-Episodic Vented Emissions

Non-episodic compressor vented emissions include rod packing and centrifugal compressor seals as well as PCs, which were discussed earlier (Section B.2.1). Rod packing and seals vent only in OP mode and are taken from prior studies or the GHGRP and are assumed steady state per compressor. In other modes, emissions are considered leaks (Table B.2).

Compressors vary in how rod packing and seal vents are connected to vent stacks, and are often plumbed together before routing to a vent stack. In both the GEF and T&S studies, when there were multiple vents on a compressor unit they were combined to generate a single emission rate for the unit. Therefore, the emission distribution is for one vent of these types on each compressor. In contrast, pocket vents (reciprocating compressors only) are typically not plumbed together, and are tracked separately based upon the count of compressor cylinders.

As with other categories, the availability of data depends upon the sector:

- *T&S*: Zimmerle et al. [33] provides emission distributions for both wet and dry seals and for rod packing.
- *Gathering*: Emission factors for reciprocating compressors are available from the GEF study [61, 62]. Too few centrifugal units were measured to update emission distributions for those compressors; emissions are modeled using data from the T&S study [33].
- *Processing plants*: No updated values are available; current model uses GHGRP factors.

For rod packing, separate emission distributions are available for emissions in OP, NOP and NOD modes. Rod packing emissions for each mode are drawn when the compressor is instantiated, and are the same each time a compressor is in that mode. However, like all leaks, rod packing may begin the simulation with no leak and then, at some point during the simulation, transition to leaking, after which the emission rate is constant for the remainder of the simulation. Mode-

dependent data is not available for centrifugal compressor seals; these are modeled to emit only in OP mode.

B.2.8 Compressor Driver Combustion Emissions

Exhaust gas composition is modeled using either AP-42 or Subpart C emission factors. Emissions are computed as in Zimmerle et al. [33] and Vaughn et al. [45]. In brief, the loading of each compressor is drawn from a distribution (see input data files) and used to compute the energy content of the fuel burned in the compressor. The discrete event simulator outputs this value to intermediate files. After time simulation, the model combines fuel gas composition and AP-42 (or Subpart C) factors to compute exhaust emissions. Required input data is described in the guide to input spreadsheets.

AP-42 provides emission factors for 2-stroke and 4-stroke, rich and lean burn, internal combustion engines and turbines. Subpart C has a single set of emission factors regardless of engine type, and is therefore less accurate than AP-42 for engines.

Combustion emissions are determined based on driver efficiency, driver size, and the driver loading.

Driver loading is chosen randomly from a distribution each time the compressor enters the OP state. This load level, as a percent of rated power, is assumed to be constant for the entire duration that the compressor is in the OP state. If loading is not specified by the user, default values are selected based on industry sector. Loading is specified in ‘per unit’ (*p.u.*) values, which is the ratio of operating load divided by rated power, and expressed as a percent:

$$P_{p.u.} = \frac{P_a}{P_r} \quad (\text{B.11})$$

where P_a is the actual power output of the engine and P_r is the rated power.

Default values vary by sector. For production and processing, compressors are assumed to be operated at rated capacity (100%). For gathering, loading is estimated to be normally distributed with a mean of 83% of rated load and a standard deviation of +/-8.9% rated load, based upon the GEF study. For the transmission sector, the default loading is assumed to be normally distributed with a mean of 90% and a standard deviation of +/-1% rated load. The default loading values can be changed by modifying the file:

input/CompressorModelData/DefaultLoading.json

The actual fuel flow for reciprocating engine drivers is calculated based on the loading and the rated fuel flow at full load:

$$Q_a = \begin{cases} (-0.4P_{p.u.} + 1.35)Q_r & P_{p.u.} < 0.75 \\ (-0.2P_{p.u.} + 1.20)Q_r & P_{p.u.} \geq 0.75 \end{cases} \quad (\text{B.12})$$

Where Q_a is the actual fuel flow rate, $P_{p.u.}$ is the relative driver loading as defined above, and Q_r is the rated fuel flow at full rated load. Using data from the EPA GHGRP, fuel rates, Q_a and Q_r , are expressed in BTU/bhp-hr.

For compressors with turbine drivers, the fuel flow rate is determined using the following equations.

$$Q_a = \begin{cases} (-1.12P_{p.u.} + 1.98)Q_r & P_{p.u.} < 0.75 \\ (-0.56P_{p.u.} + 1.56)Q_r & P_{p.u.} \geq 0.75 \end{cases} \quad (\text{B.13})$$

For a known loading, the energy input rate can be computed using the loading, fuel rate, and rated power, and then converted to kJ/s or kW:

$$F = P_{p.u.} \cdot P_r \cdot Q_a \cdot \frac{1.055kJ}{BTU} \cdot \frac{hr}{3600s} \quad (B.14)$$

EPA AP-42 emission factors are given in rates of pounds per million BTU (lb/mmBTU), and EPA Subpart C combustion emission factors are given in rates of kg/mmBTU. These are converted into rates of kilograms per kilojoule (kg/kJ), and multiplied by the fuel input rate, F , to compute emissions.

B.2.9 Simulation Examples

This subsection discusses the simulation examples utilized in the main paper.

Station Characteristics

The test data set consists of 8 stations modeled after equipment seen in a recent basin-scale study [45], and data from a recent national measurement campaign at gathering stations [71]. While compressor size and drivers are taken from the study, operating characteristics of the stations were modified to provide a range of different conditions, shown in Table B.6.

Table B.6: Key characteristics of stations in test data set

ID	Sector ¹	Description	Number of Compressors	Typical Number of Compressors Offline	Compressor Size (HP)
1	T	Small station, typical operating hours	2	NA ²	4890
2	G	Low average loads per compressor	8	1	4-1340, 4-1775
3	G	High loads, typical downtime	5	0	1775
4	G	Small, very high loads, low downtime ³	2	0	1775
5	G	Small, very low loads, low downtime ³	2	1	1775
6	G	Large, all 90% avg load, mixed downtime	24	4	4-1340, 20-1775
7	G	Small, no engines ever run	2	2	1340
8	G	Typical size, typical downtime	7	3	4-1340, 4-1775

¹ T indicates transmission, G indicates gathering.

² Simulations use default annual hours from T&S study[33]

³ Stations 4 & 5 have the same typical annual operating hours.

Table B.7 summarizes all sources utilized in the simulation. The table notes the name of the source, as it appears in the data files, charts, and other tables, and the classification of emissions from that source. See notes in the table for sources which may be either fugitive or vented emissions.

Table B.7: All sources modeled in simulations

Source	Emission Category		
	Combusted	Fugitive	Vented
Combustion Sources			
Driver Exhaust	Y		
Still Vent	Y		
Component Leaks			
Connector Flanged		Y	
Connector Threaded		Y	
OEL		Y	
PRV		Y	
Pocket Vent		Y	
Regulator		Y	
Valve		Y	
Major Component Leaks and Vents			
Blowdown Valve Leak		Y	
Isolation Valve Leak		Y	
Rod Packing ¹		Y	Y
Single Unit Vent		Y	
Starter Valve Leak		Y	
Thief Hatch Leak		Y	
Episodic Vents			
Blowdown or Start			Y
Pneumatic Controllers			
High Bleed PC			Y
Intermittent PC ²		Y	Y
Low Bleed PC			Y

¹ Rod packing emissions are classified as vents when the compressor is operating (OP mode), and as fugitive emissions when it is not (NOP and NOD modes).

² The intermittent PC model used in this study simulates both properly functioning and abnormally functioning controllers. Each controller is randomly assigned a profile, and will transition between normal and abnormal operations randomly.

Simulation Runs and Timing

To illustrate behavior of the model only a small number of Monte Carlo iterations – 20 – are included in the simulation. After simulation, results were summarized at the daily level for the entire year (365 time steps), and at one-minute resolution for 26 random, half-day, periods throughout the year, as shown in Table B.8.

Table B.8: Timing of 1-minute Extracts

Set	Day
1	15
2	23
3	37
4	59
5	64
6	70
7	71
8	80
9	147
10	159
11	170
12	182
13	198
14	208
15	222
16	225
17	246
18	248
19	251
20	253
21	263
22	287
23	302
24	306
25	327
26	339

Inserting Large Emitters

Exceptionally large emitters seen in field campaigns – commonly called ‘super-emitters’ – are not included in the emission distributions from the GEF or T&S studies. The intent is for these emitters to be handled as a separate category. As they are rare, there is anecdotal evidence that they have a different life cycle than smaller emissions, and appear randomly on various component types. Since it was desirable to have these emitters visible in some simulations, particularly for the gas composition simulations, the GEF distributions were modified to cause large emitters to occur at frequencies seen in the GEF field campaign.

In the GEF field campaign, the 435 compressors that were screened and measured exhibited 5 large emitters (1.1%), 6 large emitters of 251 screened and measured tanks (2.3%), and 3 large emitters on other equipment on the 176 stations (1.7%). Frequencies are given in Tables S3-36 and S3-37 and the emitter size is estimated in Table S3-38 of Zimmerle et al. [62]. Only the distribution for single-unit vent was utilized to extend distributions for the example problem.

For tanks, the distribution was extended to include a 1.4% probability of leak with a mean of 3675 scfh whole gas, while for compressors the distribution was extended to include a 1.1% probability of a leak with a mean of 2451 scfh whole gas.

The large emitters resulting from the above modifications, and other similarly sized emitters, are summarized in Table B.9, which shows all emitters with THC emissions ≥ 10 kg/h whole gas

Table B.9: Fugitive Emitters In Sample Iterations with Emissions ≥ 10 kg/h THC

Station ID	Major Equipment Type	Source Type	Monte Carlo Iteration	Start Day	End Day	Maximum Emission Rate (kg/h)
2	Tank	Single Unit Vent	11	137	364	182
6	Compressor	Starter Valve Leak	8	1	364	125
2	Compressor	Starter Valve Leak	6	1	54	116
6	Compressor	Starter Valve Leak	20	1	57	114
6	Tank	Single Unit Vent	11	62	364	103
8	Compressor	Starter Valve Leak	4	30	249	103
2	Compressor	Starter Valve Leak	17	72	162	99.9
2	Tank	Single Unit Vent	1	130	364	76.8
4	Tank	Single Unit Vent	8	128	364	48.9
2	Compressor	Valve	1	7	115	17.4
6	Compressor	Valve	20	1	325	17.1
1	Compressor	Isolation Valve Leak	1	1	145	16.8
8	Compressor	Valve	5	226	283	16.7
6	Compressor	Valve	2	1	229	15.8
5	Compressor	Valve	4	1	249	14.7
2	Compressor	Valve	12	1	35	14.5
8	Compressor	Valve	6	22	332	14.5
1	Tank	Single Unit Vent	18	1	276	14.1
1	Tank	Single Unit Vent	19	1	364	14.1
1	Tank	Single Unit Vent	14	343	364	14.1
1	Tank	Single Unit Vent	1	1	342	11.5
1	Tank	Single Unit Vent	14	1	109	11.5
1	Tank	Single Unit Vent	17	201	364	11.5
6	Compressor	PRV	16	293	364	11.3
2	Compressor	Valve	20	356	364	11.3

Tracer Simulation

The tracer method, described briefly in the main paper, makes several transects downwind of a facility, during which time high-resolution concentration measurements are made of the target analyte (methane and ethane in this case), and the tracer gas(es) being emitted on site at known rates (nitrous oxide and acetylene are often used). Each downwind transect consists of a time series of concentration measurements starting at approximately background conditions, extending through the downwind plume of emissions from the site, and ending when the concentrations again revert to background levels. Using various techniques [27, 30], the concentration profiles are utilized to estimate an emission rate at the site, commonly called a *plume*.

All of the complexity of a tracer flux measurement [27] cannot be captured in the simulation performed here. However, it is possible to capture the essential dynamics of the method. The simulation utilized the following assumptions:

- Emissions are estimated at each site, for each MC iteration, using 2 to 8 plumes, uniformly distributed.
- Each plume is the average of 3 minutes of total emissions from the site. This is longer than an actual plume transect, but accounts for the smoothing effect of dispersion and mixing between the site and the transect.
- A gap of 5 to 15 minutes between plumes, uniformly distributed. In field conditions, this time is required to drive the measurement unit to safely reposition the measurement unit before making another transect. Additionally, since some plume are unsuccessful, the spacing between plumes is variable.
- The standard deviation of the error in each plume estimate is 15% taken from field testing of the tracer method [27, 30].

The simulation completed 500 MC iterations focusing only on methane emissions. For each iterations, the process was:

1. Draw random start time for the tracer measurement (i.e. an offset into the 1-minute emissions time series for the site).
2. Draw the number of successful transects – i.e. the transects with that meet acceptance criteria and can be classified as ‘plumes’ – uniformly distributed on range of 2 to 8 transects.
3. Randomly select which transects were successful plumes.
4. Draw uniform random starting time offsets, relative to the start time in (1), for each transect. This offset is uniformly distributed over 5 to 15 minutes.
5. Compute average emissions over 3 minutes for each transect.
6. Discard all unsuccessful transects using vector generated in (3).
7. Apply random errors using the standard deviation of errors for each successful transect. The result is a set of 2 to 8 plumes with normally distributed errors.
8. Compute mean of plume estimates as the measurement estimate.

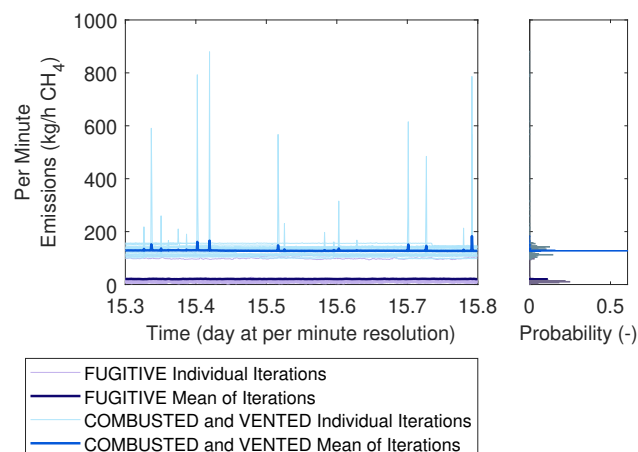
The result was a family of 500 tracer estimates for each $\frac{1}{2}$ day of 1-minute data samples, for each station. Table B.10 summarizes the confidence intervals and mean error for these estimates by station. Center columns identify the fraction of tracer estimates which were outside the confidence interval of emissions for the station, using daily total emissions for an entire year. Note that the set of compressor stations simulated here are not representative of the national mix of compressor stations (see Section B.2.9), and may not represent method accuracy on a regional or national basis.

Table B.10: Tracer Simulation Results

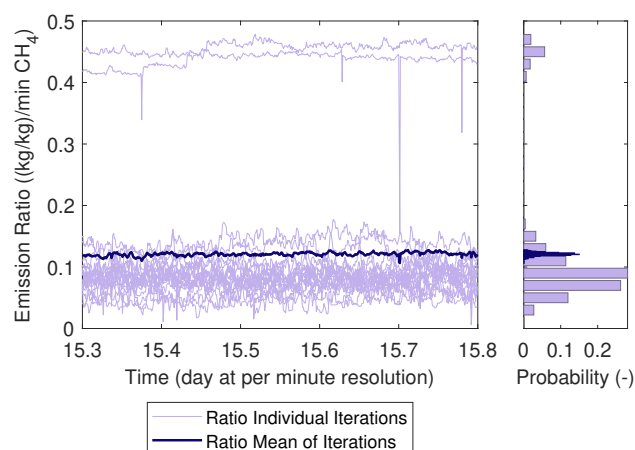
Station ID	Fraction Outside Daily CI	Fraction Above Upper Daily CI	Fraction Below Lower Daily CI	Mean Tracer Error
1	38%	11%	27%	0.31%
2	29%	3.6%	25%	0.46%
3	36%	7.8%	28%	0.55%
4	24%	3.6%	20%	-0.3%
5	30%	5.9%	24%	-0.92%
6	34%	3.6%	31%	-0.84%
7	11%	3.8%	6.8%	0.9%
8	36%	3%	33%	-2%
All	28%	7%	21%	-0.5%

Continuous Monitoring Simulation

Continuous monitoring solutions were simulated utilizing a simplistic model that includes knowledge of both total emissions (fugitive + combusted + vented) and an estimate of vented + combusted emissions. We assume the monitoring solution makes some type of measurement of facility emissions each minute (emissions were extracted at 1-minute resolution). The solution also computes the current vented emission rate of the station. In actual practice, both calculations have error – often substantial. Autonomous downwind emission measurements often have measurement errors in excess of $\pm 50\%$, while it is difficult to quantify vented *or* combusted emissions over a 1-minute period.

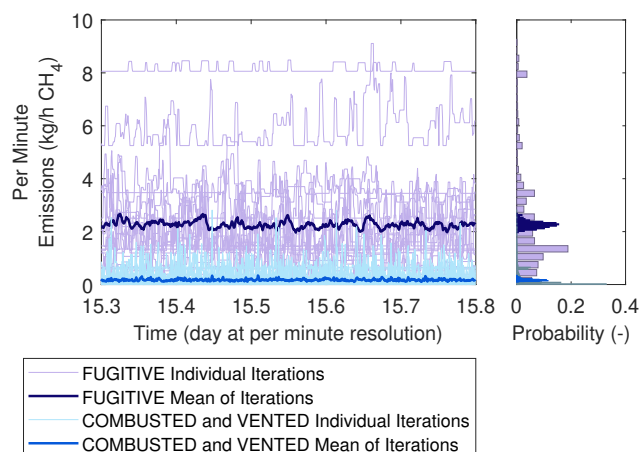


(a) Emissions time series for Station 6. Left panel: Time series at 1 minute resolution. Blue is the sum of all vented and combusted emissions while purple is fugitive emissions, barely visible at bottom of plot. Lighter lines represent individual Monte Carlo iterations; dark is the mean across all iterations, at each time step, for both types of emissions. Right panel: Histogram of 1-minute emissions for the same time period with the same color coding.

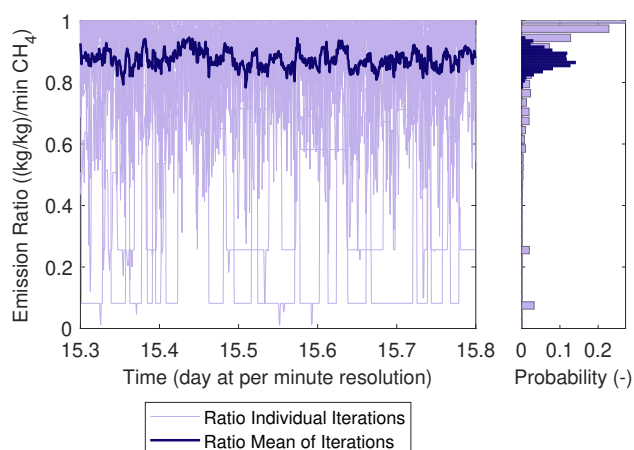


(b) Ratio of fugitive to total emissions for Station 6. Left panel: Time series at 1 minute resolution. Lighter color represents individual Monte Carlo iterations; dark is the mean across all iterations, at each time step. Right panel: Histogram of 1-minute ratios for the same time period with the same color coding.

Figure B.5: Continuous monitoring simulation for Station 6. Panel (a) displays simulated emissions, while panel (b) displays the ratio fugitive to total emissions. Station 6 is large station where one or more compressors operates at all times. Fugitive emissions are a small fraction of total emissions, and while the mean ratio hovers around 0.15, the mean is dominated by two Monte Carlo iterations with large fugitive emissions, while remaining iterations exhibit fugitive emissions that are low relative to combustion slip.

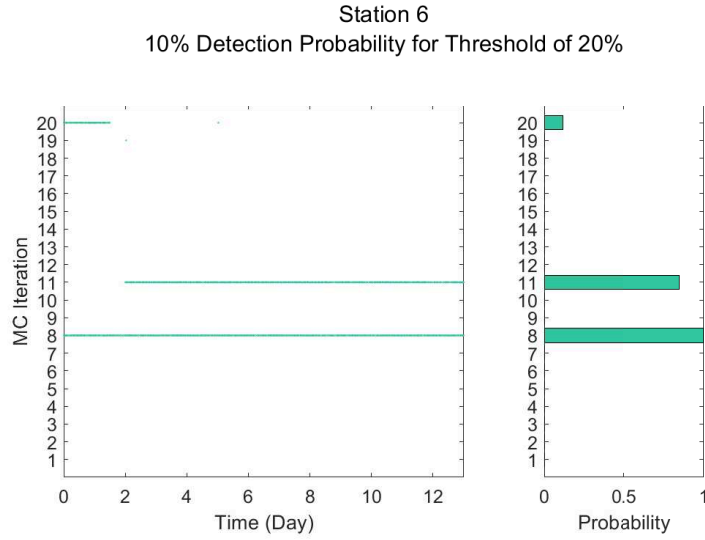


(a) Emissions time series for Station 7. Left panel: Time series at 1 minute resolution. Purple is fugitive emissions while blue is the sum of all vented and combusted emissions, barely visible at bottom of plot. Lighter lines represent individual Monte Carlo iterations; dark is the mean across all iterations, at each time step, for both types of emissions. Right panel: Histogram of 1-minute emissions for the same time period with the same color coding.

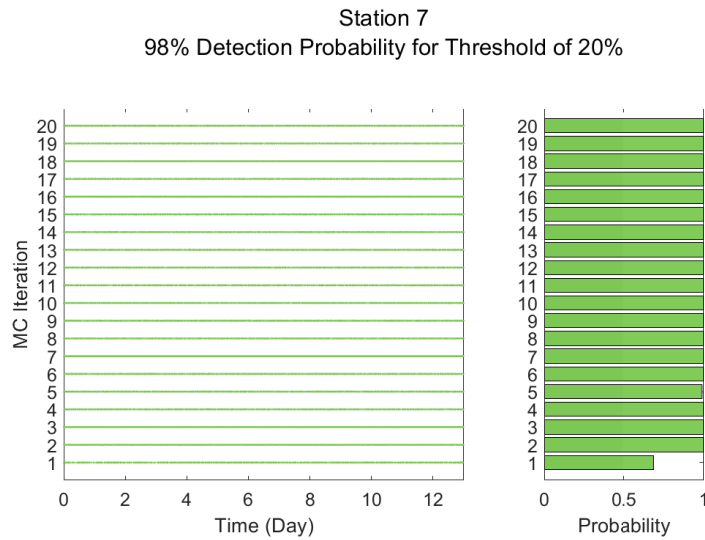


(b) Ratio of fugitive to total emissions for Station 7. Left panel: Time series at 1 minute resolution. Lighter color represents individual Monte Carlo iterations; dark is the mean across all iterations, at each time step. Right panel: Histogram of 1-minute ratios for the same time period with the same color coding.

Figure B.6: Continuous monitoring simulation for Station 7. Panel (a) displays simulated emissions, while panel (b) displays the ratio fugitive to total emissions. Station 7 has no operating compressors. Since remaining gas pneumatic emissions make up the bulk of vented emissions, fugitive emissions represent a large fraction of total emissions.



(a) Detection pattern for station 6



(b) Detection pattern for station 7

Figure B.7: Detection pattern for Stations 6 & 7, for a continuous monitor capable of detecting fugitive emissions in excess of 20% of total emissions. These two stations represent extremes of detectability with continuous monitors. Right panel shows a composite of all 1-minute simulation periods. Left shows the fraction of time, as a probability, that fugitive emissions are detected. Positive detection periods for Station 6 correspond to MC iterations where the station has large emitters (see Table B.9) with emissions larger than 100 kg/h.

Appendix C

Real-Time Distributed Simulation

C.1 Precision Timing on Linux Microcomputers

Precision Timing on Low-Cost Linux Microcomputers

Gerald P. Duggan
Electrical and Computer Engineering
Colorado State University
Fort Collins, CO
Email: dugganj@engr.colostate.edu

Peter Young
Electrical and Computer Engineering
Colorado State University
Fort Collins, CO
Email: pmy@engr.colostate.edu

Abstract—Recent years have seen a significant number of low-cost (sub-\$100) Linux-based microcontroller systems introduced to the market. Initially targeted to the educational and hobbyist markets, the low cost, small physical package, low power consumption and high compute capability of these microcontrollers has made them interesting to research and technical communities. Many of the applications these platforms are being used for require precise timing to measure physical phenomena and coordinate with other computers.

This paper examines the timing characteristics of these systems. Jitter for a periodic task is defined as the difference between the desired time interval for the task and the time the task actually took. The jitter of a sleep task is examined under a number of system variables, including system type, load average, and desired target interval.

The paper then uses a feedback control algorithm to improve jitter using a Proportional-Integral (PI) controller. Results show the jitter reducing from 90 μ seconds to just under 5 μ seconds using this controller on a typical low-cost Linux-based microcontroller.

I. MOTIVATION

There has been a dramatic proliferation of low-cost Linux microcontrollers in recent years, including the Raspberry Pi [5], the Beaglebone series [1], and the Pduino [8]. Initially targeted at educational use, these platforms are making inroads into Wireless Sensor Networks (WSN) and other cyberphysical systems. Low system cost, compact physical design, and minimal power consumption are three reasons why these platforms are of interest to the technical community.

Precision timing, the ability to perform periodic computations on a regular basis with small variations in the time period, is critical for many WSN applications. Measuring physical phenomena often needs high speed sensing with predictable error margins. Similarly, real-time simulations require precise time measurements. A common technique to extend the battery life of sensor nodes in a WSN is to have the nodes sleep for most of their duty cycle, waking periodically to perform measurements and communicate with other nodes. Improving the precision of timing on such nodes will increase the accuracy of their measurements and improve their ability to extend batter life.

A. Related Work

Linux has had real-time capabilities since at least 2007 [9], and jitter in the Linux RT patch has been examined by

[7]. Real-time scheduling using modern feedback controller theory is proposed in [6]. A variation of this approach was examined in [4], where a user-space jitter control system was proposed. This work builds on [4], implementing a user-space jitter control algorithm on low-cost Linux microcontrollers, using modern Linux high-resolution timing primitives.

II. RESULTS

The target platform for this experiment is a Beaglebone Black, available for approximately \$55.00 at many locations on line [1]. For comparison, the experiment was also run on a server class machine. Details for the platforms is shown in table I.

TABLE I
TEST PLATFORMS

Platform	Processor Speed	Memory	Cores	Clock Resolution	Cost
Beaglebone	1.0 GHz	512KB	1	1 ns	\$55
Server	2.0 GHz	16 GB	6	1 ns	\$4500

A high level block diagram of the control system is presented in figure 1. Algorithm 1 is the method used to calculate jitter. The algorithm uses high resolution timers available in Linux to get the current time and sleep to the clock resolution of the machine, 1 nanosecond [3]. Adjustment of the load average is performed using the Linux *stress* utility [2].

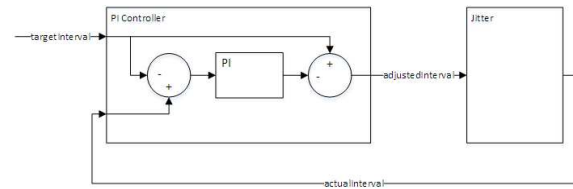


Fig. 1. Block Diagram of Control System

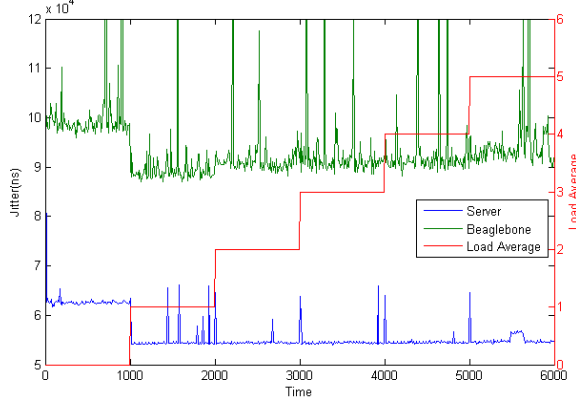


Fig. 2. Baseline Data

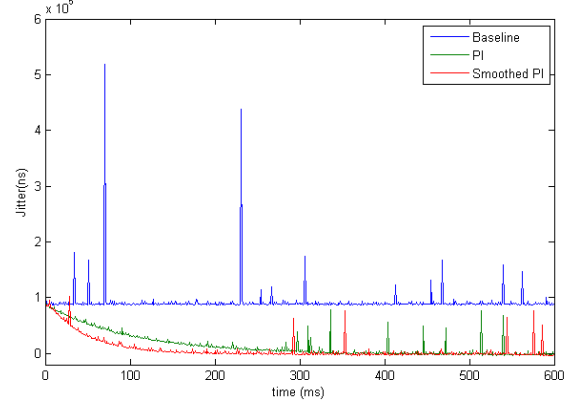


Fig. 3. Jitter for 1 millisecond target interval, load avg = 1

Algorithm 1 Jitter calculation algorithm

```

1: function JITTER(interval) ▷ Specified in ns
2:    $t_0 \leftarrow \text{current time in ns}$ 
3:    $\text{nanosleep}(\text{interval})$ 
4:    $t_1 \leftarrow \text{current time in ns}$ 
5:    $\text{actualInterval} \leftarrow t_1 - t_0$ 
6:   return  $\text{actualInterval}$ 
7: end function

```

A baseline was established by running the jitter algorithm with no correction factor for 600 data points, increasing the load average by one every 100 points. The results are shown in figure 2. The Beaglebone has significantly more variability and a higher average jitter than the server. One surprise was the high jitter at 0 load displayed by both the Beaglebone and the Server. Speculation is that the operating system is putting the processor into a power saving mode at low load averages, and coming out of this mode causes a slight delay.

The baseline data show a base jitter of approximately 90,000 nanoseconds for the Beaglebone. To compensate for this, the project team implemented a simple PI control algorithm, shown in Algorithm 2. A smoothing version of this controller, which took a rolling average of the previous 10 jitter values was also implemented.

Algorithm 2 PI feedback control

```

1: function PICONROLLER(targetInterval, actualInterval)
2:   static  $\text{integral} \leftarrow 0$  ▷ maintains value across calls
3:    $\text{error} \leftarrow \text{actualInterval} - \text{targetInterval}$ 
4:    $\text{integral} \leftarrow \text{integral} + \text{error}$ 
5:    $\text{adjustment} \leftarrow K_p * \text{error} + K_i * \text{integral}$ 
6:    $\text{adjustedInterval} \leftarrow \text{targetInterval} - \text{adjustment}$ 
7:   return  $\text{adjustedInterval}$ 
8: end function

```

III. CONCLUSION

Results of this implementation, run at a load average of 1 and a target interval of 100,000 ns is shown in figure 3 ($K_p = K_i = .01$). After approximately 320 milliseconds, the PI algorithm has reached its steady state, while the smoothed PI run has reached steady state after approximately 150 milliseconds. Table II shows the jitter values across various platforms and algorithms. The values for the server platform baseline show a fairly low mean of 54885 ns, and, more importantly, a very small standard deviation which suggests a low variability on this platform. The Beaglebone has a higher mean jitter and a significantly higher standard deviation.

TABLE II
TEST RESULTS, LOAD AVERAGE = 1

Platform	Algorithm	Jitter (ns)			
		Mean	Min	Max	σ
Server	Baseline	54885	53792	66211	1324
Beaglebone	Baseline	90957	86583	519209	19231
Beaglebone	PI	9047	-4583	92667	19355
Beaglebone	Smoothed PI	4566	-6168	297667	17956

Ultimately, the baseline mean values provide a minimum value for precision timing intervals, because the adjustment factors are subtracted from the target intervals. With a mean value of close to 55 μ seconds, the server platform can provide more precise timing than the Beaglebone mean value of 91 μ seconds. However, considering that the Beaglebone platform costs almost one-hundredth of the server platform, this can be a reasonable tradeoff for many applications.

A. Future Work

The K_p and K_i parameters to the PI control loop were obtained by trial and error. A more rigorous derivation of these values could provide a better response.

There are a number of other low-cost Linux platforms available. Trials could be run on these platforms to see how

they behave. In particular, the Pcdino3 [8] provides a dual-core system, which could provide tighter timing intervals.

One of the driving motivations for using a low-cost micro-controller is for wide deployment of sensor nodes. These type of nodes are often run on batteries, and thus power consumption is a critical design consideration. It would be interesting to evaluate the precision timing algorithms described in this paper with regard to their power consumption.

Acknowledgements

The authors would like to thank Brian Fromme and Dann Frazier for their assistance in understanding Linux kernel behavior at zero load average.

REFERENCES

- [1] beagleboard.org. *Beaglebone Black*. URL: <http://beagleboard.org/black>.
- [2] die.net. *stress(1) - Linux man page*. URL: <http://linux.die.net/man/1/stress>.
- [3] die.net. *time(7) - Linux man page*. URL: <http://linux.die.net/man/7/time>.
- [4] A. Dubey, G. Karsai, and S. Abdelwahed. "Compensating for Timing Jitter in Computing Systems with General-Purpose Operating Systems". In: *Object/Component/Service-Oriented Real-Time Distributed Computing, 2009. ISORC '09. IEEE International Symposium on*. Mar. 2009, pp. 55–62. DOI: 10.1109/ISORC.2009.28.
- [5] Raspberry Pi Foundation. *Raspberry Pi*. URL: <http://www.raspberrypi.org>.
- [6] Chenyang Lu et al. "Feedback control real-time scheduling: Framework, modeling, and algorithms*". In: *Real-Time Systems* 23.1-2 (2002), pp. 85–126.
- [7] Pavel Moryc and Jindřich Černohorský. "Task jitter measurement under RTLinux operating system". In: *Journal of Automation Mobile Robotics and Intelligent Systems* 3 (2009), pp. 62–65.
- [8] pcdino. *pcDuino3*. URL: <http://www.pcdino.com/pcduino-v3>.
- [9] Steven Rostedt and Darren V Hart. "Internals of the RT Patch". In: *Proceedings of the Linux symposium*. Vol. 2007. Citeseer. 2007.

C.2 Mitigating Communication Delays in Remotely Connected Hardware-in-the-Loop Experiments

Mitigating Communication Delays in Remotely Connected Hardware-in-the-Loop Experiments

James L. Cale¹, Senior Member, IEEE, Brian B. Johnson, Member, IEEE, Emiliano Dall'Anese², Member, IEEE, Peter M. Young³, Senior Member, IEEE, Gerald Duggan, Poorva A. Bedge⁴, Student Member, IEEE, Daniel Zimmerle, Member, IEEE, and Leah Holton

Abstract—This paper introduces a potential approach for mitigating the effects of communication delays between multiple, closed-loop hardware-in-the-loop experiments which are virtually connected, yet physically separated. The approach consists of an analytical procedure for the compensation of communication delays, along with the supporting computational and communication infrastructure. The control design leverages tools for the design of observers for the compensation of measurement errors in systems with time-varying delays. The proposed methodology is validated through computer simulation and hardware experimentation connecting hardware-in-the-loop experiments conducted between laboratories separated by a distance of over 100 km.

Index Terms—Delay compensation, drift observability, hardware-in-the-loop (HIL).

I. INTRODUCTION

SYSTEM emulation using hardware-in-the-loop (HIL) is a frequently used research and validation tool because it enables dynamic cosimulation of models with physical hardware in real time, see [1]–[7]. Several national laboratories, universities, and industrial companies are now pursuing approaches in forming virtually connected simulation testbeds, whereby individual HIL experiments (representing subsystems) at distant locations share state information through communication links to emulate larger, virtually connected systems. Motivation for

this research is driven largely by the possibility of sharing experimental resources across large geographical distances without the need for physical relocation of (potentially scarce, sensitive, or massive) research equipment.

Previous research has been performed in connecting geographically separated testbeds in real time, with broad applications. For instance, in [8], a power system for a hybrid electric vehicle using a driver-in-the-loop motion simulator located in Michigan was connected in real time to a hybrid power train in California. To support naval power systems research, the work in [9] used (software-only) simulation of a shipboard power system where components were geographically dispersed and the simulation communicated between laboratories in Mississippi and South Carolina.

To evaluate distributed energy penetration on the U.S. electric grid, a joint experiment analyzing high penetration photovoltaics (PV) on an electrical distribution feeder was performed between two U.S. national laboratories using remote HIL, wherein a large-scale grid simulation (software-only) at a laboratory in Washington was virtually connected with a set of physical, residential PV inverters operating at another laboratory in Colorado [10].

More recently, real-time (software only) simulation of two large ac power systems connected through an HVdc link was performed between laboratories in Norway and Germany [11] and between Germany and South Carolina [12]. Researchers at a national laboratory in Australia performed a closed-loop, remote HIL with a national laboratory in Colorado, where a power network simulator and a physical PV+battery inverter system in the U.S. were virtually connected to a physical PV inverter in Australia to demonstrate coordinated solar PV firming on electrical distribution feeders [13].

However, previous research has not adequately addressed strategies for mitigating communication latencies specific to remotely connected HIL experimentation. As the complexity, number, and physical distances between remotely connected HIL systems increase, it is expected that communication delays will begin to adversely impact the time resolutions possible in multisystem(location) HIL experiments. In particular, communication sampling times between remotely located processors place a fundamental limitation on the effective bandwidth of the combined experiment due to the Nyquist frequency cutoff. Moreover, communication latencies from network traffic and security firewalls at potential host locations can be a significant

Manuscript received February 19, 2017; revised June 25, 2017 and February 1, 2018; accepted March 7, 2018. Date of publication March 30, 2018; date of current version July 30, 2018. This work was supported in part by the National Renewable Energy Laboratory operated for the Department of Energy by the Alliance for Sustainable Energy, LLC under Contract DOE-AC36-08-GO28308. (Corresponding author: James L. Cale.)

J. L. Cale, G. Duggan, P. A. Bedge, D. Zimmerle, and L. Holton are with the Department of Mechanical Engineering, Colorado State University, Fort Collins, CO 80523 USA (e-mail: jcale@colostate.edu; jerry.duggan@colostate.edu; poorva.bedge@colostate.edu; dan.zimmerle@colostate.edu; leah.haling@colostate.edu).

B. B. Johnson is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: brianbj@uw.edu).

E. Dall'Anese is with the National Renewable Energy Laboratory, Golden, CO 80401 USA (e-mail: emiliano.dallanese@nrel.gov).

P. M. Young is with the Department of Electrical Engineering, Colorado State University, Fort Collins, CO 80401 USA (e-mail: peter.young@colostate.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIE.2018.2821618

0278-0046 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

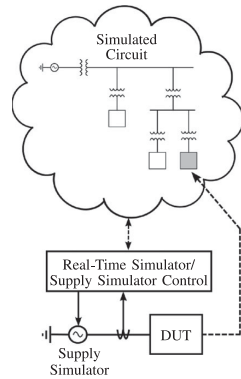


Fig. 1. Configuration for a single HIL experiment.

source of delay. Advanced methods for mitigating communication latencies are needed to enable larger and more complex virtually connected testbeds.

This paper describes a potential control methodology, computational and communications architecture for helping to mitigate the effects of communication delays between remote HIL experiments, and validates the method on an example system through simulation and hardware demonstration.

Key contributions of this paper are: 1) introducing concepts from the control of drift-observable, time-delayed systems as a potential framework and approach for mitigating communication delays in remotely connected HIL systems; and 2) demonstrating the approach in simulation and hardware.

The remainder of this paper is as follows. Section II provides a brief background on HIL and previous research in relevant delay compensation methods, with particular emphasis on observers for time-varying delay systems. This is followed in Section III by a description of the design approach introduced in this paper. The approach is demonstrated on an example system in Section IV. Section V demonstrates a validation of the concept in hardware. This paper concludes with a summary of findings and suggestions for future research in Section VI.

II. BACKGROUND

A. Brief Overview of HIL Simulation

HIL platforms enable realistic evaluation of physical hardware in the context of a modeled system. Using real-time simulators, models are executed and exchange signals with a physical device under test (DUT) in a closed-loop fashion; an example configuration is depicted in Fig. 1.

In Fig. 1, the supply simulator is a controllable power supply, which is physically connected to the DUT. The real-time simulator computes the response of the simulated circuit (bubbled area in Fig. 1) and provides control signals to the supply simulator. Measurements from the experiment are fed back to the circuit model on the real-time simulator, thus “closing the loop.” In this way, the dynamic response of the DUT inside the model is simulated, i.e., the DUT *appears* to be inside the circuit model (depicted by gray box in Fig. 1).

Single location real-time HIL, including controller HIL and power HIL, have been developed extensively for closed-loop simulations. Examples include the simulation of physical controllers and power devices [1], [2] and for investigating demand-side management techniques for providing grid ancillary services [3]. These simulations have included multiphysics domains as well, e.g., see [4] and [5].

Closed-loop *remote* HIL consists of two or more of the individual HIL systems depicted in Fig. 1, which are physically separated, yet exchange state information through communication links to provide a virtually connected, closed-loop simulation. However, communication latencies between HIL instances are a critical factor in ensuring that the desired bandwidth of the emulated experiment is met. The following sections describe prior research in delay compensation techniques for time-varying delay control systems, which, as will be proposed, provides a basis for a method to mitigate communication latencies in remotely connected HIL systems.

B. Delay Compensation Methods in HIL Simulation

Prior research has been performed on delay compensation for single location HIL simulations. In [14], current filters were incorporated in the feedback path to improve the stability of the closed-loop software ↔ hardware simulation. Additional approaches to prevent instabilities of the resultant closed-loop system were proposed in [15]. However, the approaches in [14]–[23] focused on single HIL testbeds, and did not address the issue of connecting remote systems through HIL.

The real-time simulation platform developed in [24] was utilized to emulate the operation of cyber-physical energy systems, where control signals for energy resources were dispatched and exchanged through a communication network. The platform combined real-time simulations of dynamic electric models and communication systems, with the goals of assessing the impact of noninstantaneous communications on distributed control tasks and evaluating reconfiguration strategies for the communication systems. Communication protocols were emulated within the HIL platform, but remote communication between HIL instances was not considered.

Communication and processing delays in wide area control systems were evaluated in [25]. Communication delays were estimated from empirical measurements from a transmission system; operational delays were then estimated and assessed via (single location) HIL.

The work in [26] considered remote access to HIL simulators to control a dynamical system. The framework was demonstrated for remote control of a robotic arm. The main objective was to provide control and simulation capabilities without physical existence of the equipment in the laboratory. The framework was close in spirit to the method presented in this paper; however, the work in [26] did not consider utilizing HIL for testing and emulating remote HIL systems or assess the impact of communication delays on the control and emulation tasks.

Focusing on general (non)linear dynamical systems, an approach for the construction of a state observer for nonlinear systems when the output measurements have nonnegligible time

delays was proposed in [27]. Algorithms for estimating the system state in the presence of different communication delays were proposed, and conditions ensuring global exponential convergence to zero of the estimation error for any given value of the delay in the measurements were derived. Closely related work in [28]–[30] described the design of observers that are guaranteed to asymptotically converge to the values of the actual (nondelayed) system output.

As described in the following sections, the theory for observer design for systems with time-varying delays described in [27]–[30] provides a potential building block for mitigating the effect of communication delays between remote HIL experiments, as described in this paper. Combined with a cyber-physical architecture for the HIL configuration(s), design, and selective placement of embedded observers, the proposed approach leverages the theory in [27]–[30] to provide a framework for compensating delays between remotely connected HIL simulations to produce a system simulation that appears virtually connected (i.e., colocated).

C. Observers for Time-Varying Delay Systems

This section provides a condensed summary of key research findings on observer design for systems with time-varying delays described in [27]–[30], which will be referred to in subsequent sections.

Notation: Upper-case (lower-case) boldface letters will be used for matrices (column vectors); $(\cdot)^T$ denotes transposition; $|\cdot|$ denotes the absolute value of a number or cardinality of a set; ∇ stands for the gradient operator; $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product of vectors $\mathbf{x}$ and $\mathbf{y}$. For a given $N \times 1$ vector $\mathbf{x} \in \mathbb{R}^n$, x_i denotes its i th component, and $\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^T \mathbf{x}}$. $\mathcal{C}^\infty$ denotes the field of infinitely differentiable functions.

Consider the following dynamical system describing the evolution of a state vector $\mathbf{x}(t) \in \mathbb{R}^n$:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t)) \quad (1a)$$

$$\bar{y}(t) = h(\mathbf{x}(t - \delta(t))) \quad (1b)$$

where $\mathbf{u}(t) \in \mathbb{R}^m$ is a known input vector, functions $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, and $h : \mathbb{R}^n \rightarrow \mathbb{R}$ are in $\mathcal{C}^\infty$, and $\bar{y}(t)$ represents a measurement of the *delayed* state vector $\mathbf{x}$ at time $t - \delta(t)$. Particularly, $\delta(t) \in [0, \Delta]$, $\forall t$ represents a known, time-varying and bounded measurement delay. The objective of [27]–[30] is to design an observer to predict $\mathbf{x}(t)$ by processing the delayed measurement $\bar{y}(t)$.

Let $\xi : \mathbb{R}^n \rightarrow \mathbb{R}$ be an infinitely differentiable function, and consider a vector field $\mathbf{v} : \mathcal{S} \rightarrow \mathbb{R}^n$, $\mathcal{S} \subseteq \mathbb{R}^n$. Suppose that $\mathbf{v}(\mathbf{x}) \in \mathcal{C}^\infty$. The Lie derivative of the function $\xi(\mathbf{x})$ along the vector field $\mathbf{v}$ is defined as follows:

$$L_{\mathbf{v}}\xi(\mathbf{x}) := \langle \nabla \xi(\mathbf{x}), \mathbf{v}(\mathbf{x}) \rangle = \sum_{i=1}^n \frac{\partial \xi(\mathbf{x})}{\partial x_i} v_i(\mathbf{x}). \quad (2)$$

The k th Lie derivative of $\xi(\mathbf{x})$, denoted $L_{\mathbf{v}}^k \xi(\mathbf{x})$, is obtained by k -times repeated iteration of $L_{\mathbf{v}}\xi(\mathbf{x})$ and $L_{\mathbf{v}}^0 \xi(\mathbf{x}) := \xi(\mathbf{x})$.

Consider the following mapping associated with functions $\mathbf{f}(\cdot)$ and $h(\cdot)$ in (1):

$$\Phi(\mathbf{x}) := [h(\mathbf{x}) \quad L_{\mathbf{f}}h(\mathbf{x}) \quad \cdots \quad L_{\mathbf{f}}^{n-1}h(\mathbf{x})]^T. \quad (3)$$

System (1) is said to be *globally drift-observable* if $\Phi(\mathbf{x})$ is a diffeomorphism on $\mathbb{R}^n$. Drift-observability of (1) implies that the Jacobian $\mathbf{J}(\mathbf{x})$ associated with $\Phi(\mathbf{x})$ is nonsingular for all $\mathbf{x} \in \mathbb{R}^n$, in which case the mapping $\mathbf{z} = \Phi(\mathbf{x})$ defines a global change of coordinates.

Suppose that system (1) is globally drift-observable, and has the following additional properties: (P1) The triple $(\mathbf{f}, \mathbf{g}, h)$ has *uniform observation degree* at least equal to n defined as follows:

$$L_{\mathbf{g}}L_{\mathbf{f}}^k h(\mathbf{x}) = 0, \quad k = 0, \dots, n-2, \quad \forall \mathbf{x} \in \mathbb{R}^n \quad (4)$$

$$L_{\mathbf{g}}L_{\mathbf{f}}^{n-1}h(\mathbf{x}) \neq 0, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n \quad (5)$$

in which case, the following function is well-defined

$$p(\mathbf{z}, \mathbf{u}) = (L_{\mathbf{f}}^n h(\mathbf{x}) + L_{\mathbf{g}}L_{\mathbf{f}}^{n-1}h(\mathbf{x})\mathbf{u})_{\mathbf{x}=\Phi^{-1}(\mathbf{z})}. \quad (6)$$

(P2) Function $p(\mathbf{z}, \mathbf{u})$ in (6) is globally uniformly Lipschitz continuous with respect to $\mathbf{z}$, and the Lipschitz coefficient $\gamma(\|\mathbf{u}\|)$ is a nondecreasing function of $\|\mathbf{u}\|$; i.e., for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$, it holds that

$$\|p(\mathbf{z}_1, \mathbf{u}) - p(\mathbf{z}_2, \mathbf{u})\| \leq \gamma(\|\mathbf{u}\|)\|\mathbf{z}_1 - \mathbf{z}_2\|. \quad (7)$$

If (1) is globally uniformly Lipschitz drift-observable (GULDO) with properties (P1) and (P2), then the following observer associated with system (1) can be constructed

$$\begin{aligned} \dot{\hat{\mathbf{x}}}(t) &= \mathbf{f}(\hat{\mathbf{x}}(t)) + \mathbf{g}(\hat{\mathbf{x}}(t), \mathbf{u}(t)) \\ &\quad + \mathbf{J}^{-1}(\hat{\mathbf{x}}(t))\mathbf{k}_\delta [\bar{y}(t) - h(\hat{\mathbf{x}}(t - \delta(t)))], \quad t \geq 0 \end{aligned} \quad (8a)$$

$$\mathbf{k}_\delta = e^{-\rho\delta} \mathbf{k}_0 \quad (8b)$$

where vector $\mathbf{k}_0 \in \mathbb{R}^n$ and scalar $\rho \geq 0$ are design parameters. Theorem 1 in [28] asserts that if input $\|\mathbf{u}(t)\| \leq u_M$ for some constant u_M , then for a decay rate $\rho \geq 0$ and bounded delay $\delta(t) \in [0, \Delta]$ there exists a vector $\mathbf{k}_0 \in \mathbb{R}^n$ such that

$$\|\mathbf{x}(t) - \hat{\mathbf{x}}(t)\| \leq ce^{-\rho t}, \quad t \geq 0 \quad (9)$$

for some constant c , i.e., $\hat{\mathbf{x}}(t)$ asymptotically converges to the ideal (nondelayed) system state $\mathbf{x}(t)$. Furthermore, the theorem states that if vector $\mathbf{k}_0$ satisfies the matrix inequality

$$\begin{aligned} &(\mathbf{A}_n - \mathbf{k}_0 \mathbf{C}_n)^T \mathbf{P} + \mathbf{P}(\mathbf{A}_n - \mathbf{k}_0 \mathbf{C}_n) \\ &\quad + (2\rho + \beta + \kappa)\mathbf{P} + \gamma_M^2 (\mathbf{B}_n^T \mathbf{P} \mathbf{B}_n) \mathbf{I}_{n \times n} \leq \mathbf{0} \end{aligned} \quad (10)$$

where $(\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n)$ are a Brunowski triple of order n , γ_M is the Lipschitz coefficient associated with u_M , $\beta > 0$ and $\kappa > 1$ design parameters, and $\mathbf{P}$ is symmetric positive definite, then exponential-decay state tracking [cf., (9)] is guaranteed for

$$\Delta \leq \bar{\Delta} := \frac{\beta}{2 + (\mathbf{k}_0^T \mathbf{P} \mathbf{k}_0) \|\mathbf{P}^{-1}\| (1 + \rho^2 + \mathbf{C}_n^2 \mathbf{k}_1^2) \kappa}. \quad (11)$$

Remark: it is important to note that the theory derived in [27]–[30] and summarized in (1)–(11) assumes control of a single system. While we use the structure of the observer design

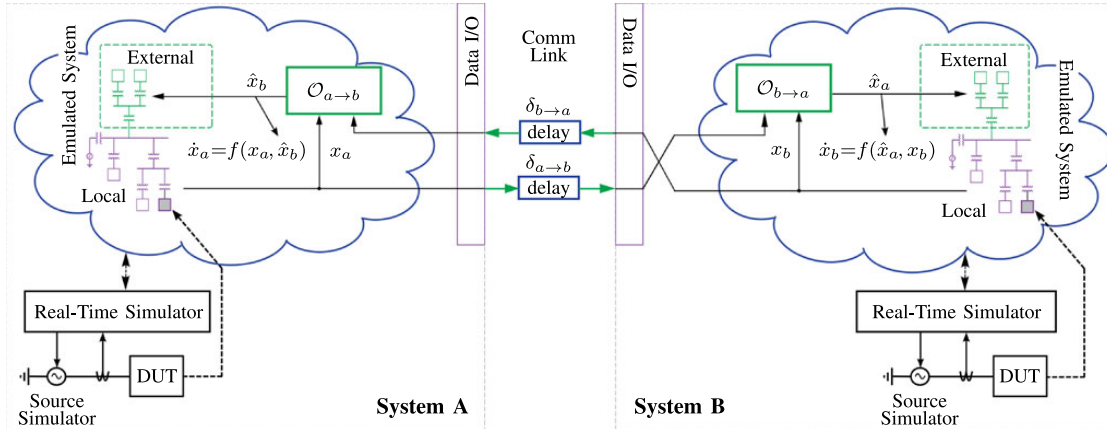


Fig. 2. Conceptual model of ODC architecture for two HIL systems connected via a communication link.

in (8) as a guide for the design of observers for remote HIL applications (with justification explained in the next section), convergence for multiple HIL systems must be verified in simulation until the convergence guarantee in (11) is extended and rigorously proven for multiple HIL instances; this an objective of future research.

III. EMBEDDED OBSERVERS FOR REMOTE HIL APPLICATIONS

This section explains how the theory in the previous section can be adapted to help mitigate communication delays in remote HIL applications, hereafter referred to as the observer delay compensation (ODC) approach. The central concept is depicted in Fig. 2. As shown in Fig. 2, the mathematical description of System A(B) is a function of both locally obtained (measured or computed) state information $x_{a(b)}$ and estimates of remote states $\hat{x}_{b(a)}$. Estimates for remote states occurring at System A(B) are computed in *observers* located at System B(A). In control theory, observers are commonly used to estimate actual state variables based on known information, such as measurements [31]. The notation $\mathcal{O}_{x \rightarrow y}$ denotes an observer at System X that estimates delayed states received from System Y; $\delta_{x \rightarrow y}$ denotes the communication delay between System X and System Y.

In Fig. 2, delayed measurements from the remote system(s) along with knowledge of its mathematical representation are used to provide an estimate of the nondelayed remote states, through the use of embedded observers. Errors in the predicted states are then dynamically corrected as new information is obtained from the remote system(s). For simplicity, we assume herein that locally measured signals do not contain delay; only communication between remote systems is considered.

Lacking a formal convergence guarantee for multiple HIL instances corresponding to [27]–[30] at the time of the experimental work, the observer designs were performed *quasi-heuristically* with (1)–(8) as a guide. In particular, we assumed that if: 1) Systems A and B in Fig. 2 could both be shown to be GULDO and possess properties (P1) and (P2); 2) observers for the system were constructed using (8) (which is always

possible if the first condition is satisfied); and 3) the upper bound on the communication delay between the two systems could be estimated with reasonable certainty. Then, design parameter $\mathbf{k}_0$ in (8) *may* exist such that all states in the complete system converged asymptotically, which we could validate through computer simulation before hardware implementation.

The assumptions above were based on the following reasoning. When designing observers to estimate the state vectors $\mathbf{x} \in \mathbb{R}^N$, $N \leq n$ at each location for an n -dimensional total system, drift-observability of each subsystem ensures the Jacobian $\mathbf{J}(\mathbf{x})$ in (8) is nonsingular for all $\mathbf{x} \in \mathbb{R}^N$, the guarantee of observation degree $\geq n$ ensures the subsystems are observable, and global uniform Lipschitz continuity ensures adjustment (i.e., feedback control) of states occurs along smooth and well-defined trajectories.

Guided by this reasoning, the design approach for constructing embedded observers to implement the ODC method consists of the following steps.

- 1) Derivation of the mathematical description of the virtually connected subsystems.
- 2) Check of GULDO and properties (P1) and (P2) for each subsystem.
- 3) Construction of observers using (8) and selection of $\mathbf{k}_0$ for each location.
- 4) Validation of system convergence (e.g., in simulation).

What follows is a demonstration of the above steps on an example system partitioned into two subsystems.

IV. DESIGN EXAMPLE

A. Mathematical Description

Consider the electrical circuit shown in Fig. 3 containing two current sources $i_{1(2)}$ in parallel with shunt capacitors $C_{1(2)}$, and connected through a series RL impedance.

The equation describing each of the $k \in \{1, 2\}$ sources is

$$\frac{di_k}{dt} = \eta_k (v_k - e_k^*) + \alpha_k i_k \quad (12)$$

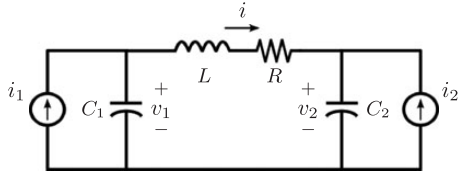


Fig. 3. Example electrical circuit (ideal).

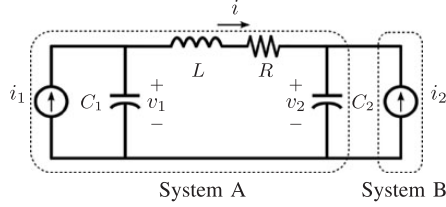


Fig. 4. Partitioned system (partition boundaries shown as dashed lines).

where it is assumed that the sources represent *composite* current supplies which include externally connected voltage sources downstream; this is used to introduce transient inputs to the simulation. Here, e_k^* represents connected (and potentially time-varying) voltage input signals, η_k are voltage error gains, and $\alpha_k < 0$ are damping factors.

This system can be represented in state-space form

$$\frac{dx}{dt} = \mathbf{A}x + \mathbf{B}u \quad (13)$$

where the vector of states $\mathbf{x} = [i_1 \ i_2 \ v_1 \ v_2 \ i]^T$

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & 0 & \eta_1 & 0 & 0 \\ 0 & \alpha_2 & 0 & \eta_2 & 0 \\ 1/C_1 & 0 & 0 & 0 & -1/C_1 \\ 0 & 1/C_2 & 0 & 0 & 1/C_2 \\ 0 & 0 & 1/L & -1/L & -R/L \end{bmatrix}, \mathbf{B} = \begin{bmatrix} -\eta_1 & 0 \\ 0 & -\eta_2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (14)$$

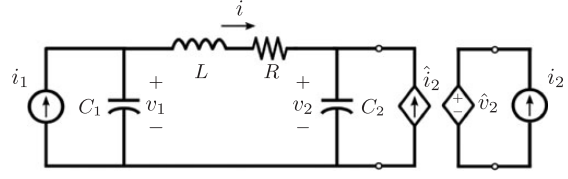
and the input vector $\mathbf{u} = [e_1^* \ e_2^*]^T$.

Suppose we wish to partition the system in Fig. 3 so portions can be simulated remotely, yet remain coupled through mutual exchange of state information. First, consider a partitioning of the system into two (sub)systems, referred to herein as “System A” and “System B,” as shown in Fig. 4.

Next, consider the physical separation and control of Systems A and B where each is controlled in the following manner. First, replace System B in Fig. 4 with a controllable current source that injects estimated input current $\hat{i}_2$ into System A, as shown in the left of Fig. 5. Then, replace System A in Fig. 4 with a controllable voltage source that provides estimated voltage $\hat{v}_2$ to System B as shown on right of Fig. 5.

System A in Fig. 5 can be described in state-space as

$$\frac{dx_a}{dt} = \mathbf{A}_a x_a + \mathbf{B}_a u_a \quad (15)$$

Fig. 5. Systems A and B with remote inputs $\hat{i}_2$ and $\hat{v}_2$, respectively.

where the vector of states $\mathbf{x}_a = [i_1 \ v_1 \ i \ v_2]^T$

$$\mathbf{A}_a = \begin{bmatrix} \alpha_1 & \eta_1 & 0 & 0 \\ 1/C_1 & 0 & -1/C_1 & 0 \\ 0 & 1/L & -R/L & -1/L \\ 0 & 0 & 1/C_2 & 0 \end{bmatrix}, \mathbf{B}_a = \begin{bmatrix} -\eta_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/C_2 \end{bmatrix} \quad (16)$$

and the input vector $\mathbf{u}_a = [e_1^* \ \hat{i}_2]^T$. The nominal output of System A at System A is $v_2(t)$, the delayed output $\bar{y}_a(t) = v_2(t - \delta(t))$ in (1), is obtained by defining matrix $\mathbf{C}_a$ as

$$\begin{aligned} \bar{y}_a(t) &= \mathbf{C}_a x_a(t - \delta(t)) \\ &= [0 \ 0 \ 0 \ 1] x_a(t - \delta(t)) \\ &= v_2(t - \delta(t)). \end{aligned} \quad (17)$$

For System B, the state-space representation is

$$\frac{dx_b}{dt} = \mathbf{A}_b x_b + \mathbf{B}_b u_b \quad (18)$$

where the state vector $\mathbf{x}_b = [i_2]$

$$\mathbf{A}_b = [\alpha_2], \mathbf{B}_b = [-\eta_2 \ \eta_2] \quad (19)$$

and the input vector $\mathbf{u}_b = [e_2^* \ \hat{v}_2]^T$. The nominal output of System B at System B is $i_2(t)$, so matrix $\mathbf{C}_b$ is defined as

$$\begin{aligned} \bar{y}_b(t) &= \mathbf{C}_b x_b(t - \delta(t)) \\ &= [1][i_2(t - \delta(t))] \\ &= i_2(t - \delta(t)). \end{aligned} \quad (20)$$

For System A note from comparison of (1) and (15)–(17)

$$\mathbf{f}(\mathbf{x}(t)) \mapsto \mathbf{A}_a x_a(t) \quad (21)$$

$$\mathbf{g}(\mathbf{x}(t))\mathbf{u}(t) \mapsto [-\eta_1 e_1^*(t) \ 0 \ 0 \ 0]^T \quad (22)$$

$$\mathbf{h}(\mathbf{x}(t - \delta(t))) \mapsto \mathbf{C}_a x_a(t - \delta(t)) = v_2(t - \delta(t)). \quad (23)$$

The mapping $\mathbf{z}_b = \Phi(\mathbf{x}_a)$ for System A, with $n = 4$ is

$$\Phi(\mathbf{x}_a) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1/C_2 & 0 \\ 0 & 1/LC_2 & -R/LC_2 & -1/LC_2 \\ \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \end{bmatrix} x_a = \mathbf{\Lambda}_a x_a \quad (24)$$

where $\sigma_1 = 1/LC_1C_2$, $\sigma_2 = -1/LC_1C_2 - R/L^2C_2$, $\sigma_3 = -R^2/L^2C_2 - 1/LC_2^2$, and $\sigma_4 = R/L^2C_2$. Note that the mapping $\Phi(\mathbf{x}_a)$ is diffeomorphic if $\mathbf{\Lambda}_a$ defined in (24) is invertible.

which is determined by the values of the circuit parameters. From (24)

$$\mathbf{J}_a(\mathbf{x}_a) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1/C_2 & 0 \\ 0 & 1/LC_2 & -R/LC_2 & -1/LC_2 \\ \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \end{bmatrix} = \mathbf{A}_a \quad (25)$$

which is indeed nonzero for all $\mathbf{x}_a \in \mathbb{R}^4$. For property (P1), it is easy to verify that $L_g L_f^k \mathbf{h}(\mathbf{x}) = 0$ for $k = 0, \dots, n-2$, $\forall \mathbf{x} \in \mathbb{R}^n$. The second condition in (P1) holds since $L_g L_f^3 \mathbf{h}(\mathbf{x}_a) = \mathbf{\Gamma} \mathbf{g}(\mathbf{x}(t)) \mathbf{u}(t)$ where $\mathbf{\Gamma} := [\varsigma_1 \ \varsigma_2 \ \varsigma_3 \ \varsigma_4]$, $\varsigma_1 = 1/LC_1 C_2$, $\varsigma_2 = -R/L^2 C_2$, $\varsigma_3 = (-1/LC_1 C_2 + R^2/L^2 C_2 - 1/L^2 C_2)$, $\varsigma_4 = R/L^2 C_2$. Therefore, $L_g L_f^3 \mathbf{h}(\mathbf{x}_a) = -\eta_1 \varsigma_1 e_1^*(t) \neq 0$ for some $\mathbf{x}_a \in \mathbb{R}^4$, in particular whenever the signal $e_1^*(t) \neq 0$. As expected, $p(\mathbf{z}_a, \mathbf{u}_a)$ is well-defined

$$\begin{aligned} p(\mathbf{z}_a, \mathbf{u}_a) &= (L_f^4 \mathbf{h}(\mathbf{x}_a) + L_g L_f^3 \mathbf{h}(\mathbf{x}_a) \mathbf{u}_a)_{\mathbf{x}_a = \Phi^{-1}(\mathbf{z}_a)} \\ &= \mathbf{\Gamma} \mathbf{x}_a - \eta_1 \varsigma_1 (e_1^*(t))^2 \\ &= \mathbf{\Gamma} \mathbf{A}_a^{-1} \mathbf{z}_a - \eta_1 \varsigma_1 (e_1^*(t))^2. \end{aligned} \quad (26)$$

Considering property (P2) for System A, note that $p(\mathbf{z}_a, \mathbf{u}_a)$ is globally uniformly Lipschitz continuous with respect to $\mathbf{z}_a$; its derivative with respect to this variable is equal to (constant) $\mathbf{\Gamma} \mathbf{A}_a^{-1}$. The Lipschitz coefficient can be determined from (7) to derive the requirement that $\gamma \geq \|\mathbf{\Gamma} \mathbf{A}_a^{-1}\|$ and is therefore nondecreasing in $\|\mathbf{u}_a\|$.

For System B, note from comparison of (1) and (18) and (20),

$$\mathbf{f}(\mathbf{x}(t)) \mapsto \mathbf{A}_b \mathbf{x}_b(t) = \alpha_2 i_2(t) \quad (27)$$

$$\mathbf{g}(\mathbf{x}(t)) \mathbf{u}(t) \mapsto -\eta_2 e_2^*(t) \quad (28)$$

$$\mathbf{h}(\mathbf{x}(t - \delta(t))) \mapsto \mathbf{C}_b \mathbf{x}_b(t - \delta(t)) = i_2(t - \delta(t)). \quad (29)$$

The mapping $\mathbf{z}_b = \Phi(\mathbf{x}_b)$ for System B with $n = 1$ is

$$\Phi(\mathbf{x}_b) = \mathbf{x}_b = i_2(t). \quad (30)$$

Functions of the form $\Phi(x) = x$, $x \in \mathbb{R}$ are diffeomorphic and $\Phi(\mathbf{x}_b)$ is a diffeomorphism for $\mathbf{x}_b \in \mathbb{R}$. The Jacobian for System B

$$\mathbf{J}_b(\mathbf{x}_b) = 1 \quad (31)$$

which is indeed nonzero for all $\mathbf{x}_b \in \mathbb{R}$.

We now examine whether System B possesses properties (P1) and (P2). For the first property, note that the first condition in (6) doesn't apply since $n = 1$. The second condition is true since $L_g L_f^0 \mathbf{h}(\mathbf{x}_b) = -\eta_2 e_2^*(t) \neq 0$, for some $\mathbf{x}_b \in \mathbb{R}$, in particular whenever $e_2^*(t) \neq 0$. Note the function

$$\begin{aligned} p(\mathbf{z}_b, \mathbf{u}_b) &= (L_f^n \mathbf{h}(\mathbf{x}_b) + L_g L_f^{n-1} \mathbf{h}(\mathbf{x}_b) \mathbf{u}_b)_{\mathbf{x}_b = \Phi^{-1}(\mathbf{z}_b)} \\ &= \alpha_2 \mathbf{z}_b - \eta_2 (e_2^*(t))^2 \end{aligned} \quad (32)$$

is indeed well-defined. Considering property (P2) for System B, note that $p(\mathbf{z}_b, \mathbf{u}_b)$ is globally uniformly Lipschitz continuous with respect to $\mathbf{z}_b$; its derivative with respect to $\mathbf{z}_b$ is equal to constant α_2 for all $\mathbf{z}_b \in \mathbb{R}$. The Lipschitz coefficient can be

determined from the inequality

$$\begin{aligned} \|p(\mathbf{z}_{b,1}, \mathbf{u}_b) - p(\mathbf{z}_{b,2}, \mathbf{u}_b)\| &= \|\alpha_2\| \|\mathbf{z}_{b,1} - \mathbf{z}_{b,2}\| \\ &\leq \gamma(\|\mathbf{u}_b\|) \|\mathbf{z}_{b,1} - \mathbf{z}_{b,2}\| \end{aligned} \quad (33)$$

from which we derive that $\gamma \geq \|\alpha_2\|$, and is therefore nondecreasing in $\|\mathbf{u}_b\|$.

B. Observer Construction and Parameter Selection

The observer $\mathcal{O}_{a \rightarrow b}$ estimates input current $\hat{i}_2$ at System A based on measurements of signal $i_2(t - \delta_{b \rightarrow a}(t))$ received from System B and is constructed as

$$\begin{aligned} \frac{d\hat{i}_2(t)}{dt} &= \mathbf{A}_b \hat{i}_2(t) + \mathbf{B}_b \mathbf{u}_b + \mathbf{J}_b^{-1}(\hat{i}_2(t)) \mathbf{k}_{0b} e^{-\rho \delta_{b \rightarrow a}} \\ &\quad (i_2(t - \delta_{b \rightarrow a}(t)) - \hat{i}_2(t - \delta_{b \rightarrow a}(t))). \end{aligned} \quad (34)$$

The observer $\mathcal{O}_{b \rightarrow a}$ estimates voltage $\hat{v}_2(t)$ based on measurements of signal $v_2(t - \delta_{b \rightarrow a}(t))$ received from System A and is constructed as

$$\begin{aligned} \frac{d\hat{\mathbf{x}}_a}{dt} &= \mathbf{A}_a \hat{\mathbf{x}}_a(t) + \mathbf{B}_a \mathbf{u}_a + \mathbf{J}_a^{-1}(\hat{\mathbf{x}}_a(t)) \mathbf{k}_{0a} e^{-\rho \delta_{a \rightarrow b}} \\ &\quad (v_2(t - \delta_{a \rightarrow b}(t)) - \hat{v}_2(t - \delta_{a \rightarrow b}(t))) \end{aligned} \quad (35)$$

where $\hat{\mathbf{x}}_a = [\hat{i}_1 \ \hat{v}_1 \ \hat{i}_2 \ \hat{v}_2]^T$. In (35) and (36), $\mathbf{k}_{0a(b)}$ denotes the parameter $\mathbf{k}_0$ in (8) for the observer that estimates states physically located at System A(B).

A reasonable guess for selecting suitable $\mathbf{k}_{0a(b)}$ (which must be validated, as shown in the next subsection) is to compute the feedback vector for the so-called Luenberger observer [32] corresponding to each of the subsystems. This is the method that was used in this work, where the MathWorks `place` command was used to place closed-loop poles at desired locations to yield satisfactory total system response. Note that it is always possible to obtain these feedback vectors since the subsystems are confirmed to be observable in step B) of the design process.

C. Validation in Simulation

The delay compensation technique was applied to the partitioned circuit in Fig. 4 using the parameter values shown in Table I. [Note that these parameters are not intended to represent a specific system—they were chosen only to clearly visualize the difference in the compensated and uncompensated responses.] The simulations were performed in MathWorks SimPowerSystems toolbox, version 5.7 [33].

The passive circuit elements, error gains and damping factors in Table I were chosen arbitrarily, except that $\mathbf{A}_a$ in (25) was validated to be invertible. The communication delay T_d was chosen to be several times greater than the empirically measured network delay (approximately 30 ms per round trip, as described in Section V). The update rate τ_u was chosen to be an order of magnitude greater than T_d ; the desired convergence rate ρ was arbitrarily chosen to visually observe the response. The vector $\mathbf{k}_{0a}$ was calculated as $[4.06 \ 6.18 \ 4.18 \ 1.06]^T$ for poles placed at $[-1.00 \ -1.01 \ -1.02 \ -1.03]^T$. Similarly, $\mathbf{k}_{0b}$ was calculated as $[1.0]$ for a pole placed at $[-1.0]$. The selection of

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Capacitances, C_1, C_2	1.0 [F]
Resistance, R	0.5 [Ω]
Inductance, L	1.0 [H]
Voltage Error Gain, η_1	-1.0 [$\text{s}\Omega$] ⁻¹
Voltage Error Gain, η_2	-1.0 [$\text{s}\Omega$] ⁻¹
Damping Factor, α_1	-1.0 [s] ⁻¹
Damping Factor, α_2	-1.0 [s] ⁻¹
Communication Delay, T_d	0.20 [s]
External Command Update Rate, τ_u	2.0 [s]
Observer Convergence Rate, ρ	4.0 [ms]

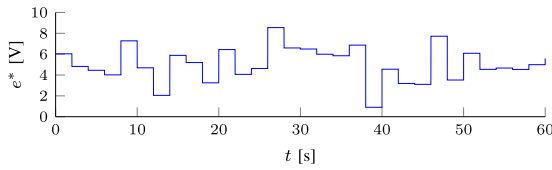


Fig. 6. Signal representing external sources.

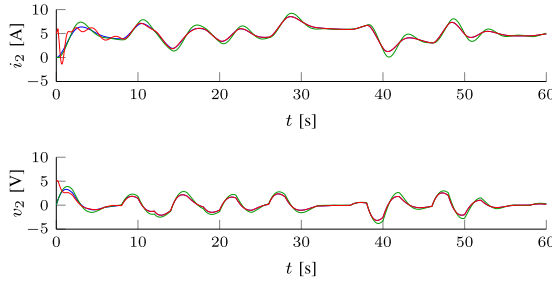


Fig. 7. Simulation results from example partitioned system. Blue lines signify the ideal response (from the ideal circuit in Fig. 3), green lines the uncompensated response, and red lines the observer output.

poles was chosen only to ensure that the full system simulation would be stable and the dynamics would be visible for the choice of system parameters; exploration of system stability with various system and observer parameters choices is a topic of future research.

For simplicity, the input signals $e_1^*(t)$ and $e_2^*(t)$ representing externally connected sources in (12) were both set equal to the waveform $e^*(t)$ depicted in Fig. 6. This function held a (randomly sampled) constant value which was updated every time interval τ_u after the start of the simulation.

The resulting simulation waveforms are shown in Fig. 7. In Fig. 7, the blue lines signify the ideal response (from the ideal circuit in Fig. 3), green lines the uncompensated response, and red lines the observer output. As can be seen, the ideal and observer based (“compensated”) signals converge for both voltage and current, despite continued changes in the input signals $e_1^*(t) = e_2^*(t) = e^*(t)$ throughout the simulation.

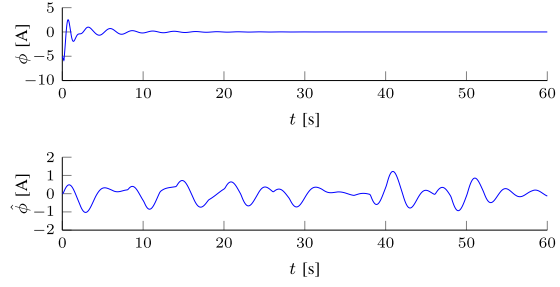


Fig. 8. Errors in i_2 with (upper) and without (lower) delay compensation.

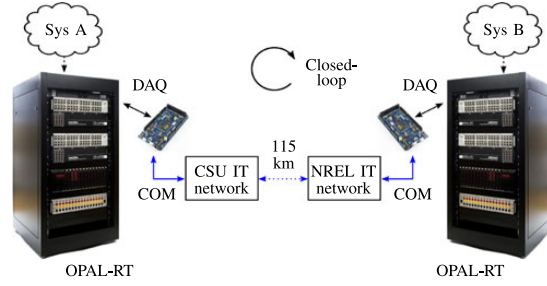


Fig. 9. Hardware in remote Sys A $\leftrightarrow$ Sys B HIL demonstration.

Errors between compensated and uncompensated responses are plotted in Fig. 8 (for current i_2). In particular, $\phi(t)$ is defined as the difference between the compensated and ideal response (blue and red lines in Fig. 7, upper); $\hat{\phi}(t)$ is the difference between the uncompensated and ideal response (green and red lines in Fig. 7, upper).

As shown in Fig. 8, the error in the compensated current signal decreases exponentially to zero, as predicted by (9). Remarkably, once the controller error converges to zero it stays at zero even after the external inputs $e_1^*(t)$, $e_2^*(t)$ continue to change. Note that continued convergence to zero in the compensated systems, despite changing external input signals, are likely possible in this example because models for the full and partitioned systems were known exactly.

V. HARDWARE RESULTS

This section describes experimental validation of the ODC method in hardware. During this experiment, a closed-loop, remote HIL experiment emulating the system shown in Fig. 5, employing the architecture shown in Fig. 2, was performed. This experiment provided a virtual connection of equipment physically located at Colorado State University (CSU), Fort Collins, CO, USA, with equipment located remotely at the National Renewable Energy Laboratory (NREL), Golden, CO, USA. The physical separation between these locations was approximately 115 km. A depiction of the overall experiment is shown in Fig. 9.

In Fig. 9, the System A (“Sys A”) and System B (“Sys B”) portions of the electrical circuit in Fig. 4 were simulated on OPAL-RT real-time digital simulators [34] at CSU and NREL,

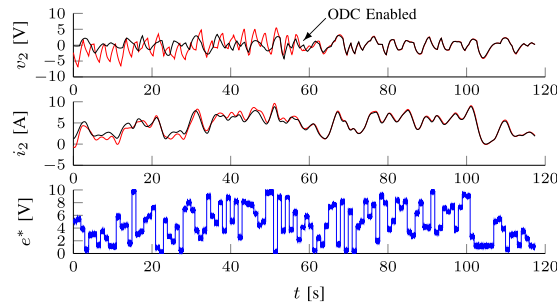


Fig. 10. Measurements from hardware validation of ODC method: compensated signals (red); ideal signals (black); and external input signals (blue).

respectively. The observers for both systems were executed on Arduino EtherDUE [35] at each location. The EtherDUE microcontrollers had a built-in Ethernet port in addition to analog and digital I/O channels.

The observer calculations were executed on the microcontroller boards rather than on the real-time simulators in order to isolate the electrical simulations from the observer computations. This approach provided the following advantages:

- 1) software for network packet manipulation and time functions were accessible through standard libraries readily installed on the microcontroller boards;
- 2) processing network signals externally to the real-time simulators eliminated the need for using limited memory and computational resources on the real-time simulators;
- 3) isolation of the electrical simulation and observers made the architecture more maintainable, since components of the system could be modified and tested independently.

Measurements and observer values were passed between colocated OPAL-RT and Arduino microcontrollers using digital and analog I/O channels. State information was passed between remote locations through the EtherDUE boards using TCP/IP communication. Synchronization between systems was implemented using algorithms on the Arduino boards in a master/slave, hand-shaking configuration.

As mentioned in Section IV, selection of observer parameters and stability analysis for this experiment required an estimate of the network delay between Systems A and B. Since network delays between two systems are generally dependent on many factors (e.g., network load and routing algorithms), the estimated delay was determined experimentally using repeated ping algorithms on the microcontrollers. The sample mean of the round-trip communication delay was found to be ≈ 30 ms.

Measurements obtained during the experiment are shown in Fig. 10. The compensation algorithms were activated at each location at approximately $t = 57$ s. The measurements show that the compensated currents and voltages (red lines) converged to the ideal response signals (black lines) with little error after the delay compensation algorithms converged. Note that as predicted in simulation, once the observer-based delay compensators converged, the error remained near zero even when the external input signal continued to change.

VI. CONCLUSION AND FUTURE RESEARCH

This paper introduced a potential method and architecture for helping to mitigate communication delays between remotely connected HIL experiments. Motivation for this paper was the desire to connect geographically distant software/hardware testbeds while minimizing the loss of experimental fidelity arising from communication latencies. The approach was validated through computer simulation and a remote HIL experiment where the subsystems were separated by a physical distance of 115 km and communication delay of 30 ms. However, we note that the convergence results shown in this paper assumed perfect knowledge of the underlying system models. Primary challenges in implementing the method are generating sufficiently accurate mathematical descriptions of the subsystems and determining controller parameters. Automated state-model generation or average value models can possibly be used for more complex systems to address the first challenge. Theoretical development to extend the convergence guarantee for drift-observable, time-delayed observer design for multiple remote systems and design guidelines for optimal parameter selection are needed to resolve the second challenge. Additional future research topics include the addition of robust control techniques to account for potential parameter variations and imperfect system knowledge.

REFERENCES

- [1] M. O. Faruque *et al.*, "Real-time simulation technologies for power systems design, testing, and analysis," *IEEE Power Energy Technol. Syst. J.*, vol. 2, no. 2, pp. 63–73, Jun. 2015.
- [2] G. F. Lauss, M. Faruque, K. Schoder, C. Dufour, A. Viehweider, and J. Langston, "Characteristics and design of power hardware-in-the-loop simulations for electrical power systems," *IEEE Trans. Ind. Electron.*, vol. 63, no. 1, pp. 406–417, Jan. 2016.
- [3] M. H. Syed, P. Crolla, G. Burt, and J. Kok, "Ancillary service provision by demand side management: A real-time power hardware-in-the-loop co-simulation demonstration," in *Proc. 2015 Int. Symp. Smart Elect. Distrib. Syst. Technol.*, Vienna, Austria, 2015, pp. 492–498.
- [4] B. Lu, X. Wu, H. Figueroa, and A. Monti, "A low-cost real-time hardware-in-the-loop testing approach of power electronics controls," *IEEE Trans. Ind. Electron.*, vol. 54, no. 2, pp. 919–931, Mar. 2007.
- [5] A. Benigni and A. Monti, "A parallel approach to real-time simulation of power electronics systems," *IEEE Trans. Ind. Electron.*, vol. 30, no. 9, pp. 5192–5206, Sep. 2015.
- [6] B. Lundstrom and S. Chakraborty, "An advanced platform for development and evaluation of photovoltaic inverters using hardware-in-the-loop," *Int. J. Distrib. Energy Resources Smart Grids*, vol. 9, no. 1, Mar. 2013.
- [7] B. Lundstrom, B. Mather, M. Shirazi, and M. Coddington, "Implementation and validation of advanced unintentional islanding testing using power hardware-in-the-loop (PHIL) simulation," in *Proc. 2013 IEEE 39th Photovoltaic Spec. Conf.*, Tampa Bay, FL, USA, Jun. 2013, pp. 3141–3146.
- [8] M. Compere, J. Goodell, M. Simon, W. Smith, and M. Brudnak, "Robust control techniques enabling duty cycle experiments utilizing a 6-DOF crewstation motion base, a full scale combat hybrid electric power system, and long distance internet communications," SAE International, Pittsburgh, PA, USA, Tech. Rep. 2006-01-3077, Nov. 2006.
- [9] J. L. Bastos, J. Wu, N. Schulz, R. Liu, and A. Monti, "Distributed simulation using the virtual test bed and its real-time extension," in *Proc. Summer Comput. Simul. Conf.*, San Diego, CA, USA, Jan. 2007, vol. 2, pp. 757–765.
- [10] B. Palmintier, B. Lundstrom, S. Chakraborty, T. Williams, K. Schneider, and D. Chassin, "A power hardware-in-the-loop platform with remote distribution circuit cosimulation," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2236–2245, Apr. 2015.
- [11] M. Stevic, S. Vogel, A. Monti, and S. D'Arco, "Feasibility of geographically distributed real-time simulation of HVDC system interconnected with AC networks," in *Proc. 2015 IEEE PowerTech Eindhoven*, Eindhoven, The Netherlands, Jun. 2015, pp. 1–5.

- [12] M. Stevic *et al.*, "Virtual integration of laboratories over long distance for real-time co-simulation of power systems," in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc.*, Florence, Italy, Oct. 2016, pp. 6717–6721.
- [13] B. Lundstrom, B. Palmintier, D. Rowe, J. Ward, and T. Moore, "Transoceanic remote power hardware-in-the-loop: Multi-site hardware, integrated controller, and electric network co-simulation," *IET Gener. Transmiss. Distrib.*, vol. 11, no. 18, pp. 4688–4701, Jul. 2017.
- [14] G. Lauss, F. Lehfuss, A. Viehweider, and T. Strasser, "Power hardware in the loop simulation with feedback current filtering for electric systems," in *Proc. 37th Annu. Conf. IEEE Ind. Electron. Soc.*, Melbourne, VIC, Australia, Nov. 2011, pp. 3725–3730.
- [15] A. Viehweider, G. Lauss, and F. Felix, "Stabilization of power hardware-in-the-loop simulations of electric energy systems," *Simul. Model. Practice Theory*, vol. 19, pp. 1699–1708, Sep. 2011.
- [16] M. Steurer, C. Edrington, M. Sloderbeck, W. Ren, and J. Langston, "A megawatt-scale power hardware-in-the-loop simulation setup for motor drives," *IEEE Trans. Ind. Electron.*, vol. 57, pp. 1254–1260, May 2010.
- [17] A. Monti, F. Ponci, Z. Jiang, and R. Dougal, "Hardware-in-the-loop testing platform for distributed generation systems," *Int. J. Energy Technol. Policy*, vol. 5, no. 2, pp. 241–257, Jan. 2007.
- [18] L. Gauchia and J. Sanz, "A per-unit hardware-in-the-loop simulation of a fuel cell/battery hybrid energy system," *IEEE Trans. Ind. Electron.*, vol. 57, pp. 1186–1194, May 2010.
- [19] S. Grubic, B. Amlang, W. Schumacher, and A. Wenzel, "A high performance electronic hardware-in-the-loop drive-load simulation using a linear inverter," *IEEE Trans. Ind. Electron.*, vol. 57, pp. 1208–1216, May 2010.
- [20] S. Suryanarayanan, W. Ren, M. Steurer, F. Ribeiro, and G. Heydt, "A real-time controller concept demonstration for distributed generation interconnection," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, Montreal, QC, Canada, Jun. 2006, 3 pp.
- [21] P. Crolla, A. J. Roscoe, A. Dysko, and G. Burt, "Methodology for testing loss of mains detection algorithms for microgrids and distributed generation using real-time power hardware-in-the-loop based technique," in *Proc. IEEE 8th Int. Conf. Power Electron. ECCE Asia*, Jeju, South Korea, Jun. 2011, pp. 833–838.
- [22] B. Lundstrom, M. Shirazi, M. Coddington, and B. Kroposki, "An advanced platform for development and evaluation of grid interconnection systems using hardware-in-the-loop: Part III—grid interconnection system evaluator," in *Proc. IEEE Green Technol. Conf.*, Denver, CO, USA, Apr. 2013, pp. 392–399.
- [23] O. Mohammed, M. A. Nayeem, and A. Kaviani, "A laboratory based microgrid and distributed generation infrastructure for studying connectivity issues to operational power systems," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, Minneapolis, MN, USA, Aug. 2010, pp. 1–6.
- [24] F. Guo and L. Herrera, "Comprehensive real-time simulation of the smart grid," *IEEE Trans. Ind. Appl.*, vol. 49, no. 2, pp. 899–908, Mar./Apr. 2013.
- [25] F. Zhang, Y. Sun, L. Cheng, X. Li, J. Chow, and W. Zhao, "Measurement and modeling of delays in wide-area closed-loop control systems," *IEEE Trans. Power Syst.*, vol. 30, no. 3, pp. 2426–2433, Sep. 2015.
- [26] T. Singaraju, A. Turan, M. Gokasan, and S. Bogosyan, "Hardware-in-the-loop simulation of PUMA 560 via internet," in *Proc. 32nd Annu. Conf. IEEE Ind. Electron.*, Paris, France, 2006, pp. 5426–5432.
- [27] A. Germani, C. Manes, and P. Pepe, "A new approach to state observation of nonlinear systems with delayed output," *IEEE Trans. Autom. Control*, vol. 47, no. 2, pp. 96–101, Feb. 2002.
- [28] F. Cacace, A. Germani, and C. Manes, "An exponential observer with delay-dependent gain for a class of nonlinear systems with time-varying measurement delay," in *Proc. IEEE Conf. Dec. Control*, Maui, HI, USA, Dec. 2012, pp. 2364–2369.
- [29] F. Cacace, A. Germani, and C. Manes, "An observer for a class of nonlinear systems with time varying observation delay," *Syst. Control Lett.*, vol. 59, pp. 305–312, May 2010.
- [30] F. Cacace, A. Germani, and C. Manes, "A chain observer for nonlinear systems with multiple time-varying measurement delays," *SIAM J. Control Optim.*, vol. 52, no. 3, pp. 1862–1885, Jun. 2014.
- [31] K. Ogata, *Modern Control Engineering*. 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
- [32] Z. Gajic and M. Lelic, *Modern Control Systems Engineering*. London, U.K.: Prentice-Hall, 1996.
- [33] "Simscape power systems." Mathworks, Natick, MA, USA, 2016. [Online]. Available: <https://www.mathworks.com/products/simpower.html>
- [34] "Hardware Overview," OPAL-RT Technologies, Montreal, QC, USA, 2016. [Online]. Available: <https://www.opal-rt.com/hardware-overview/>
- [35] "Homepage," Somerville, MA, USA, 2018. [Online]. Available: <https://www.arduino.cc>



James L. Cale (SM'16) received the B.S. degree (*summa cum laude*) in electrical engineering from the University of Missouri–Rolla, Rolla, MO, USA, in 2001, the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 2005 and 2007, respectively.

From 2007 to 2010, he was a Lead Electrical Engineer, R&D, Orbital ATK. From 2010 to 2011, he was a Member of Technical Staff with Advanced Energy. From 2011 to 2017, he was a Senior Research Engineer and Group Lead with the National Renewable Energy Laboratory. He is currently an Associate Professor with Colorado State University, Fort Collins, CO, USA. His research interests include energy conversion, power electronics, finite-inertia power systems, microgrids, and power and controller hardware-in-the-loop.



Brian B. Johnson (M'13) received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2010 and 2013, respectively.

From 2013 to 2018, he was an Electrical Engineer with the National Renewable Energy Laboratory, Golden, CO, USA. He is currently an Assistant Professor with the Department of Electrical Engineering, University of Washington in Seattle, WA, USA. His research interests include renewable energy, power electronics, and control systems.

Dr. Johnson was the recipient of National Science Foundation Graduate Research Fellowship in 2010. He is an Associate Editor for the IEEE TRANSACTIONS ON ENERGY CONVERSION.



Emiliano Dall'Anese (M'11) received the Ph.D. degree in information engineering from the University of Padua, Padua, Italy, in 2011.

From 2009 to 2010, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA, where he was a Postdoctoral Associate with the Department of Electrical and Computer Engineering, from January 2011 to November 2014. Since December 2014, he has been a Senior Researcher with the National Renewable Energy Laboratory, Golden, CO, USA. His research interests include advanced theory, algorithms, and analysis for optimization, control, and monitoring of large-scale cyber-physical systems.



Peter M. Young (SM'13) received the B.A. degree from Oxford University, Oxford, U.K., in 1985, the M.S. degree from the University of Florida, Gainesville, FL, USA, in 1988, and the Ph.D. degree from California Institute of Technology, Pasadena, CA, USA, in 1993, all in electrical engineering.

From 1985 to 1986, he was an Executive Engineer with British Telecom Research Laboratories, and from 1993 to 1995, he was a Postdoctoral Associate with Massachusetts Institute of Technology. He is currently a Professor with Colorado State University, Fort Collins, CO, USA. His research interests include the development of analysis and design techniques for large scale uncertain systems, and robust learning controllers, as well as a number of specific application areas. These include control of building energy management systems, optimal utilization of energy assets, and microgrids with high penetration renewable energy.



Gerald Duggan received the B.S. degree in math and computer science from New Mexico State University, Las Cruces, NM, USA, and the M.S. degree in computer science from the University of Utah, Salt Lake City, UT, USA, in 1983 and 1985, respectively.

From 1985 to 2009, he was a Senior Software Architect with Hewlett Packard, defining standards and implementing software for large-scale IT management systems. He is currently the Lab Manager of the Smart Village Microgrid Lab, Energy Institute of Colorado State University, Fort Collins, CO, USA. His research interests include village-scale microgrids, dc power distribution systems, and ac–dc efficiency studies.



Daniel Zimmerle received the B.S. and M.S. degrees in mechanical engineering from North Dakota State University, Fargo, ND, USA, in 1983 and 1984, respectively.

From 1985 to 2005, he worked in various roles at Hewlett Packard and Agilent Technologies including both a division R&D and General Manager. From 2006 to 2007, he was the Chief Operating Officer with Spirae, Inc. He is currently a Senior Research Associate and the Director of the Electric Power System Laboratory, Energy Institute, Colorado State University, Fort Collins, CO, USA. His research interests include microgrids and the development, integration, and field assessment of distributed, renewable, and waste heat electrical generation systems.



Poorva A. Bedge (S'15) received the B.S. degree in electrical engineering from the Government College of Engineering Pune, Pune, India, in 2004, and the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2008. She is currently working toward the Ph.D. degree in systems engineering with Colorado State University, Fort Collins, CO, USA.

She worked in Mumbai and Delhi, India with Reliance Energy as a Senior Engineer on the electric power distribution system. At Midwest ISO, she was an Outage Coordination Engineer, and at California ISO she was a Senior Operations Planning Engineer. Her research interests include developing tools and solutions to modernize the electric grid to accommodate the rapid increase in renewable energy.



Leah Holton received the B.S. degree in mathematics and the B.S. degree in engineering from James Madison University, Harrisonburg, VA, USA, in 2012, and the M.S. degree in mechanical engineering from Arizona State University, Tempe, AZ, USA, in 2014.

She is currently a Solar Engineer with the Independent Engineering Group, ICF, Denver, CO, USA. She supports solar energy project owners and investors with due diligence and owners engineering services.

Appendix D

Data Validation Taxonomy

It is very common to use spreadsheets or Comma Separated Variable (CSV) files to exchange data between groups. Tools for creating and editing these files are very common and easy to use. However, this ease of use means that it is very easy to introduce data errors, as there are typically no data validation features built-in. This section lists some of the common data errors found when using these files.

Some of the issues below are dependent on the language used to implement the solutions. In this case, it will refer to Python using Pandas [140].

- **Missing or reordered fields.** When a person is editing a spreadsheet, even if they are supplied a template, it is very common for them to delete columns that are not relevant to them, or to reorder the columns to make them more usable for their purposes. Software that is reading these files should be flexible for reordered columns, and should handle missing columns by either providing default values or throwing an error.
- **Column name mismatch.** If a person has to enter the column names on their own, or decides to edit them for some reason, it is very easy to get the column names wrong – even if this includes differences in capitalization. In this case, the processing software should signal an error that describes the expected name of the column and the column name that was actually used.
- **Spaces at the end of column names or data values.** It is very easy to inadvertently add a space to the end of a column name or a data value, which causes column name mismatches or

other data interpretation issues. This error is surprisingly difficult to detect – it often requires writing code to read the data in and explicitly check for improper spaces included.

- **Types of data fields & empty data values.** When pandas reads a file into a dataframe, it will by default infer the types of the fields, which means reading them in as integers, floats, or strings. If all the values in a column are of the same type, this behaves as expected. If they are of mixed types (e.g. strings intermixed with integers), pandas will read each individual field as its inferred type. If the software expects them to be a single type, it will have to convert them manually.

When pandas reads in a column with empty data values, the default action will be to read it in as a floating point Not a Number (NaN) value. This can cause surprising behavior, particularly if the software reading the dataframe attempts to convert the column to a string or integer. In the former case, the cell will contain the string literal value 'nan', and the latter case will raise an exception as NaN cannot be converted into an integer.

- **Key value mismatch.** It is common to have mismatches on key values relating rows from one dataframe to another, particularly if the keys are manually entered. One way to check this is to use set arithmetic, where one set is the keys from the first dataframe and another set is the keys from the second dataframe. By using set differencing operations, it is possible to determine keys which have no referent, and thus address simple formatting issues.
- **Date and time.** Handling dates and times can be very problematic because of difference in formatting (e.g. MM/DD/YYYY vs DD/MM/YY, etc.), timezones, leap years, and daylight savings / standard time shifts. A rule of thumb for performing calculations on these values is to never write routines to do them – instead, use built-in libraries to do so. These built-in libraries will handle the many, many special cases that arise in these calculations.

Another rule of thumb is to immediately convert any date / time values read in to its equivalent value in Coordinated Universal Time (UTC), then perform all calculations in the stan-

dard value. This completely sidesteps the special cases described above. If necessary, convert to local time only when needed for display or other output (graphs, etc.).

- **Values out of range.** Some range errors can be easy to detect, for example, empty values, negative values, or values that are orders of magnitude above legal or expected limits. Other types of errors are much more difficult to detect, and require either domain knowledge or statistical techniques. An issue with statistical techniques is that it is difficult to determine whether an aberrant data value is an error or a true outlier. In the former case, the value should be rejected. In the latter case, it should be kept in the analysis data set. Techniques for statistical outlier detection include z-scores [133], interquartile range [132] or Statistical Process Control [114].
- **Units.** It is important that the units for data values be clearly understood, otherwise calculation errors are extremely likely. Techniques for specifying the units include entries in a data dictionary, augmenting the column header with unit information, or including a column in the input data specifying the units.