THESIS

ON THE CERTAINTY FRAMEWORK FOR CAUSAL NETWORK DISCOVERY WITH APPLICATION TO TROPICAL CYCLONE RAPID INTENSIFICATION

Submitted by

Michael DeCaria

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2022

Master's Committee:

Advisor: Peter Jan van Leeuwen

Christine Chiu
Elizabeth Barnes
Imme Ebert-Uphoff

ABSTRACT

ON THE CERTAINTY FRAMEWORK FOR CAUSAL NETWORK DISCOVERY WITH APPLICATION TO TROPICAL CYCLONE RAPID INTENSIFICATION

Causal network discovery using information theoretic measures is a powerful tool for studying new physics in the earth sciences. To make this tool even more powerful, the certainty framework introduced by van Leeuwen et al. (2021) adds two features to the existing information theoretic literature. The first feature is a novel measure of relative strength of driving processes created specifically for continuous variables. The second feature consists of three decompositions of mutual information between a process and its drivers. These decompositions are 1) coupled influences from combinations of drivers, 2) information coming from a single driver coupled with a specific number of other drivers ($m$links), and 3) total influence of each driver. To represent all the coupled influences, directed acyclic hypergraphs replace the standard directed acyclic graphs (DAGs).

The present work furthers the interpretation of the certainty framework. Measuring relative strength is described thermodynamically. Two-driver coupled influence is interpreted using DAGs, introducing the concept of separability of drivers' effects. Coupled influences are proved to be a type of interaction information. Also, total influence is proved to be nonnegative, meaning the total influences constitute a nonnegative decomposition of mutual information. Furthermore, a new reference distribution for calculating self-certainty is introduced. Finally, the framework is generalized for variables that are continuous with one discrete mode, for which partial Shannon entropy is introduced.

The framework was then applied to the rapid intensification of Hurricane Patricia (2015). The hourly change in maximum tangential windspeed was used as the target. The four drivers were outflow layer (OL) maximum radial windspeed (u_u), boundary layer (BL) radial windspeed at radius of maximum wind (RMW) (u_l), equivalent potential temperature at BL RMW (θ_e), and the temperature difference between the OL and BL (ΔT). All variables were azimuthally averaged. The drivers explained 45.5% of the certainty. The certainty gain was 35.8% from θ_e, 24.5% from ΔT, 24.0% from u_u, and 15.7% from u_l. The total influence of θ_e came mostly from inseparable effects, while the total influence of

u_u came mostly from separable effects. Physical mechanisms, both accepted in current literature and suggested from this application, are discussed.

ACKNOWLEDGMENTS

First and foremost, I want to express what an honor and privilege it is to work with Professor Peter Jan van Leeuwen. Working on our framework has been an amazing experience. I have gained so much from studying the various applications he has found. My understanding and appreciation for the physics of the atmosphere, earth science in general, and the interactions tying them all together have grown deeper, and my ability to convey science has increased exponentially as he challenges every last detail. This work would not be possible without him, his guidance, and his quest for knowledge.

I want to express my gratitude to Professor Christine Chiu and Matthew Lang for their role in studying convective initiation. Their unique knowledge set has created an interesting project. And, their patience for my implementing and revising the framework has already yielded a more robust implementation, and it will yield a very unique application of our framework when we resume this project in the future.

Many thanks to Dandan Tao and Professor Michael Bell for guiding our study of Hurricane Patricia! Yet again was I repeatedly met with patience and a wealth of knowledge while studying such a highly complex system.

Thank you to Chih-Chi Hu and WeiTing Hsiao for being amazing scientific peers as well as my COVID circle. Thank you to Elizabeth Dulac for C++ assistance, and to my brother Victor DeCaria for mathematical guidance. And, thank you to the many friends at ATS for making the department a home.

And, I would be remiss to not thank the very understanding members of my committee, Professor Christine Chiu, Professor Elizabether Barnes, and Professor Imme Ebert-Uphoff. I look forward to your thoughts and insights into this work.

Studying the rapid intensification of Hurricane Patricia (2015) was supported by the Programs of Research and Scholarly Excellence (PRSE) summer 2020 fellowship. The data came from a 60-member ensemble Weather Research and Forecasting model simulation designed by Yue “Michael" Ying and analyzed by Dandan Tao. The drop-sonde and radar data with which the ensemble was initialized came from the Office of Naval Research Tropical Cyclone Intensification field campaign (2015), which was made available by Michael Bell.

This thesis is typeset in LaTeX using a document class created by Leif Anderson and modified by Christopher Slocum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Causal Network Discovery

Causal inference is now an exact science. Within causal inference lie many subfields, of which this work focuses on causal network discovery (CND). As the name implies, causation can be expressed in terms of separate processes connected to one another via causal pathways, and CND seeks to discover the full network of causal pathways. There are many methods for CND, from interventional (Pearl 2000) to observational (Wiener 1956; Granger 1963, 1969) forms of analysis. The present work uses the recent observational framework from van Leeuwen et al. (2021), which 1) introduces how select processes nonlinearly couple to drive a single target variable and 2) determines the completeness of a study. As many processes in the natural world interact with one another rather than act individually, introducing how to measure nonlinear coupling is a great leap forward for CND, especially in the earth sciences. Beyond this, the completeness of a study indicates how thoroughly the select processes represent the underlying physics that drives the target. Thus, van Leeuwen et al. (2021) revolutionizes CND while adding to existing CND methods.

This chapter introduces the necessary components for CND. Section 1.1 introduces what a causal network is, how to represent it, and briefly how to use it. Section 1.2 discusses discovery methods, focusing on measuring the strength of a causal connection. In Section 1.3, current attempts to represent and measure coupled causation are discussed. Section 1.4 gives a few examples of using CND in the earth sciences. Section 1.5 briefly introduces the necessary information theoretic measures for the rest of the thesis.

Beyond this chapter, the thesis is organized as follows. Chapter 2 summarizes the certainty framework in (van Leeuwen et al. 2021), while Chapter 3 furthers the framework's interpretations. Chapter 4 discusses implementing the framework in code. In Chapter 5 is the application to tropical cyclone rapid intensification. Chapter 6 generalizes the framework to targets that are mostly continuous with one discrete mode. Chapter 7 summarizes this work and suggests related future work.

1.1 Causal Networks

A causal network (CN) shows how processes affect one another. For instance, how does the El Niño Southern Oscillation (ENSO) affect land surface temperatures in different places around the world? How does land surface temperature affect daily precipitation rates? How do local precipitation rates affect the global hydrological cycle? The uses of such a CN are vast and important. Including economic

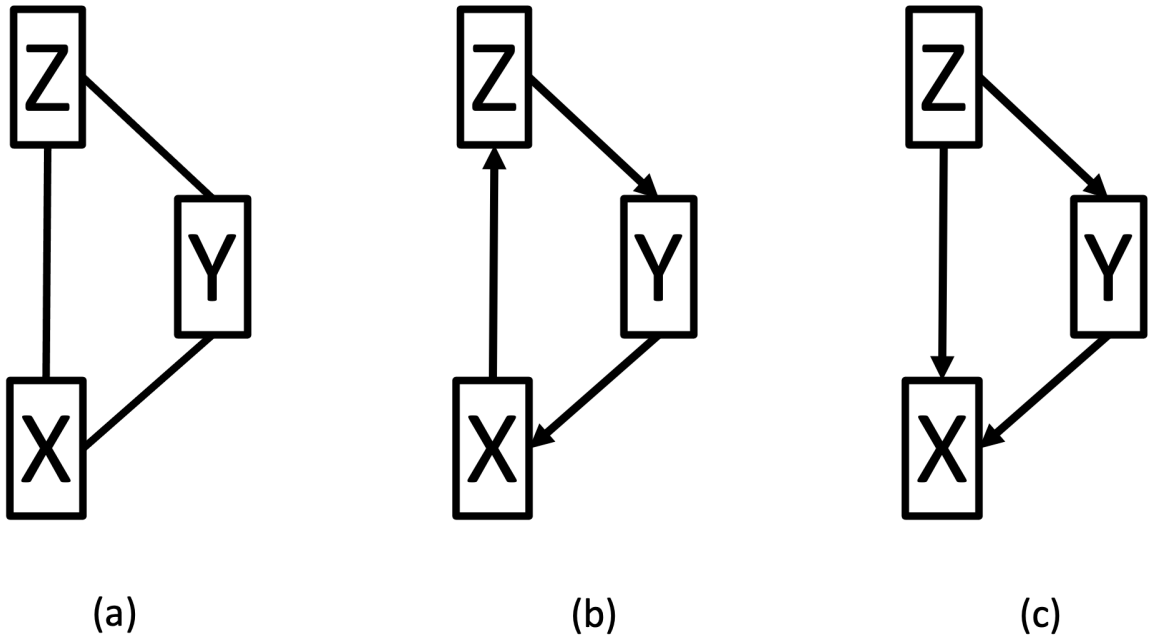

FIG. 1.1. Different types of graphs. From left to right, the graphs are undirected, directed but cyclic, and directed and acyclic.

activity in this CN could help determine humanity's effect on this part of the natural world. CNs, and how to discover them, are thus an essential element of causal inference.

Often, graphical models are used to represent CNs. Most often, these models are directed acyclic graphs (DAGs). A *graph* contains nodes and edges connecting the nodes. (See Fig. 1.1 (a).) For a CN, the nodes represent processes, and the edges represent the relationship between processes. In a *directed* graph, or *digraph*, each edge is an arrow, pointing from one process to another. (See Fig. 1.1 (b).) For two processes X and Y, Y causing X is written as $Y \to X$. In the case that Y indirectly causes X, i.e. Y drives processes that either directly or indirectly drive X, then there is a *forward*, or *causal*, *path* from Y to X, written $Y \to \cdots \to X$. A single directed edge also constitutes a forward path. When Y indeed causes X, Y is an *ancestor* of X, and X is a *descendent* of Y. When Y directly causes X, then Y is a *parent* of X, and X is a *child* of Y. Furthermore, when X and Y are connected by a common ancestor, then the path through the common ancestor is a *backdoor path*. *Acyclic* means that, if there is a forward path from Y to X, then there cannot be a forward path from X to Y. (See Fig. 1.1 (c).) There is a small caveat when representing feedbacks, but that is beyond the scope of this work. I revisit DAGs in Section 1.3 in discussing how to represent coupled causation.

Two forms of analysis can be used to calculate how changes in one process affect another process in the same network. The most popular is *interventional analysis*, which calculates the results of an intervention. The standard method for intervention by experimentation for the past several decades is the randomized control trial (RCT). In an RCT, there is a control group and one or more experimental groups. The control group receives the standard treatment, which may be no treatment at all, and is therefore *controlled*. Each experimental group, however, receives a different nonstandard treatment. Which group receives which treatment is randomized prior to experimentation, theoretically removing variation between groups. Assuming, then, that the groups are identical prior to the experiment, statistically significant differences between groups must be the result of the differing treatments. In the earth sciences, where 1) control is difficult, 2) small interventions may have little effect, and 3) large interventions are unethical, numerical sensitivity studies permit analyzing interventions without doing experiments in the physical world. An example could be anomalously heating a certain region and letting the simulation run.

In light of the difficulty, ethics, and expense of experiments, as well as attempting to implement artificial human-like intelligence, Pearl (1995, 2000) introduced do-calculus. According to do-calculus, intervening on a process severs the ties between the process and its parents, thereby creating a different CN. Do-calculus, explicitly recognizing this, aims to use the undisturbed network to calculate the effect of an intervention. In other words, it aims to calculate interventional causation from observational data. While Pearl (1995, 2000) give example CNs on which certain interventions still require experimentation, fewer experiments are needed, and each experiment is potentially minimally invasive. The end result is what would happen if a real intervention occurred, placing do-calculus in the realm of interventional analysis.

The other form is *observational analysis*, in which interventions are not used. Common forms of observational causation include Wiener-Granger Causality (Wiener 1956; Granger 1963, 1969), nonlinear Granger Causality using transfer entropy (Schreiber 2000), and convergent cross mapping (Sugihara et al. 2012). While many CND analyses assume, either implicitly or explicitly, that causation cannot be defined in the absence of an intervention, van Leeuwen et al. (2021) argue that permitting interventions at all may hinder analysis. Because intervening on a process results in a different CN than what occurs naturally, the results from interventional analysis may not accurately reflect results from the underlying system.

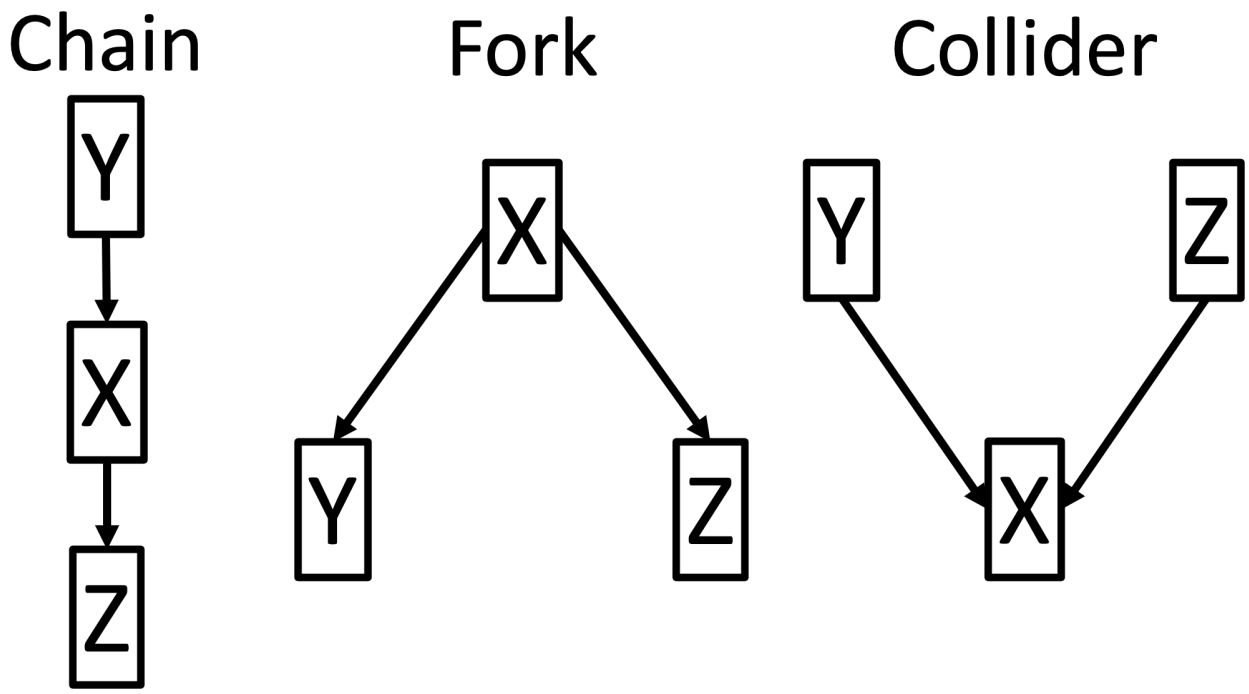

FIG. 1.2. Three simplest causal networks of three variables. From left to right, they are a chain, a fork, and a collider, where X is always between Y and Z.

To illustrate the distinction, consider a closed system. Every state of the system must be a result of the system. Observational analysis treats the state as coming from a closed system, thereby accurately reflecting this situation. Interventional analysis, however, opens the system to set the state, whether or not the previous state of the system would have generated the new state. If the environmental state does not come from the previous state, then the results may not reproduce what would occur in the intervention-free system.

1.2 Discovering Causal Networks

Discovering causal networks is much different from using a causal network. To demonstrate how to discover networks, consider the following basic causal structures. These are chains, forks, and colliders, each of which involve three processes, say X, Y, and Z. (See Fig. 1.2.) The edges $Y \to X$ and $X \to Z$ constitute a chain, $Y \to X \to Z$. Reversing the first of these edges, so $X \to Y$, results instead in a fork, $Y \leftarrow X \to Z$. Reversing both edges in a fork, so $Y \to X$ and $Z \to X$, results in a collider, $Y \to X \leftarrow Z$. This gives a sense of the flow and origin of information in a CN. In a chain, information flows from previous generations to later generations. A fork, instead, is an instance of a common source of information for two different forward paths in a CN. In contrast to both a chain and a fork, each of which has one

source of information, a collider has two or more sources of information with forward paths to the same process.

To reason with both simple and complicated CNs, we use the concept *d-separation*. For processes X, Y, and Z, X d-separates Y and Z if and only if X blocks all forward or backdoor paths between Y and Z, where X may be empty or contain many variables. In a CN, this generally means X contains no common descendents of Y and Z and only ancestors of Y or Z, including common ancestors. When given only three processes, d-separation distinguishes chains and forks from colliders. In the chain and fork, X d-separates Y and Z because it is a parent of at least one of them, and Y and Z are not otherwise related. What would then distinguish the chain from the fork are assumptions about which variables precede others. But, in the collider, Y and Z are already d-separated by the empty set, and X does not d-separate them. Instead, conditioning on X creates a path between Y and Z. While this may seem counter intuitive, consider the collider in Fig. 1.2. Conditioning on an observed value of X means keeping it fixed while Y and Z are free to vary. Thus, variations in Y must be offset by variations in Z, thereby creating a dependence between them. Thus, these simple structures and d-separation allow us to intuit effects from changes in a process. Appendix A.1 discusses examples of d-separation, and for a full discussion of d-separation, see Pearl (2000), Section 1.2.3..

To actually discover complicated causal structures, Spirtes et al. (2000) list three major causal assumptions that go beyond statistical reasoning. The first is *causal sufficiency*, which requires that all common causes of processes in a network are included in the analysis. Inversely, causes that are not common may be neglected and treated as independent noise. The second is the *causal markov condition*, which states that d-separation of two processes on the graph by one of their parent sets implies conditional independence of the two processes given the same parent set. This also means that conditional dependence of the two processes given either parent set implies they are not d-separated by either parent set. The third assumption is *faithfulness*, which essentially means that, if the causal markov condition is met, then the converse of the causal markov condition is also met. That is, conditional independence of two processes given a set of variables implies that the set of variables d-separates the two processes. To infer causal networks from time series, Runge (2018) adds the assumption of *causal stationarity*, which assumes the causal network discovered for one time is the same throughout time.

Many causal discovery algorithms incorporate these three or four assumptions. Some algorithms, e.g. PC (Spirtes and Glymour 1991), start with a fully connected CN and prune connections of d-separated variables, while others, like greedy search algorithm (Chickering 2002) or optimal causation

entropy (Sun et al. 2014), start with a minimally connected network and add strong connections. Another popular method is PC-MCI, or TIGRAMITE, in which the PC algorithm generates a preliminary causal network, and then the parent set for each process is reviewed and revised (Runge 2015).

This leads to discussing how to measure the strength of a connection in the first place to determine whether or not a connection is negligible. Strength can be measured either in an absolute sense or in a relative sense. Using absolute strength often involves some threshold value for determining whether or not a connection exists. This is how information theoretic measures are often evaluated.

Using relative strength, by contrast, is often preferable to absolute measures of strength. To show this, consider linear regression analysis for driver Y and target X. The correlation coefficient, the typical measure of strength, is given by

$$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{1.1}$$

where σ_x and σ_y are the standard deviations of X and Y, respectively, σ_{xy} is their covariance, and the value of R is between -1 and 1. Covariance is an absolute measure of strength. Meanwhile, R^2 is interpreted as the proportion of variance explained, thereby making it a measure of relative strength. Suppose X and Y have covariance $\sigma_{xy} = 3$, and further say that the variance of Y is $\sigma_y^2 = 1$. Knowing $\sigma_{xy} = 3$ reveals hardly anything about how strongly Y drives X, as the variance of X, σ_x^2, remains unknown. If $\sigma_x^2 = 9$, then the correlation coefficient $R = 1$, and Y completely determines X. If instead $\sigma_x^2 = 900$, then $R = 0.1$, or $R^2 = 0.01$, and we might question if Y even causes X. Thus, not only does R^2 suggest how completely Y determines X, but it also suggests how complete the physics in a study is overall.

As the framework in van Leeuwen et al. (2021) is based on information theoretic measures, this work necessarily focuses on these measures. When using information theoretic measures, we must first determine whether the target process is discrete or continuous. If the target is discrete, then the parent set explains the Shannon entropy of the target (McGill 1954). If the target is continuous, then the parent set explains the total certainty of the target (van Leeuwen et al. 2021). In either case, mutual information between the target and the parent set is the amount of explanation offered. Dividing by the appropriate value then shows how complete the parent set is in explaining the target. This contribution of van Leeuwen et al. (2021) will be made more intuitive in Section 3.1, while Chapter 6 generalizes the framework to handle variables like precipitation, which are partly continuous with one discrete mode.

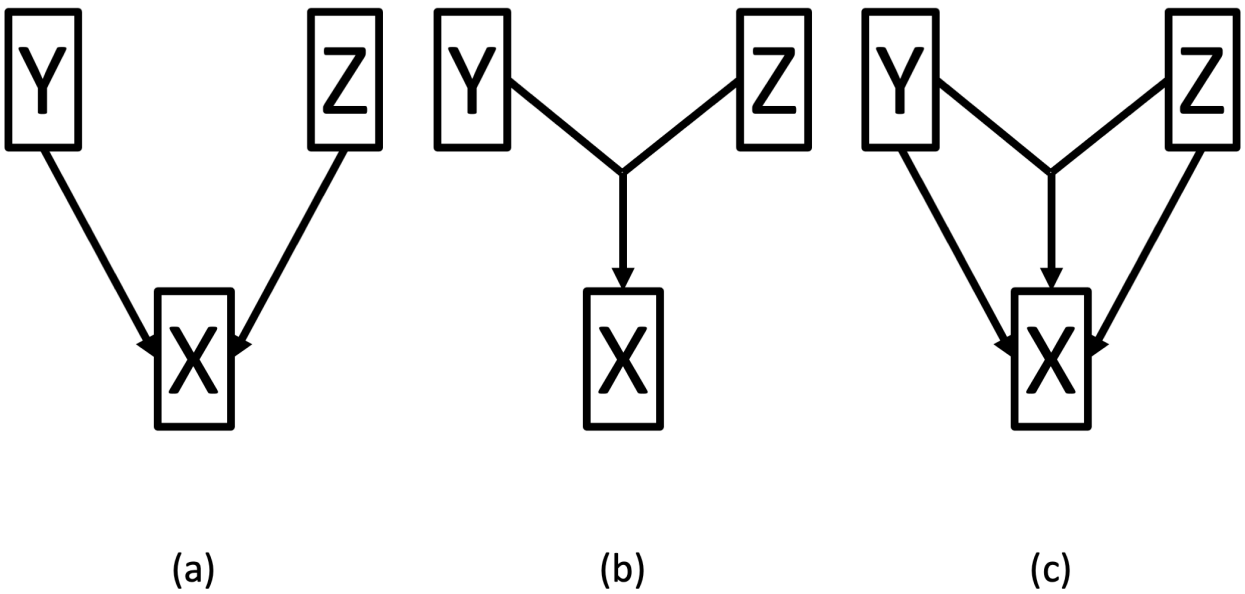

FIG. 1.3. Comparison of (a) a directed graph and (b,c) directed hypergraphs. In the graph, Y and Z each have individual edges directed into X. In (b), however, Y and Z instead have a single shared edge directed into X. In (c), another valid hypergraph, all edges are present

1.3 Coupled Causation

Calculating coupled causation explicitly recognizes that a mechanism which causes a process may require more than one driver process. To motivate the necessity of coupled causation, consider as target convective initiation (CI), and as drivers convective available potential energy (CAPE) and convective inhibition (CIN). A standard graph representation is a collider, CAPE→CI←CIN. (See Fig. 1.3 (a), where Y and Z are CAPE and CIN, and X is CI.) The implication is almost that of logical operator OR, i.e. either CAPE is large enough or CIN is small enough for CI to occur. The reality is that both thresholds need to be satisfied.

Representing coupled causation requires generalization beyond the standard graph. This requires introducing directed acyclic *hyper*graphs, called *causal webs* in van Leeuwen et al. (2021). In the example, the causal web would have the standard graph as well as lines exiting CAPE and CIN which come together at a vertex, and an arrow pointing from this vertex to CI, written as {CAPE,CIN} →CI. (See Fig. 1.3 (b).) It is also possible that shared and individual paths are present. (See Fig. 1.3 (c).) As shown in van Leeuwen et al. (2021), there are as many such vertices as there are combinations of drivers.

There are few frameworks that attempt to describe this, and only three that use information theory. One framework not using information theory is the sufficient-component cause (SCC) framework, introduced by Rothman (2017) and named such by Koopman (1981). This framework addresses something critically lacking in DAG notation: using AND versus OR. A sufficient cause is any process that, when certain conditions are met, results in a change of state in another process. The sufficient cause may contain any number of component causes, each of which is equally required for the sufficient cause. In the above example, CAPE and CIN are component causes to a sufficient cause. Not only must there be something to convect, i.e. CAPE is nonzero, but CIN cannot be too large. If only one of these conditions is met in this example, then CI will not occur. The SCC framework focuses on binary drivers and binary targets, so applying it to earth science would require generalization.

One of the frameworks using information theory is partial information decomposition (PID), proposed by Williams and Beer (2010). PID claims to be able to decompose mutual information into multiple nonnegative components, with unique contributions from individual drivers, synergistic contributions from combinations of drivers, and redundant contributions from all drivers. Unfortunately, these terms are not fully defined, so many interpretations exist. While the theory enforces that all terms are nonnegative, Barrett (2015) show that, with a multivariate Gaussian system with one target and two drivers, many popular interpretations of PID are degenerate and yield that the weakest driver contributes zero unique information. This suggests that PID is not appropriate for continuous variables in general.

Another framework using information theory is *multivariate information*, a term proposed originally by McGill (1954) and later negated and called *interaction information*. Interpreting interaction information has been troubled for decades, as it may be positive or negative. Interpretation for two drivers is relatively straight forward, which I detail in Section 3.2.1. Beyond two drivers, interpretation is still lacking.

Needless to say, the third framework using information theory to address coupled causation is the certainty framework. This is introduced in the next chapter, while Chapter 3 details how I have furthered interpretations since van Leeuwen et al. (2021).

1.4 Causal Discovery in the Earth Sciences

Causal discovery methods have benefited the earth sciences for over a decade. For example, causal discovery can evidence teleconnections in surface pressure anomalies (Runge et al. 2019a; Ebert-Uphoff and Deng 2012), feedbacks between sea ice and atmospheric patterns (Kretschmer et al. 2016;

Matthewman and Magnusdottir 2011; Strong et al. 2009), between the MJO and the NAO (Samarasinghe et al. 2021; Barnes et al. 2019), and how ENSO affects temperatures throughout the world (McGraw and Barnes 2018). This is but a very short list of causal discovery in the earth sciences to date. For a more complete review, see Runge et al. (2019b). What van Leeuwen et al. (2021) and the present thesis do differently is explicitly address the issue of variables coupling to drive a target as well as introduce information theoretic measures to evidence physical completeness of a study. What all the above studies show is that, while causation is only evidenced until physical pathways are identified, discovering connections via causal discovery in the first place can lead to hypotheses about said physical pathways!

1.5 Information Theory Primer

Information theory was created to study the degradation of information as it is transmitted via phone lines. Now, information theory has many applications in the sciences, from psychology and neuroscience to the earth sciences, often to infer causation. Familiarity with information theory is essential to understand the certainty framework, as there are several information theoretic measures relevant to the framework. These are Shannon entropy and conditional Shannon entropy, mutual information and conditional mutual information, interaction information, and Kullback-Leibner (KL) divergence. I will discuss each of them below. Throughout this primer, X is always the target, and Y and Z are drivers, all assumed continuous.

Shannon entropy is essentially a measure of uncertainty. The Shannon entropy of X is

$$H(X) = -\int p(x) \log p(x)\, dx, \tag{1.2}$$

where p is the probability density of X. One property of Shannon entropy of continuous variables is that it depends on the spread of X. For example, if we define $X' = aX$, for some $a > 0$, then

$$H(X') = -\int p(x') \log p(x')\, dx' = -\int p(x) \log\left(\frac{1}{a} p(x)\right) dx = \log a + H(X). \tag{1.3}$$

If $a > 1$, i.e. the spread of X' is larger than the spread of X, then $H(X') > H(X)$. If $a < 1$, i.e. the spread of X' is smaller than the spread of X, then $H(X') < H(X)$. A consequence of this is that, unlike Shannon entropy of discrete variables, Shannon entropy of continuous variables may be negative.

Conditional Shannon entropy is a measure of uncertainty after conditioning on a variable. Thus, the conditional Shannon entropy of X given Y is

$$H(X|Y) = -\int p(x,y) \log p(x|y)\, dx\, dy. \tag{1.4}$$

Note that $p(x,y)$ is used for the weighting, but $p(x|y)$ is inside the logarithm. Whether the variables are discrete or continuous, conditioning on a variable never increases the entropy, i.e. $H(X|Y) \leq H(X)$.

Mutual information is a measure of dependence between two variables. One way to define it is the decrease in Shannon entropy of one variable by conditioning on the other,

$$I(X;Y) = H(X) - H(X|Y) = \int p(x,y) \log \frac{p(x|y)}{p(x)}\, dx\, dy \geq 0, \tag{1.5}$$

with $I(X;Y) = 0$ only when X and Y are independent, i.e. $p(x|y) = p(x)$ for all x and y. Mutual information is symmetric, that is $I(X;Y) = I(Y;X)$, and non-negative.

Conditional mutual information is a measure of conditional dependence between two variables given a third variable. For X, Y, Z, the mutual information between X and Y conditioned on Z is

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \int p(x,y,z) \log \frac{p(x|y,z)}{p(x|z)}\, dx\, dy\, dz \geq 0, \tag{1.6}$$

attaining zero only when X and Y are conditionally independent given Z. Unlike with Shannon entropy, conditioning on a variable does not always decrease mutual information.

In fact, studying how conditioning changes mutual information is how interaction information was first conceptualized (McGill 1954). (The modern definition of interaction information negates McGill's definition of multivariate information.) The interaction information of three variables is the change in mutual information between any two conditioned on the third,

$$II(X,Y,Z) = I(X;Y) - I(X;Y|Z) = I(X;Z) - I(X;Z|Y) = I(Y;Z) - I(Y;Z|X), \tag{1.7}$$

where II is the interaction information. Because mutual information does not always decrease by conditioning on a third variable, interaction information may be positive or negative. Interaction information can be generalized to any number of variables as

$$II(X_1, \ldots, X_{n-1}, X_n) = II(X_1, \ldots, X_{n-1}) - II(X_1, \ldots, X_{n-1}|X_n), \tag{1.8}$$

where $II(X_1,\ldots,X_{n-1}|X_n)$ is a conditional interaction information, and X_n is used as the condition without loss of generality. Interpreting interaction information is addressed in Section 3.2.

KL divergence is sometimes also called relative entropy. It is a measure of difference between the observed probability density and some chosen reference density, q. In essence, q defines a base level of entropy, and any departure of p from q is a source of uncertainty, hence KL divergence is indeed *relative* entropy. KL divergence is expressed as

$$D_{KL}(p||q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx = -E_p(\log q) - H(X) \geq 0, \tag{1.9}$$

where $E_p(\log q)$ is the cross-entropy and $H(X)$ is the Shannon entropy of X. A KL divergence is zero only when $p = q$, and is positive otherwise. Mutual information can also be thought of as a specific KL divergence, measuring the departure of a variable's conditional distribution from its marginal distribution.

CHAPTER 2

The Certainty Framework

Before discussing my work beyond van Leeuwen et al. (2021), this chapter introduces the main concepts of the framework, henceforth the certainty framework. First, certainty is discussed as an information theoretic quantity, followed by the standard choice of reference density. Then, the three decompositions of mutual information are described. Third, normalizing for relative strength is discussed. Finally, my contributions to van Leeuwen et al. (2021) are made explicit. Throughout this chapter and the next, X is the target and $\mathbf{Y}$ is the set of all drivers.

The gain in explaining the evolution of the target X by a set of drivers Y is expressed as

$$W(X|\mathbf{Y}) = I(X;\mathbf{Y}) + W(X), \tag{2.1}$$

where $W(X|\mathbf{Y})$ is the *total certainty* of X and $W(X)$ is the *self-certainty* of X, as introduced in van Leeuwen et al. (2021). This equation can be compared to the equation for Shannon entropy, $H(X) = I(X;\mathbf{Y}) + H(X|\mathbf{Y})$, but has the advantage that all terms are always nonnegative. Self-certainty is a Kullback-Leibner (KL) divergence. That is,

$$W(X) = D_{KL}(p_X||q_X) \geq 0, \tag{2.2}$$

where p_X is the true distribution of X and q_X is the reference distribution. (For more on KL divergence, see the information theory primer, Section 1.5.) Mutual information between the target X and the drivers $\mathbf{Y}$ is added to the self-certainty of X to form the total certainty of X. Then, total certainty is also a KL divergence,

$$W(X|\mathbf{Y}) = D_{KL}(p_{X|\mathbf{Y}}||q_X) \geq I(X;\mathbf{Y}), \tag{2.3}$$

where $p_{X|\mathbf{Y}}$ is the distribution of X conditioned on all the drivers. This gives the interpretation that $I(X;\mathbf{Y})$ is *total certainty explained* or *certainty gain*.

Because self-certainty is a KL divergence, a reference probability density must be determined for X. In van Leeuwen et al. (2021), we suggest the reference density should be as wide and as featureless as possible. Therefore, we suggest using a Lorentz-Cauchy distribution,

$$q(x) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (x - x_0)^2}, \tag{2.4}$$

where x_0 is the sample mean of X, and the spread parameter $\gamma = \sqrt{e/8\pi}\sigma_X$, where σ_X is the standard deviation of X. Using σ_X ensures that self-certainty is determined only by the shape, and not the spread, of the target's distribution.

Included in the certainty framework are three decompositions of mutual information. The first decomposition is direct influences and coupled influences, called 1links and (multivariate) $m$links in van Leeuwen et al. (2021). The direct influence of a driver, Y, is the mutual information between it and X conditioned on all remaining drivers, that is

$$I_{Y|\mathbf{Y}\setminus\{Y\}} = (Y \to X)_{1link} = {}^{X}_{Y}M_{\mathbf{Y}} = I(X;Y|\mathbf{Y}\setminus\{Y\}), \tag{2.5}$$

where the left two notations are from van Leeuwen et al. (2021) and the M-notation is my invention. The coupled influence of Y with another driver, Z, is the mutual information they have with X conditioned on all other drivers, minus their individual direct influences. That is,

$$I_{Y,Z|\mathbf{Y}\setminus\{Y,Z\}} = {}^{X}_{YZ}M_{\mathbf{Y}} = I(X;Y,Z|\mathbf{Y}\setminus\{Y\}) - {}^{X}_{Y}M_{\mathbf{Y}} - {}^{X}_{Z}M_{\mathbf{Y}}, \tag{2.6}$$

where again the leftmost notation is from the paper. In general, for a combination of drivers $J \subseteq \mathbf{Y}$, its coupled influence on X is

$$I_{J|\mathbf{Y}\setminus J} = {}^{X}_{J}M_{\mathbf{Y}} = I(X;J|\mathbf{Y}\setminus J) - \sum_{j\subset J}^{|j|>0} {}^{X}_{j}M_{\mathbf{Y}}, \tag{2.7}$$

where $I_{J|\mathbf{Y}\setminus J}$ would be the paper's notation, and $|j|>0$ means that the subsets used in the summation are not empty. The direct and coupled influences sum to the mutual information between all drivers and the target, $I(X;\mathbf{Y})$.

The next two decompositions are derived from the direct and coupled influences. The first is $m$links, called (single-variate) $m$links in van Leeuwen et al. (2021). As shown above, the 1link of a driver is its direct influence. The 2link of a driver, Y, is the sum of all coupled influences in which Y couples with exactly one other driver. We can express this as

$$(Y \to X)_{2links} = \sum_{J\subseteq\mathbf{Y}, Y\in J}^{|J|=2} {}^{X}_{J}M_{\mathbf{Y}}. \tag{2.8}$$

Note that the summation expresses both that Y is included ($Y \in J$) and that exactly one other driver is involved ($|J|=2$). In general, $m$links for any m are calculated similarly,

$$(Y \to X)_m = \sum_{J\subseteq\mathbf{Y}, Y\in J}^{|J|=m} {}^{X}_{J}M_{\mathbf{Y}}, \tag{2.9}$$

where m appears wherever 2 did before. The 2links of Y can be seen as the influence that Y has by coupling with exactly one other driver, whatever driver that is, and so on for higher order $m$links. When summing all the $m$links of all drivers, each $m$link must be divided by m to avoid counting the value multiple times. This can also be thought of as dividing each coupled influence of m variables by m, which then is evenly distributed among the $m$links of the m drivers. This is because the m drivers contribute symmetrically to the value, so the value must be divided evenly among them. Thus, all the $m$links of all drivers, where each $m$link is divided by m, sum to $I(X;\mathbf{Y})$.

The final decomposition is into the total influences, called total contributions in van Leeuwen et al. (2021). The total influence of a driver, Y, is the sum of all its $m$links that are normalized by the number m of drivers involved. If there are n drivers in the study, then the total influence of a driver is

$$(Y \to X)_{total} = \sum_{m=1}^{n} \frac{1}{m} (Y \to X)_{mlinks}. \tag{2.10}$$

The total influences of all drivers sum to $I(X;\mathbf{Y})$.

Normalizing each term by $W(X|\mathbf{Y})$ then expresses each term's relative contribution to the certainty of X. Normalizing the coupled influence of a combination of drivers shows how much certainty comes from that particular combination of drivers. Normalizing a driver's $m$link shows the relative influence of that driver when coupled with $m-1$ other drivers. The relative strength of a driver overall is shown by normalizing that driver's total influence. Normalizing $I(X;\mathbf{Y})$ measures how complete a particular set of drivers is. By contrast, normalizing $W(X)$ indicates the relative influence of noise and unknown drivers.

My second authorship for van Leeuwen et al. (2021) is due to my overall contributions to the framework, involving myself in discussions on reference densities, and being actively involved in editing the text and formulating replies to reviewers. Specific contributions to the framework include showing that Shannon entropy cannot be used for normalization when the target is continuous, and developing equation (2.3). Specific contributions to shaping the reference density include 1) advocating that the density changes with the sample standard deviation so that distributions with equivalent shape but different spread will have equal self-certainty, and 2) the width parameter expression for the Lorentz-Cauchy reference density, which is now standard for the framework. I also implemented the code for all calculations in the paper, and this implementation is discussed in Chapter 4.

CHAPTER 3

New developments in the certainty framework

In chapter 1, I outlined causal network discovery (CND). Then, in chapter 2, the certainty framework from van Leeuwen et al. (2021) was introduced. In short, it provides a way to calculate relative causal strength using information theory for continuous variables, allowing both for easy comparison of results between studies and for more meaningful representations of the discovered relationships. Included in the framework are three decompositions of mutual information into contributions from individual and coupled processes.

In this chapter, I will outline my development of the concepts since van Leeuwen et al. (2021). Section 3.1 holds a thermodynamic interpretation of calculating relative strength for both traditional measures and for certainty. My interpretation of the coupled influences, as well as discussions of single-variable $m$links and total influences are in Section 3.2. On top of this, a new reference density for highly noisy variables is introduced and discussed in Section 3.3.

3.1 Certainty: The New Norm

In Section 1.2, I discussed how measuring relative strength is better than measuring absolute strength. The nature of the relative strength measure introduced in van Leeuwen et al. (2021), however, is different from most measures of relative strength. To illustrate and justify the distinction, I liken the differences between traditional measures and certainty to the differences between a heat engine and a heat pump, respectively, which is shown below.

3.1.1 Relative Strength and the Heat Engine

In a heat engine, a hot reservoir at temperature T_H emits heat energy Q_H, some of which the engine extracts and turns into usable work, W. Whatever is not turned into work is passed as waste heat into a cold reservoir at temperature T_C. This process is illustrated in Figure 3.1 on the left. The thermal efficiency of the engine is the ratio of work to emitted heat, W/Q_H, which is upper bounded by $(T_H - T_C)/T_H$. Friction in the engine makes it such that the theoretical maximum is never attained, i.e. $W/Q_H < (T_H - T_C)/T_H$.

The analogy of the heat engine to traditional methods of measuring relative strength starts with the observed target in place of the hot reservoir. (See Fig. 3.1, right side.) Some of the analogy is method-dependent, so I will continue the analogy using linear regression. The observed target has a variance,

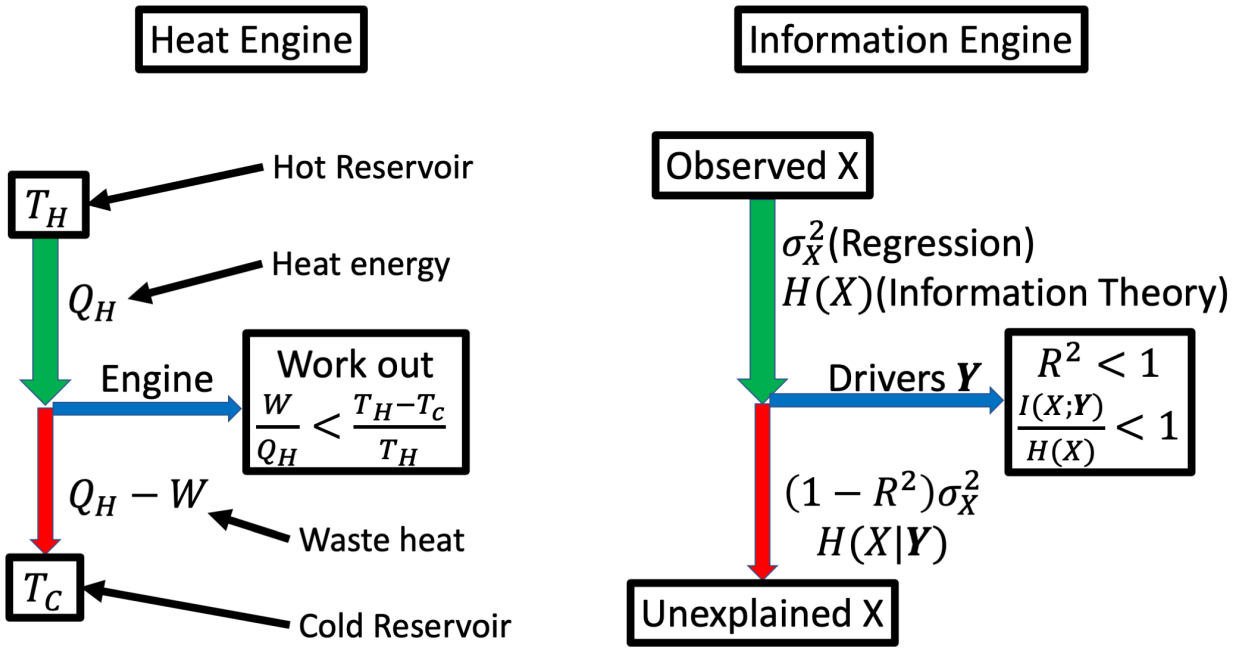

FIG. 3.1. Side by side comparison of the heat engine (left) and the analogical information engine (right). The thermal efficiency of the heat engine is shown as work extracted, W, over the total emitted heat energy, Q_H, with theoretical upper bound of $(T_H - T_C)/T_H$, where T_H and T_C are the temperatures of the hot and cold reservoirs, respectively. The information engine diagram shows the analogical measures for both regression and information theory, where the regression measures appear above the information theory measures. Note that, while the upper bound of thermal efficiency is less than 1 whenever $T_C > 0$, the upper bound for R^2 and proportion of Shannon entropy explained is always 1.

equivalent to the emitted thermal energy. The observed drivers are the engine itself, and their covariance with the target is the work extracted. What they fail to explain is equivalent to the waste heat. This wasted explanation ultimately enters the unexplained target, which is in place of the cold reservoir. The relative amount of explanation is the familiar R^2, which is the covariance squared divided by the variances of the target and drivers.

Regardless of the method, there are three main reasons the drivers fail to fully explain the target. The first is physical, in that the selected drivers are either incorrect or not complete. This means that some physics in driving the target remains hidden. Just as heat engines are specialized based on their intended use, using incorrect drivers or an incomplete driver set will not perform well at explaining a target. The next two reasons are related, as they regard noise in the drivers and target. Noise is inherent in any data, whether from observation or from numerical inaccuracies. The roles are distinct in this analogy, though. Noise in the drivers acts as friction in the engine, reducing the ability of the drivers

to extract the theoretical upper bound of information content. Noise in the target, however, acts as the temperature of the cold reservoir. Note that the upper bound of thermal efficiency is 1 when $T_C = 0$. Any other temperature, which cannot be negative, yields an upper bound less than 1. Similarly, if the target has any noise, then even a correct and complete set of noiseless drivers cannot fully explain the target.

The framework in McGill (1954) uses information theoretic measures to fill these roles. (See again Fig. 3.1, right side.) Shannon entropy of the target is used instead of variance. The drivers are still the engine, but they extract mutual information between them and the target as explanation, leaving the conditional Shannon entropy of the target as wasted explanation.

3.1.2 Relative Strength and the Heat Pump

Using information theory for continuous variables instead of discrete variables requires a new method for measuring relative strength. The reason is that mutual information between continuous variables is not upper bounded while Shannon entropy remains finite. This means that Shannon entropy specifically no longer upper bounds mutual information, and so mutual information cannot be interpreted as Shannon entropy explained. By reversing the arrows in the heat engine analogy, mutual information again has an upper bound, allowing for calculating the relative strength of explanation.

This is exactly how the heat pump differs from the heat engine. (See Fig. 3.2, left side.) A heat pump extracts energy from the cold reservoir, adds the work it does to that, and puts the total into the hot reservoir. *Certainty*, introduced in van Leeuwen et al. (2021), follows a similar concept. (See Fig. 3.2, right side.) The observed target now acts as the cold reservoir, emitting some *self-certainty*. To this self-certainty, the drivers add the mutual information between them and the target as a *certainty gain*. Their sum, the *total certainty*, is equivalent to the heat added to the hot reservoir. The more certain target is then equivalent to the hot reservoir.

Admittedly, this particular analogy with the heat pump is not as exact as for the heat engine. The issue comes from how heat pumps not only add work to the hot reservoir, but also move heat from the cold reservoir to the hot reservoir. The difficulty to add heat to the hot reservoir appears as extracting less heat from the cold reservoir for the same amount of work. In the certainty framework, however, self-certainty is independent of the certainty gain from the drivers, and the difficulty to explain the target appears as adding less certainty for the same amount of self-certainty.

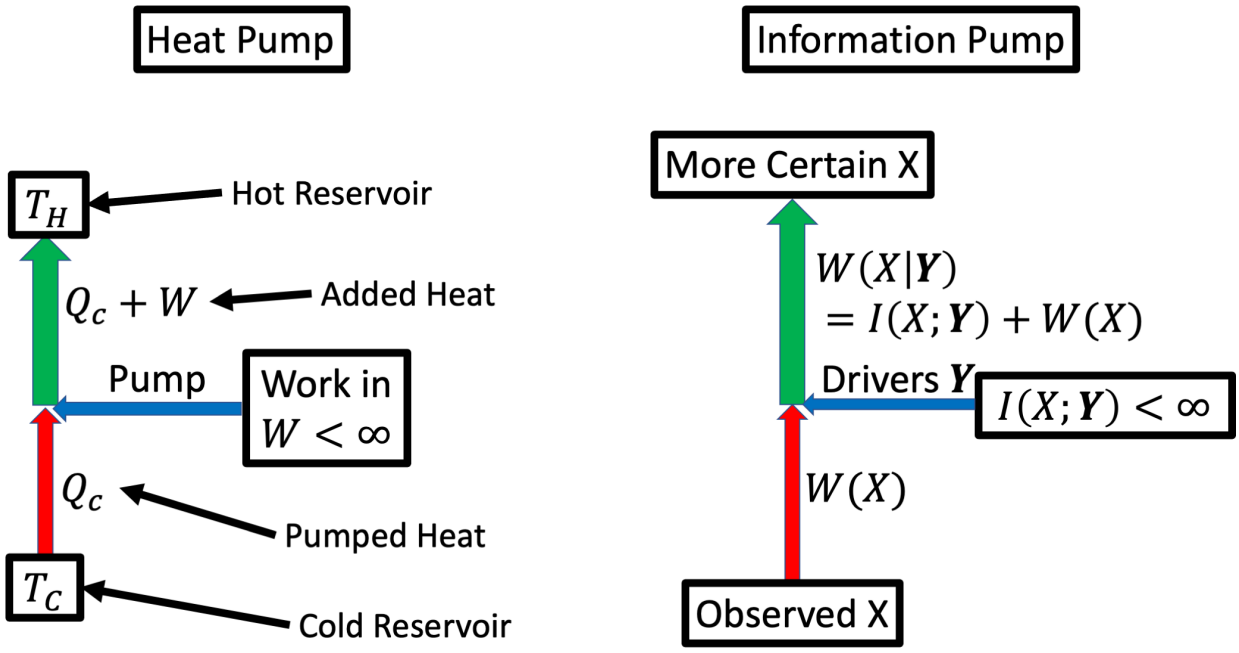

FIG. 3.2. Comparison of heat pump (left) and information pump (right). The heat pump moves heat energy, Q_C, from a cold reservoir at temperature T_C to a hot reservoir at temperature T_H. The work, W, put into the system from the heat pump may be arbitrarily large, so it has no finite upper bound. The total heat that the pump puts into the hot reservoir is $Q_C + W$. With the information pump, there is some background certainty, $W(X)$, that comes from the observation. The drivers, $\mathbf{Y}$, add an arbitrarily large amount of certainty as mutual information, $I(X;\mathbf{Y})$. Their sum, $W(X)+I(X;\mathbf{Y})$, is the full, or conditional, certainty about the target, $W(X|\mathbf{Y})$.

The analogy can be corrected, however, when we look at the relative contributions. As heat pumps create a temperature difference, the drivers create a certainty difference in the target. Just as a heat pump does little work if the temperature difference is small, the drivers poorly explain the target if the certainty difference is small. When normalized by the certainty of the more certain target, the relative contribution of the drivers is small. However, when the contribution from the drivers is large compared to the self certainty, the relative contribution of the self certainty is small. In this way, the self certainty does factor into how an information pump works, and so the analogy to the heat pump stands.

3.2 Coupled Influence, *M*links, and Total Influence

All the decompositions of certainty gain from van Leeuwen et al. (2021) were shown in the previous chapter. Since van Leeuwen et al. (2021), I proved that coupled influences are actually interaction informations. (Proof in Appendix A.5.) Interaction information was introduced in McGill (1954) to show

how drivers interact to yield more or less predictability of a target. The trivariate interaction information of X, Y, Z can be expressed as

$$II(X,Y,Z) = I(X;Y) - I(X;Y|Z) = I(X;Z) - I(X;Z|Y) = I(Y;Z) - I(Y;Z|X), \tag{3.1}$$

where II is the interaction information. The symmetry of the decomposition is clear. If X was the target of Y and Z, we could express this instead as

$$\begin{aligned} {}^{X}_{YZ}M_{\mathbf{Y}} &= I(X;Y,Z|\mathbf{Y}\setminus\{Y,Z\}) - I(X;Y|Z,\mathbf{Y}\setminus\{Y,Z\}) - I(X;Z|Y,\mathbf{Y}\setminus\{Y,Z\}) && (3.2)\\ &= I(X;Z|\mathbf{Y}\setminus\{Y,Z\}) - I(X;Z|Y,\mathbf{Y}\setminus\{Y,Z\}) && (3.3)\\ &= II(X,Y,Z|\mathbf{Y}\setminus\{Y,Z\}), && (3.4) \end{aligned}$$

where I used the relation $I(X;Y,Z|\mathbf{Y}\setminus\{Y,Z\}) = I(X;Y|Z,\mathbf{Y}\setminus\{Y,Z\}) + I(X;Z|\mathbf{Y}\setminus\{Y,Z\})$. The result for two drivers was already shown in van Leeuwen et al. (2021). In general, the coupled influence of $J \subseteq \mathbf{Y}$ on X is the interaction information of X and every driver in J, written

$${}^{X}_{J}M_{\mathbf{Y}} = II(X,J|\mathbf{Y}\setminus J), \tag{3.5}$$

the proof of which is in Appendix A.5.

3.2.1 Interpreting Coupled Influences from Two Variables

The physical interpretation of coupled influences in van Leeuwen et al. (2021) was only evidenced by examples, and there is no strong link to underlying forms of equations. Because measures of information are conserved under bijective transformations of a single variable, a link to equations might not exist. For example, for drivers Y and Z and target X, the information measures for $x = yz$ are equal to those for $\log x = \log y + \log z$ when the noise is small relative to the influences of Y and Z. Thus, to make an interpretation for coupled influences, I will simply use an additive form $x = f(y) + g(z) + \eta$, where η represents noise and f and g are some functions of Y and Z, respectively, noting that many forms of equations are represented by this one. For the rest of this section, I will use Y and Z as drivers of target X.

Assuming drivers Y and Z are independent, the graphical representation of $x = f(y) + g(z) + \eta$ is $Y \to X \leftarrow Z$, a standard collider. By the equation and the graph, we can and should conclude the effects of Y and Z are separate. In terms of mutual information, we observe $I(Y;Z) = 0$ because Y and

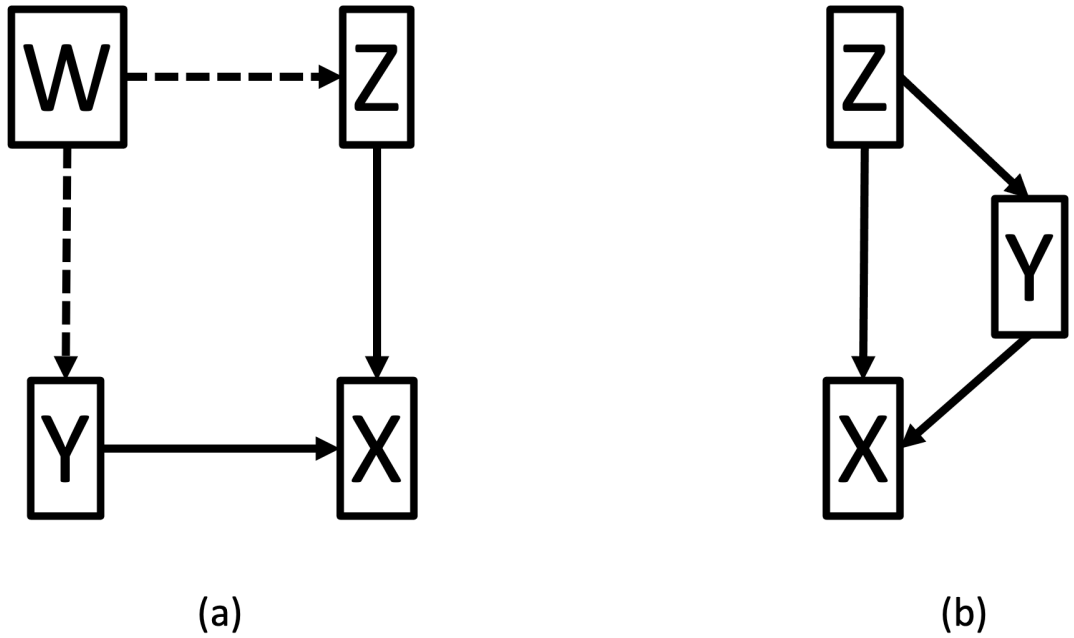

FIG. 3.3. Standard graphs where Y and Z are dependent, which would yield less negative coupled influence from them. On the left, they have a common cause, where the dashed lines imply W is not included in the study. On the right, Y mediates the indirect effect of Z.

Z are independent. Meanwhile, since Pearl (2000) shows conditioning on a target makes its parents dependent, $I(Y;Z|X)>0$. This yields $I(Y;Z)<I(Y;Z|X)$, so the coupled influence of Y and Z

$$ {}^{X}_{YZ}M_{\mathbf{Y}} = II(X,Y,Z) = I(Y;Z) - I(Y;Z|X) < 0 \tag{3.6}$$

is negative.

We should also expect this from the equation. When Z, for example, is not included in the analysis, it looks like noise. That is, $x = f(y)+(g(z)+\eta)$, in which the variability of Z is clearly part of the uncertainty of the effect of Y. Conditioning on Z means that we instead study the effect of Y on X while keeping Z constant, repeating this for each value of Z. This removes the variability of Z from the equation for X and thereby decreases the uncertainty of the effect of Y. This decrease in the uncertainty of the effect of Y appears as an increase in its mutual information with X, so $I(X;Y)<I(X;Y|Z)$. This argument also holds when calculating the mutual information between Z and X and conditioning on Y instead. Thus, when Y and Z are independent, the system $x = f(y)+g(z)+\eta$ will yield negative coupled influence, which shows that the effects of Y and Z are separable from each other.

To understand when coupled influences will be less negative and possibly positive, we again need the interaction information decomposition. That is,

$$ {}^{X}_{YZ}M_{\mathbf{Y}} = II(X,Y,Z) = I(Y;Z) - I(Y;Z|X), \tag{3.7} $$

where it is obvious that ${}^{X}_{YZ}M_{\mathbf{Y}} \leq I(Y;Z)$. When Y and Z are dependent, so $I(Y,Z) > 0$, their coupled influence is less negative and possibly positive. They may be dependent in many ways. Pure chains, like $Y \to Z \to X$ or $Z \to Y \to X$, are not allowed, however, as we would see the middle process d-separates its parent from X, which appears as a 1link or direct influence that is zero for the root process. Instead, Y and Z may have a common ancestor W, yielding the graph $Y \leftarrow W \to Z$ and $Y \to X \leftarrow Z$. (See Fig. 3.3 (a).) This means that $y = h_1(w) + \eta_y$ and $z = h_2(w) + \eta_z$, so

$$ x = f(h_1(w) + \eta_y) + g(h_2(w) + \eta_z) + \eta. \tag{3.8} $$

Conditioning on Z will again remove its variability. But, since the variability of both Y and Z contains the variability of W, conditioning on Z will decrease the variability of Y, thereby decreasing $I(X;Y|Z)$. By a similar argument, without conditioning on Z, the variability that Z has in common with Y is attributed instead to Y, thereby increasing $I(X;Y)$. The amount of change in $I(X;Y|Z)$ and $I(X;Y)$ is dependent on the strength of both $W \to Y$ and $W \to Z$. If one or both connections are weak, then Y and Z will be weakly dependent, so their coupled influence will likely remain negative. But, if both connections are strong, then Y and Z will be strongly dependent, likely yielding $I(X;Y) > I(X;Y|Z)$, or positive coupled influence. Since W is excluded from the study because $\{Y,Z\}$ d-separates it from X, the effects of Y and Z are inseparable.

Another possibility is that, instead of Y and Z being dependent because of a common ancestor, suppose Z has a direct effect on X as well as a mediated effect through Y. The graph would be $Z \to X$ and $Z \to Y \to X$. (See Fig. 3.3 (b).) If $Z \to Y$ is weak, then Y and Z are weakly dependent, so $I(X;Y|Z) > I(X;Y)$. And, if Z completely determines Y, then $I(X;Y|Z) = 0$, meaning Y would be removed from the study. But, if $Z \to Y$ is strong yet not deterministic, then their strong dependence will yield $I(X;Y) > I(X;Y|Z)$, even if $Y \to X$ is strong. Again, the effects of Y and Z appear inseparable, but there is no other process to condition on to separate their influences.

This leads to an important discussion regarding causal sufficiency. Causal sufficiency, again, requires that the common cause of two or more variables must also enter the analysis. Yet, the above argument states that W, the common cause of Y and Z, should be removed from the analysis because

it is d-separated from X. This does not go against causal sufficiency, however, because the effects of the W are still captured by Y and Z. If the direct influence of W was nonzero, then the certainty framework requires that an arrow connects W directly to X. If the direct influence is instead zero, including it in the study jeopardizes the analysis because the coupled influences W has with Y and Z would give W causal significance. In this way, causal sufficiency can be restored by calculating the coupled influence of Y and Z.

The current interpretation is clear for the two-driver case. Negative coupled influence indicates that the effects of the drivers are separable. Inversely, positive coupled influence indicates that the effects are inseparable, mostly because the drivers themselves are dependent. The more negative or positive the coupled influence is, the more separable or inseparable the effects are, respectively.

For now, this interpretation is carried over to coupled influence terms of more than two drivers. As of yet, however, exactly how the value reflects the structure of the causal web is unknown. Possible interpretations are explored in the discussion portion of the thesis (Section 7.1.2).

3.2.2 Mlinks and Total Influence

Coupled influences are combined to form both $m$links and total influence for individual variables (van Leeuwen et al. 2021). An $m$link shows how inseparable an individual driver is at a given level of coupling. Because an $m$link involves multiple coupled influences, it can indicate how active a driver is at a certain level of coupling. If a driver's largest influence comes from its 1link, then it is mostly active alone, and its effect is generally separable. If, however, the driver's largest $m$link is its 2link, then it is most active when coupled with exactly one other driver, and its effect is generally inseparable.

As stated in van Leeuwen et al. (2021), the total influence of a driver is the sum of its $m$links. Since van Leeuwen et al. (2021), I proved that a driver's total influence is nonnegative. (Proof in Appendix A.4.) This means that, even though coupled influences and mlinks may be negative, the total influences constitute a nonnegative decomposition of mutual information. A nonnegative decomposition is important because it implies that information attributed to one driver is not also attributed to another driver.

3.3 Choice of reference

In the certainty framework, the choice of reference density is the only explicitly subjective component. We have yet to determine clear recommendations for users of the framework. In fact, many

possible densities are explored in van Leeuwen et al. (2021). There will probably always be some remaining subjectivity no matter what we do.

Of special interest to the application to tropical cyclone rapid intensification (TCRI), in Chapter 5, is a reference density designed specifically for high noise targets. Since noise in any process diminishes the total mutual information, the presence of high noise may artificially make relevant drivers appear irrelevant. This problem is made worse when the self-certainty is large. Because the reference density to a large extent determines the size of the self-certainty, the reference density should be different when the target is believed to have high noise and the drivers are believed to be relevant. This is intended simply as a correction so that the results more truly evidence the relevance of the drivers.

Using the TCRI data, I experimented with many different reference densities. Gaussian references with varying spread were tried, but they all yielded a very low self-certainty. For example, using a Gaussian reference with the same variance as the target yielded $W(X) = 0.0113$. Since self-certainty is a KL divergence, a measure of departure of the target's density from the reference density, I concluded the target was too Gaussian-like to use a Gaussian reference, which is evidenced in the examples in van Leeuwen et al. (2021). Ultimately, the Lorentz-Cauchy distribution seemed to be preferable given that it is not a Gaussian and that it is featureless. After experimenting with the width-parameter expression, the expression resulting in the minimum self-certainty was

$$\gamma = \sqrt{\frac{e}{2\pi}}\,\sigma_X, \tag{3.9}$$

where σ_X is the standard deviation of the target. This is simply double the original expression.

CHAPTER 4

Implementing the Framework

Implementing the framework in code is a feat of combinatorics. The number of (conditional) mutual information terms is exponential with the number of drivers. To be exact, if the driver set, $\mathbf{Y}$, has n drivers, the number of information terms is $2^n - 1$. When including the self-certainty of the target, $W(X)$, the total number of information theoretic calculations is 2^n. The $2^n - 1$ information calculations are recombined into $2^n - 1$ coupled influence terms. These coupled influence terms are recombined according to the amount of linkage into n^2 $m$link terms, which themselves are summed over each driver to make n total influence terms. Everything is normalized by the conditional certainty, $W(X|\mathbf{Y})$, which is calculated as the sum of $W(X)$ and $I(X;\mathbf{Y})$.

The entire framework is implemented in C++ using double precision arithmetic. The main reason is that I can use C++ fluently, allowing me to quickly write the implementation. Beyond this, object orientation allows for easy transition from concepts to code, and C++ offers potentially the most computational efficiency for object orientation. Also, compiled code is automatically faster than interpreted code, and most C++ compilers additionally offer a lot of optimization. A Python wrapper as well as a FORTRAN implementation are future projects.

The rest of this chapter is organized as follows. First, estimating mutual information is discussed in Section 4.1. Combining the mutual information to calculate coupled influences is discussed in Section 4.2, while calculating $m$links and total influences is discussed in Section 4.3. Finally, calculating self-certainty is discussed in Section 4.4.

4.1 Calculating Mutual Information

Calculating mutual information for the framework entails two main parts. The first is determining which drivers are active and which are conditioned on. To do this, the drivers are first given a set order. (One such ordering is detailed at the end of Section 4.1.2.) Then, to efficiently populate the active set, the implementation loops through the integers between 1 and $2^n - 1$ inclusively, where n is the number of drivers. In the computer, bits are either ON or OFF. The integer 1 is represented by only the smallest bit being ON, in which case only the first driver in the driver set is active and all other drivers are conditioned on. The integer 2 is represented instead by only the second smallest bit being ON, so only the second driver is active and all other drivers are conditioned on. The integer 3 has both the

smallest and second smallest bits ON, so the first two drivers are active and the rest are conditioned on. The integer $2^n - 1$ has the first n bits ON, so all drivers are active and the condition set is empty.

The other main part is implementing the mutual information estimators themselves. The implementation uses k-nearest neighbors (kNN) estimators as developed by Kraskov et al. (2004) (mutual information) and Vejmelka and Paluš (2008) (conditional variant). Specifically, the mutual information estimator follows the first algorithm in Kraskov et al. (2004),

$$I(X;Y) \approx \psi(k) + \psi(N) - <\psi(n_x+1) + \psi(n_y+1)>, \tag{4.1}$$

where ψ is the digamma function, k is the nearest-neighbors parameter, N is the length of the time series, and n_x is the number of neighbors in X within the kNN distance of the set $\{X, Y\}$, and likewise for n_y and Y. The kNN distance, ϵ, from each point, (x_i, y_i), in the time series is the the k^{th} smallest distance to any other point (x_j, y_j). We chose the Chebychev, or maximum, norm to calculate distance. Finally, the values of $\psi(n_x+1)+\psi(n_y+1)$ are averaged over all points (x_i, y_i), yielding the term $<\psi(n_x+1)+\psi(n_y+1)>$.

The algorithm from Vejmelka and Paluš (2008) closely parallels this. It is,

$$I(X;Y|Z) \approx \psi(k) - <\psi(n_{xz}+1) + \psi(n_{yz}+1) - \psi(n_z+1)>, \tag{4.2}$$

where n_{xz} is the number of neighbors in the combined set $\{X, Z\}$ within the kNN distance of the set $\{X, Y, Z\}$, and likewise for n_{yz} and $\{Y, Z\}$.

Choosing the mutual information estimator that parallels the conditional estimator helps reduce error. Further reducing error can happen by 1) transforming the data, 2) addressing large active sets, and 3) choosing the optimal value of k. (The transformation for self-certainty estimation is different, which is detailed is Section 4.4.) These separate parts are detailed below.

4.1.1 Transforming the Data

While kNN estimators do not assume any underlying distribution for the drivers and target, they are sensitive to highly peaked probability densities. Furthermore, since the Chebychev distance is the maximum value in a vector, variables with larger scales are over represented while those with smaller scales are under represented. Since the true value of mutual information is insensitive to single-variable bijective transformations, applying any such transformation to individual drivers is allowed to fix the

above issues. Experimentation suggests the estimators yield their maximum values when each variable is marginally Gaussian. Thus, for the information estimation, each variable is transformed to a truncated standard Gaussian.

The transformation starts with replacing a value in a time series with its overall rank in value in the time series. Repeated values have the same rank, and the next highest value has a rank that is greater by the number of repeated values. These ranks are then divided by the length of the time series, making it approximately uniformly distributed over the interval $[0,1]$. From here, the inverse cumulative distribution function of the standard Gaussian is applied to the time series, using the approximation to the inverse error function in Winitzki (2008), such that the tails of the distribution are equally truncated. Many amounts of truncation were tried. The estimators were maximized by removing 4.25% of the full distribution from each tail such that the central 91.50% of the Gaussian distribution is occupied.

4.1.2 Overcoming the Curse of Dimensionality

Dimensionality refers to the number of time series comprising the target and active set. In the estimators, X, Y, and Z may comprise multiple time series, making implementation very simple at first. But, with even a few active drivers, the estimators became highly inaccurate, a phenomenon known as the *curse of dimensionality*. Kraskov et al. (2004) recognized this and offered an alternate estimator in such cases, but it is less consistent conceptually with the conditional mutual information estimator. Furthermore, only one of the information terms is a regular mutual information, and the other $2^d - 2$ terms are conditional mutual informations. Therefore, the implementation follows the original algorithm.

To alleviate the issue, both estimators are a sum of mutual informations with only one target time series and one active time series at a time. The formula for two active time series Y_1, Y_2 is

$$I(X; Y_1, Y_2|Z) = I(X; Y_1|Y_2, Z) + I(X; Y_2|Z), \tag{4.3}$$

and for m active time series $Y_1, \ldots, Y_m$,

$$I(X; Y_1, \ldots, Y_m|Z) = I(X; Y_1|Y_2, \ldots, Y_m, Z) + \cdots + I(X; Y_m|Z), \tag{4.4}$$

and similarly for multiple target time series. Any information estimation with more than one active or target time series becomes a summation of information estimations between one target and one driver time series at a time.

Computing the mutual information using this decomposition may over represent strong drivers and under represent weak drivers in the estimation. The information terms calculated toward the beginning of this decomposition will have more time series in the condition set, while those toward the end of the decomposition will have less. If a strong driver is active with less processes in the condition set, the resulting value may be greater than it should be. Inversely, if a weak driver is active with more processes in the condition set, the resulting value may be less than it should be. The order of decomposition often changes the final value, showing that the systematic biases do not necessarily cancel.

To address this, the drivers are reordered by their 1links, where the driver with the greatest 1link is first, and the driver with the least 1link is last. This way, when an estimation is decomposed, the calculations in which the strong drivers are active will always have more processes in the condition set than calculations in which the weak drivers are active. This reduces both the over representation of the strong drivers and the under representation of the weak drivers. Thus, the overall systematic bias should be reduced.

4.1.3 Determining the Number of Neighbors Parameter

Part of the art of using kNN estimators is figuring out what number of neighbors, k, to use for a given dataset. The user must tell the program which k to use. The implementation does not automatically choose the value of k, though the default is $k = 3$.

But, the implementation does have functionality to suggest a k based on two criteria. The first criterion was that k is less than 1% of the time series length, based on Kraskov et al. (2004) and Vejmelka and Paluš (2008). The second criterion was that the estimation of the total certainty gain had the least dependency on the order of decomposition.

This dependency was evaluated in the following way. For each k, the kNN estimators above were used. This means that the drivers were first reordered based on their 1links for each value of k. Second, $I(X;\mathbf{Y})$ was estimated, where X is the target and $\mathbf{Y}$ is the set of all drivers. Third, for each driver $Y \in \mathbf{Y}$, both $I(X;Y|\mathbf{Y}\setminus\{Y\})$ and $I(X;\mathbf{Y}\setminus\{Y\})$ were estimated, the sum of which should be $I(X;\mathbf{Y})$. Finally, the code calculated the relative deviation of each $I(X;Y|\mathbf{Y}\setminus\{Y\}) + I(X;\mathbf{Y}\setminus\{Y\})$ from $I(X;\mathbf{Y})$, then squared and summed the relative deviations. The relative deviation for the first driver is always 0. This square relative deviation evidences how strongly an estimate depends on the order of decomposition, with greater values indicating greater dependency.

When using this functionality of the implementation, the program prints the k that is less than 1% of the time series length and has the least square relative deviation as described above. The user must then run the framework and explicitly input the desired value of k.

4.2 Calculating the Coupled Influences

To calculate coupled influences from the mutual information terms, I prove in section A.2 that the coupled influence of a set of drivers $J \subseteq \mathbf{Y}$ is

$$ {}^{X}_{J}M_{\mathbf{Y}} = \sum_{j \subseteq J}^{|j|>0} (-1)^{|J|-|j|} I(X; j | \mathbf{Y} \setminus j). \tag{4.5} $$

In words, the conditional mutual information between X and every nonempty subset $j \subseteq J$ conditioned on all drivers not in j is calculated. Then, the mutual information is added if $|j|$ and $|J|$ are both even or both odd, or it is subtracted otherwise. Writing the influences as a summation of the mutual informations removes the dependencies imposed by the framework's recursive definition of influence.

Instead of finding the subsets of a set, J, to calculate the coupled influence of J, the implementation finds the supersets of J to iteratively adjust their coupled influences. A superset of a set is the reverse relationship of a subset of a set. Just as a subset of J is fully contained in J, J is fully contained by its supersets. The term $I(X; J|\mathbf{Y} \setminus J)$ will not appear in the coupled influence of any proper subset of J, but it will appear in the coupled influence of J and all supersets of J. The implementation loops through the sets as previously described. The conditional mutual information for each set is first added to the set's coupled influence. Then, the OFF bits are turned ON and OFF in a recursive scheme such that the supersets are updated in increasing order of their integer representation. When the size of the set and a superset are both even or both odd, the set's mutual information is added to the superset's coupled influence, and subtracted otherwise. While other implementations for combining the information estimates into coupled influences may be faster, this was the first one that made sense to me, and I did not explore others.

4.3 Calculating Mlinks and Total Influences

The $m$link and total influences are calculated from the coupled influence terms as shown in section 3.2.2. Alternate calculations for $m$link and total influences using mutual information terms as basic units are developed in sections A.3 and A.4, respectively. While these equations might prove useful

to avoid round off errors from adding and subtracting terms repeatedly, double precision arithmetic makes such errors negligible. Thus, the implementation no longer uses them.

These results do merit discussion, however, as the result for total influence is important for the framework. Sometimes the implementation yields negative total influences, which was not expected. The result for total influences in Section A.4, which uses the result for $m$links in Section A.3, shows that the true value of total influence can never be negative. Thus, a calculated negative total influence is due to numerical error.

4.4 Calculating Self-certainty

The self-certainty is a simple KL divergence, written as

$$W(X) = D_{KL}(p_X || q_X) = \int p_X(x) \log \frac{p_X(x)}{q_X(x)} \, dx = -E_{p_X}(\log q_X) - H(X), \tag{4.6}$$

where p_X is the distribution of X, q_X is the reference distribution, and the expectation of $\log q_X$ is called the cross entropy while $H(X)$ is the familiar Shannon entropy. To avoid implementing a KL divergence estimator, I used the fact that KL divergence, like mutual information, is conserved under single-variable bijective transformations. Any such transformation applied to the target's time series is a transformation applied to the reference density. Thus, the implementation applies a transformation to the target such that the cross entropy is zero in the transformed space. Because the implementation lacks a KL divergence estimator, each reference density must have a separate transformation implemented. Currently, the supported reference densities are the uniform distribution, both the original and wide Lorentz-Cauchy (LC) distributions, and the Gaussian distribution.

Applying the cumulative density function of the reference density makes the cross entropy zero. The reference density in the transformed space is then uniform on $[0,1]$. As the density is 1 everywhere, its logarithm is 0, and so its cross entropy is zero. This leaves

$$W(X) = -H(X_{trans}), \tag{4.7}$$

where X_{trans} is the transformed target. This method is used for the uniform distribution and the LC distributions.

The transformation for the uniform distribution scales and shifts the data. The minimum observed value is mapped to zero. The maximum observed value is mapped to one.

For the LC distributions, the transformation is

$$X_{trans} = \frac{1}{\pi} \arctan\left(\frac{X - x_0}{\gamma}\right) + \frac{1}{2}, \tag{4.8}$$

where x_0 is the sample mean of X, and γ is the spread parameter. The formula for γ is either the original or the wide variant (from Section 3.3), depending on which the user says to use.

The transformation for the Gaussian reference shifts and scales the data. With Gaussian references, the cross entropy is the Shannon entropy of the reference Gaussian, $1/2 \log(2\pi e\sigma^2)$, where the implementation uses the variance of the target for σ^2. Thus, transforming the target to have variance $\sigma^2_{trans} = (2\pi e)^{-1}$ makes the cross entropy zero. Because the formula for the cross entropy is known, the implementation does not need to transform the data. The transformation in this case is only for the sake of consistency.

CHAPTER 5

A Case Study of the Rapid Intensification of Hurricane Patricia (2015)

Tropical cyclone (TC) rapid intensification (RI) is defined as an incease of at least 30 knots (about 35 mph or 55 kph) within 24 hours in sustained maximum tangential wind. Predicting whether or not a hurricane undergoes RI is one part of the problem, though the vast majority do. Assuming the conditions allow for RI, the questions then become when and by how much the TC will rapidly intensify. While TC dynamics and thermodynamics are governed by processes at all levels of the troposphere, the present analysis focuses on processes in outflow and boundary layers.

The importance of the outflow layer is well recognized (Gray 1968; McBride and Zehr 1981; Merrill 1988a,b; Davis and Bosart 2004; McTaggart-Cowan et al. 2008; Kimberlain et al. 2016). While low vertical shear over the center of the storm is preferable to prevent ventilation and loss of heat, the ability for air to exit the system is critical (Gray 1968; McBride and Zehr 1981). Furthermore, synoptic forcing, such as interactions with upper-level troughs and ridges, is responsible for the formation and intensification of many TCs (Davis and Bosart 2004; McTaggart-Cowan et al. 2008). In fact, of the several modes of cyclogensis that McTaggart-Cowan et al. (2008) identify, many involve synoptic forcing of the outflow layer. The local vertical shear gradients may create large cyclonic vorticity fields to spinup the TC (McBride and Zehr 1981; McTaggart-Cowan et al. 2008; Kimberlain et al. 2016). Merrill (1988b) also shows that the "azimuthal mean radial outflow tends to be stronger for intensifying hurricanes... [than for nonintensifying hurricanes]," implying that intensification entails a net removal of air from the center of the storm. Many more complex dynamics certainly occur in the outflow layer, but the importance of the outflow itself cannot be ignored.

The boundary layer also exhibits dynamics and thermodynamics peculiar for TC intensification. Early studies even suggested that TCs could form solely based on boundary layer dynamics and thermodynamics (Gray 1968). Though this was later debunked, the role of the boundary layer remains important (Ooyama 1969; McBride and Zehr 1981). Specifically, for both intensification and maintenance, TCs need water vapor supplied to the core of the storm by warm oceans, which is well known since Gray (1968). The friction of the circulation against the ocean induces convergence in the boundary layer, which carries the water vapor toward the center. The core lies within the radius of maximum wind (RMW), pointing to the importance of the radial wind at RMW in conjunction with water vapor.

The thermodynamics of TC development are just as rich. For instance, a mature TC can be viewed as a Carnot engine, linking TC intensity to the temperature difference between the outflow and the sea surface temperature (Emanuel 1986, 1991). Later, Emanuel and Rotunno (2011) and Emanuel (2012) related the outflow temperature to intensification. In an idealized model, Hu and Wu (2020) showed that high equivalent potential temperature (θ_e) between RMW and 3 RMW leads to intensification, while high θ_e beyond 3 RMW instead leads to stronger rainbands. Furthermore, Vigh and Schubert (2009) show that diabatic heating is more efficient at spinning up circulation when applied inside than outside RMW, stressing the importance of moving high θ_e air into the core of the storm. The thermodynamics are much more complex, but this is what is most relevant to this study.

Kimberlain et al. (2016) show that Hurricane Patricia (2015), the most rapidly intensifying and rapidly weakening TC on record for both the northeast Pacific and north Atlantic basins, was highly favored by synoptic forcing. Even though the initial low was slow to develop, it combined with a tropical wave and later intensified from convergence of cyclonic vorticity generated by a Tehuantepec gap wind event. On top of this, the rising branch of an eastward propagating Madden-Julian Oscillation possibly increased the deep convection of the system prior to RI. More synoptic forcing and steering occurred to turn this depression into a storm, which was also slow to develop. Then, Patricia finally passed through an environment with a patch of anomalously warm water, high humidity, and low vertical shear, where it underwent RI. In many ways, Patricia was a very ideal storm to study RI.

5.1 Methodology

5.1.1 Data

The data comes from a 60-member Weather and Research Forecasting (WRF) ensemble forecast which simulates Hurricane Patricia during its RI, from 21:00 UTC on October 21 to 00:00 UTC on October 23. The horizontal grid resolution was 1km, while there were 42 vertical levels using the eta vertical coordinate. The ensemble was initialized using data from typical observations as well as data from the Office of Naval Research Tropical Cyclone Intensification 2015 field campaign. Simulation data was output hourly, yielding 28 time steps per member. See Tao et al. (2020) for more information on these simulations.

For each member, two time steps were removed. The initial and final time steps were removed from the drivers' time series. The initial time step seemed to experience shock from the data assimilation.

The final time step was removed because the drivers' time series lagged the target's time series by an hour. Accordingly, the target's first two time steps were removed. In total, the time series had 26 time steps per member.

The data was then azimuthally averaged. To obtain the azimuthal averages, the WRF output was first interpolated to fixed vertical coordinates. Variables derived from potential temperature, pressure, and/or moisture were calculated after the interpolation but before the azimuthal averaging. A 21km by 21km moving average was applied to the surface perturbation pressure, and the location of the minimum was used as the center of the storm. Using the perturbation pressure helped account for any topographical effects, as higher altitudes are naturally at a lower pressure. From there, the processes were azimuthally averaged in 1km bins from 1km to 120km.

5.1.2 Selected Processes

Because the time series length was relatively short (26 time steps per member $\times$ 60 members = 1560 time steps), we limited the study to considering only four drivers for the sake of accurate information estimations. Another criterion was proximity to the core of the storm. Thus, the processes were 1) the upper-level radial wind, u_u, 2) the cross-RMW boundary layer (BL) radial wind, u_l, 3) the BL equivalent potential temperature at RMW, θ_e, and 4) the temperature difference between the surface and outflow layers, ΔT. The reasoning behind choosing these drivers is described first. Then, generating the representative time series is described. Defining RMW is detailed in the next section (5.1.3).

Based on the importance of the outflow, u_u was selected as the proxy variable. We experimented with using the actual outflow, which is the radial wind weighted by mass, instead of the radial wind. The results are not shown, but u_u was a better predictor than mass outflow. There are two reasons we think are plausible. The first is that, since the maximum value at each time was chosen, the value for u_u tended to be closer to the eye, while outflow peaks further from the center because the mass weighting increases linearly with radius. Proximity to the eye of the storm is crucial when the time lag is only one hour. The other reason is that, as the storm intensifies, the central pressure drops. Because mass and pressure are positively related, the decrease in pressure may offset changes in radial wind, meaning that the net outflow of air may not change much. Directly using radial wind avoids this issue.

To gauge whether or not warm moist air is able to reach the core of a TC, u_l was used. There were some times when u_l was directed out of the core. While air may have entered the core at other places at that time, the azimuthal average indicates that air was ultimately leaving the core.

Using θ_e was based on Hu and Wu (2020). Their study showed that θ_e near RMW in the boundary layer has a high partial correlation with the rate of intensification. Given the strong linear dependence between θ_e and intensification, θ_e was included as a potential driver.

Using ΔT to study intensification was new, as far as we could tell. Again, a TC is a Carnot engine, so ΔT is linked to intensity once the storm matures and is in a steady state (Emanuel 1986). But, the TC also physically links the outflow and surface layers. As with any physical link between two bodies of different temperatures, the rate of heat exchange is proportional to the temperature difference. Since the TC does the heat exchange, perhaps ΔT drives intensification until a steady state is reached. Any statement beyond this would be mere speculation, but this was why ΔT was a potential driver.

The time series for these processes were generated in the following ways. First, the value of u_u was the maximum value of a moving 3km wide by 1km tall average of the radial component u over the outflow layer. Second, the value of u_l was a 3km wide by 1km tall average at the surface centered on the RMW. Third, the value of θ_e was a 3km wide by 1km tall average at the surface centered at RMW. Finally, the value of ΔT was the difference between the average BL temperature and the average outflow layer temperature. The region for averaging the BL temperature was the bottom 1km, from RMW to RMW+10km, including only where $u < 0.95u_l$. The region for averaging the outflow layer temperature was wherever $u > 0.95u_u$.

5.1.3 Choosing the target and the reference

The hourly change in maximum tangential wind, Δv_{max}, was the target variable. The hourly intensification is a natural proxy for instantaneous intensification, and the assumed small memory in the process means its past can be neglected as a potential driver. The maximum tangential windspeed itself, v_{max}, was defined by first a 3km wide moving average of the tangential windspeed over the 600m to 1km heights, from which the maximum value was used. The radius where v_{max} occurred was used as the radius of maximum wind (RMW).

Because Δv_{max} was believed to have high noise, the wide Lorentz-Cauchy distribution, as described in section 3.3, was used for the reference density. The belief of high noise is because of the following. Assuming that the errors of the v_{max} time series are independent, the error variance in Δv_{max} time series is doubled. Meanwhile, the maximum value of the ratio between Δv_{max} and v_{max} is about $1/7$, meaning the noise-to-signal ratio is roughly 14 times larger for Δv_{max} than for v_{max}. This and other experiments not shown here suggest that Δv_{max} had high noise.

5.1.4 Calculating the information quantities

The implementation as described in Chapter 4 was used. To determine the best number of neighbors, k, the criteria in Section 4.1.3 were followed. This yielded $k = 7$ for the number of neighbors.

5.2 Results

The framework yielded $W(\Delta v_{max}) = 0.1575$ and $I(\Delta v_{max}; \mathbf{Y}) = 0.1316$, where $\mathbf{Y} = \{u_u, u_l, \theta_e, \Delta T\}$. This yielded a conditional certainty $W(\Delta v_{max}|\mathbf{Y}) = 0.2891$, 45.5% of which came from the drivers. The process u_u accounts for 10.9% of $W(\Delta v_{max}|\mathbf{Y})$, u_l for 7.1%, θ_e for 16.3%, and ΔT for 11.2%. Overall, the drivers left 54.5% unexplained. These totals are shown in figure 5.1.

The causal web (Fig. 5.2) shows the direct and coupled influences normalized by $I(\Delta v_{max}; \mathbf{Y})$. Normalizing by this instead of the conditional certainty helps highlight the origin of the certainty gain itself. The largest direct influence is from u_u while the smallest is from u_l. There are six two-driver coupled influences. Coupled influences involving θ_e were all positive. Specifically, u_u or ΔT coupled with θ_e yielded large influences, 18.0% and 16.1% respectively. The coupled influence of u_l and ΔT was moderately negative, -8.0%. The other two-driver coupled influences were relatively negligible.

Of the four three-driver coupled influences, two are worth noting for the size of contribution. Coupling $\{u_u, u_l, \theta_e\}$ yielded -9.4%, while $\{u_l, \theta_e, \Delta T\}$ yielded 29.9%. The coupled influence of $\{u_u, \theta_e, \Delta T\}$ was -1.9%, a small but negative term. This means two of the three three-driver couplings involving both u_u and ΔT yielded negative values, while the third was weakly positive.

Finally, the coupled influence of all drivers was -17.1%, a large negative value. This is how the certainty gain, i.e. $I(\Delta v_{max}; \mathbf{Y})$, decomposed into coupled influences.

The $m$links and total influences are shown in table 5.1. Recall they are calculated by summing all m-driver coupled influences where the driver listed in the row header is included, then divided by the number of processes. The total influences are the $m$links summed across the row.

Both the smallest direct and smallest total influences come from u_l, at 10.4% and 15.7%, respectively. It does contribute roughly as much as its direct influence via its 2link and 3link, but much of this gain is corrected by the 4link. Most of its moderate 3link comes from the $\{u_l, \theta_e, \Delta T\}$ coupled influence. In net, about one third of its total influence comes from interaction terms.

While u_u yields the largest direct influence, it largely acts alone, making it have the third largest total influence. Its two-driver coupled influences with ΔT or u_l largely cancel, making the coupled

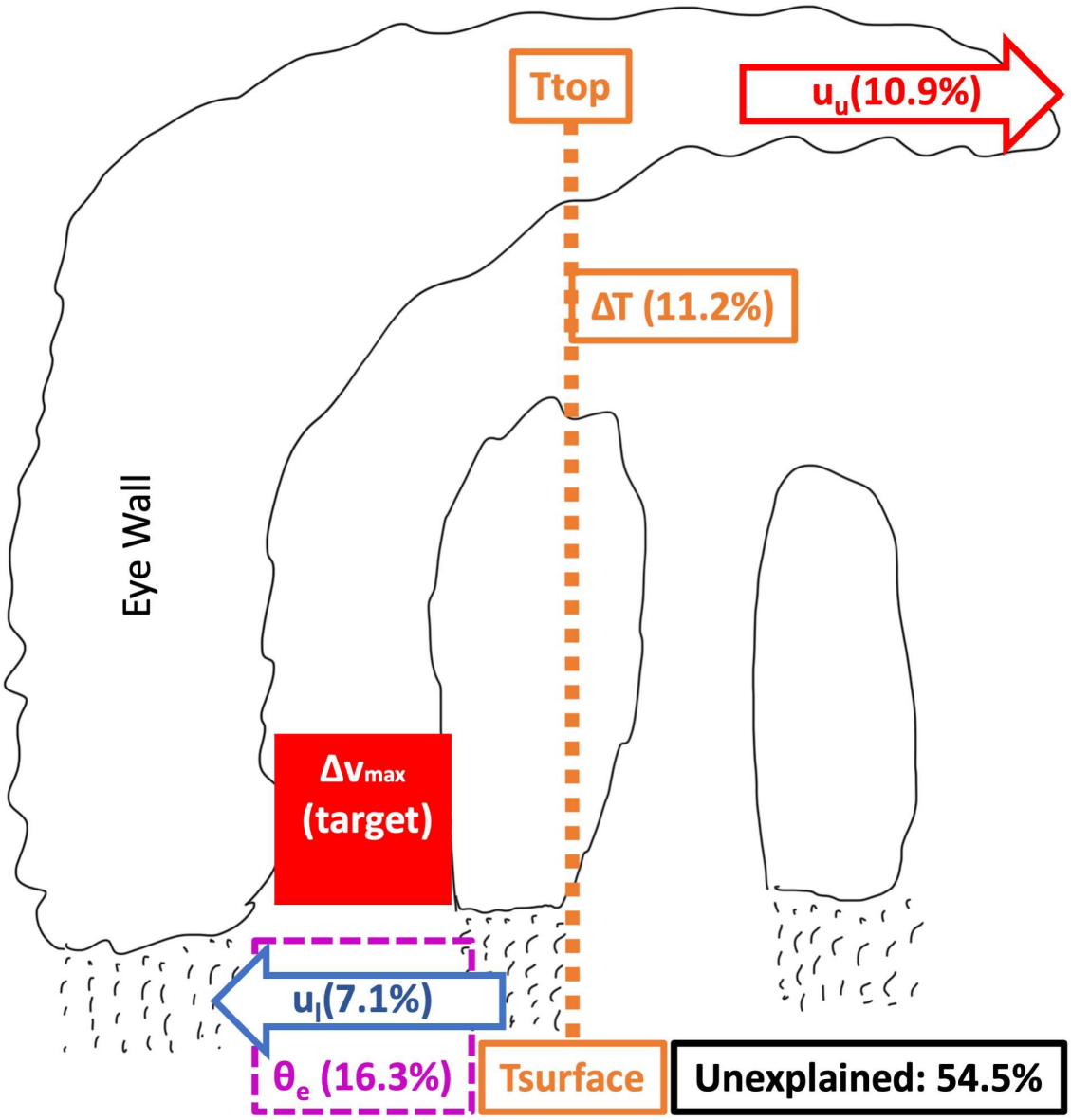

FIG. 5.1. Axisymmetric cross section of a generic hurricane with total influences superimposed approximately where the processes were located. Upper-level radial wind, u_u accounted for 10.9% of our certainty, boundary layer θ_e at the radius of maximum wind (RMW) for 16.3%, top-bottom temperature difference, ΔT, for 11.2%, and boundary layer cross-RMW radial wind, u_l, for 7.1%. Overall, the processes left 54.5% of the certainty unexplained.

interaction with θ_e the main source of its 2link. But, moving from 21.1% in direct influence to 24.0% in total, u_u contributes least of all via interactions.

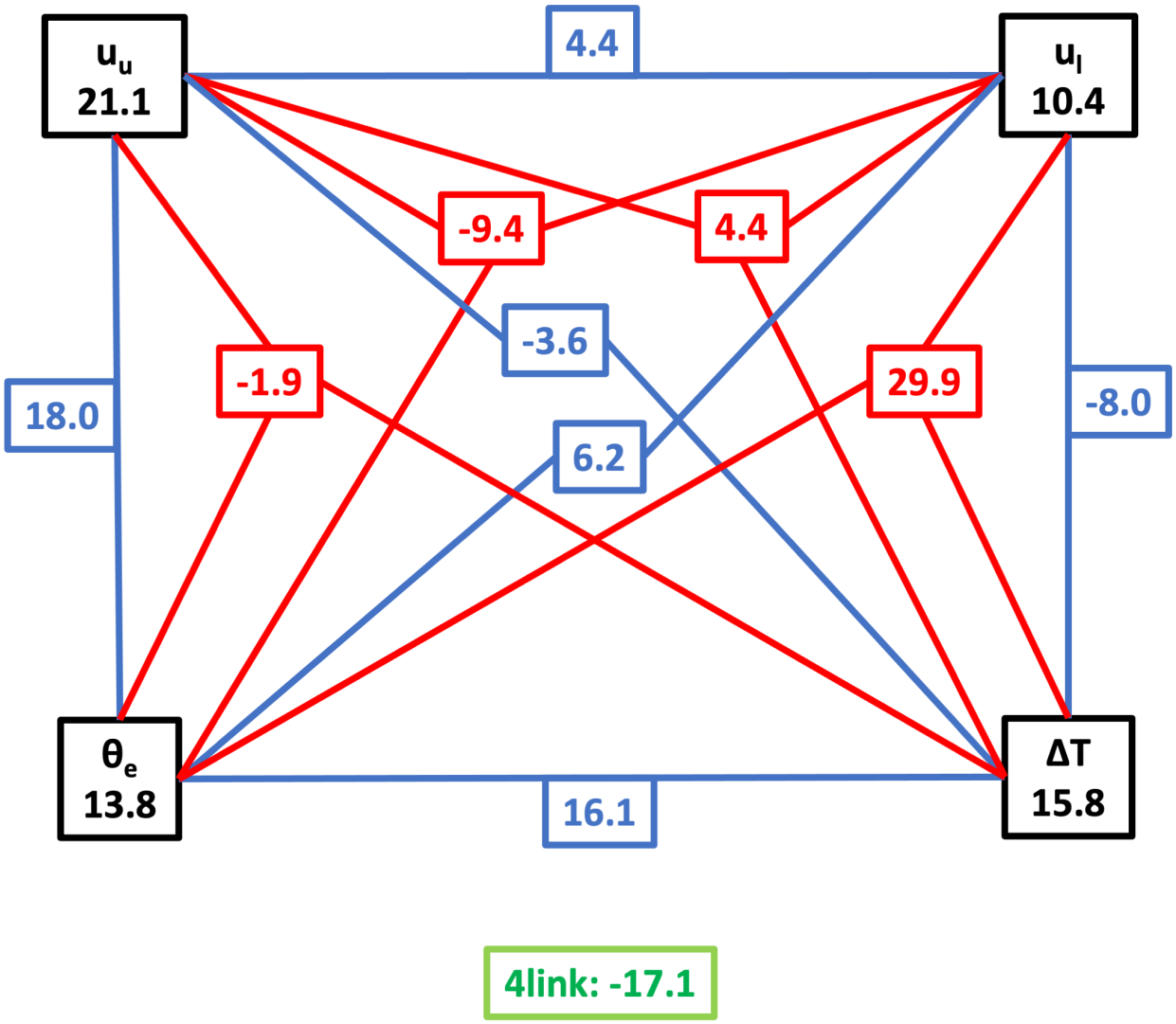

FIG. 5.2. Causal web showing direct and coupled influences as percentages of the 45.5% certainty explained. The target, Δv_{max}, and the lag of each driver, which was 1hr, are implied. The direct influences are shown in the black boxes containing the driver labels. The influences from two drivers are shown in the blue boxes attached to blue lines which connect the two constituent drivers. The influences from three drivers are shown in the red boxes at the intersection of three red lines, each line connecting to one of the constituent drivers. The influence from all four drivers together is shown in the green box below, and not connected to, the rest of the web.

The source of the second largest direct influence, ΔT, remained second in total influence. Its total influence is only slightly greater than that of u_u, though. But, its coupled influences yield positive 2links and 3links, increasing its total influence by about 13%. Like with u_l, the 3link its strongest level of coupling mostly due to the $\{u_l, \theta_e, \Delta T\}$ term. In net, interactions increase the influence of ΔT from 15.8% to 24.5%.

TABLE 5.1. Table decomposing the certainty gain into $m$link influences from each driver. All values are percentages of the certainty gain. Note that the 1link influences are simply the direct influences. Meanwhile, the 4link influence is the same for every driver, as it is the coupled influence of all four drivers divided four ways.

Process	1link	2link	3link	4link	Total
u_u	21.1	9.4	-2.3	-4.3	24.0
u_l	10.4	1.3	8.3	-4.3	15.7
θ_e	13.8	20.1	6.2	-4.3	35.8
ΔT	15.8	2.2	10.8	-4.3	24.5

Though θ_e contributed the third largest direct influence, it contributed the most in total influence. It was the only process contributing more via interactions than by direct action. Specifically, it seemed most active at the 2link level, where all its coupled influences are positive. Almost half of this comes from its coupling with u_u, with its coupling with ΔT being another large contribution. Of all the positive 3links, the 3link from θ_e was the smallest, largely because two of its three three-driver coupled influences were negative. Overall, θ_e directly contributed 13.8%, but its total influence accounted for 35.8% of the certainty gain.

5.3 Discussion

The results for u_u appear to corroborate existing narratives. A large upper-level outflow removes air aloft, thereby decreasing the surface pressure and sharpening the pressure gradient. This sharper gradient then allows for a more intense storm in order to maintain a balanced state. The storm will also contract, which will intensify the storm to conserve angular momentum. The large direct influence from u_u, which is the second largest influence in this study, evidences just how important this role of u_u is for driving intensification.

The role of u_u in this particular ensemble, however, may suggest something more. At the initialization time of the ensemble, synoptic-scale forcing was present. While this may be the source of u_u at the beginning, the memory of the synoptic-scale forcing diminishes as the simulation continues. Thus, in the later stages, when most of the ensemble members experience a second intensification, the simulated hurricane itself may be the source of u_u. This suggests the possibility of a self-intensifying storm as u_u would then feedback to intensify the storm.

If the outflow is at least partially storm-driven, anamolous warming of the core may explain most of it. This would cause the air in the core to expand, causing pressure to rise aloft. The circulation aloft

then becomes super gradient. This triggers a larger outflow over the core of the hurricane, thereby intensifying the storm.

This pushes the burden of explanation onto finding a source for the anamolous warming, and preferably one driven by the storm itself. Once the eye forms, there is subsidence within the eye, which could serve to further warm the eyewall. But as Vigh and Schubert (2009) showed, diabatic, i.e. latent, heating in the eyewall itself efficiently drives intensification. This is evidenced by the coupled influence of $\{u_u, \theta_e\}$, which is the third largest contribution to our certainty gain. A large u_u over the eyewall allows high θ_e air to enter and rise. As it rises, the water condenses and releases large amounts of heat into the upper layers. This high θ_e air could not have been as effective if u_u was not also acting simultaneously to remove air aloft, which would explain the large coupled term. The positive coupled influence from u_l and θ_e also supports this warming narrative.

The radial winds were expected to be very inseparable because of how the secondary circulation is defined. Instead, the winds are only weakly inseparable, possibly due to where the variables are defined. The location of u_l is always defined by the RMW, making it always part of the secondary circulation. Meanwhile, the location of u_u is sometimes over the eyewall, and it is sometimes within the eye. When it is over the eyewall, u_u is part of the secondary circulation, thereby making it inseparable from u_l. When it is within the eye, however, u_u represents air moving out of the center, and therefore is not part of the secondary circulation. This effect is therefore separable from the effect of u_l, so their overall effects are only weakly inseparable.

There is an interesting narrative involving u_u, u_l, and θ_e. Note that the two-driver couplings $\{u_l, u_u\}$, $\{u_l, \theta_e\}$, and $\{u_u, \theta_e\}$ all suggest inseparable effects, but the effects of $\{u_u, u_l, \theta_e\}$ are separable. This implies that the effects of two-driver couplings themselves are separable, even if the effects of the drivers involved in each two-driver coupling are not separable. This suggests that u_l and u_u each couple with θ_e, but these two-driver coupled effects are separable. Physically, this may mean that u_u and u_l each present valid pathways for high θ_e air to enter the core, and they do not need to cooperate to govern the flow of high θ_e air. That is, the effect of u_l bringing high θ_e air into the core is separable from the effect of u_u allowing the air to rise. There also may be a more complex story based on the timing of events like there was with u_u and u_l.

The direct influence of ΔT evidences its possible role of determining at least the maximum rate of intensification. And, this could in turn suggest that the view of a hurricane as a Carnot engine, while

correct, is incomplete. Of course, more explicitly physical studies are necessary to test this. One possible nonphysical explanation is that ΔT is a proxy for the vertical motion of the secondary flow. Note that u_u and u_l are components of the secondary flow. The two-driver coupled influences of ΔT with either u_u or u_l are negative, while their three-driver coupled influence is positive. Unlike what happened with u_u, u_l, and θ_e, the two-driver coupled effects are inseparable even though the effects of ΔT are separable from either u_u or u_l. Thus, the three drivers together are inseparable, suggesting they may together represent the secondary circulation.

Despite the enigmatic role of ΔT, it is involved with by far the largest contribution toward the certainty gain. In fact, this could also evidence that ΔT contains information about the vertical motion in the secondary circulation, as that dictates the path of high θ_e air beyond the boundary layer. What this term does clearly evidence, along with the large term from $\{\theta_e, \Delta T\}$, is that thermodynamics played an important role in Patricia's RI.

Aside from the four-driver coupled influence, the other negative coupled influences arise when u_u and θ_e are coupled with either u_l or ΔT, and when ΔT is coupled with either u_u or u_l. Together, they may evidence the changing importance of processes during the RI of Patricia. When Patricia was first starting, the mechanical driving, and so dynamical variables like u_u and u_l, may have been more important. Later, when Patricia was formed and the dynamics were stable, the thermodynamics, represented here by θ_e and ΔT, gave her a second wind. These modes of driving are vastly different, which could lead to the large negative values.

In the end, however, θ_e proved to be the most responsible of these four for driving intensification. It drove RI mostly be being worked on. The hot, humid air did little on its own, as evidenced by only one storm in the same region four months prior to Patricia (Kimberlain et al. 2016). But, with u_u allowing it to rise, u_l directing it into the core, and ΔT governing the connection between the boundary and outflow layers, the hot humid air was able to realize its full potential. This agrees strongly with the findings by Hu and Wu (2020).

One thing that needs mention is the relatively large self-certainty of the target. With 45.5% of our conditional certainty coming from the drivers, they failed to explain even a simple majority. This does not, however, reduce the above discussion to nothing. For example, having only 45.5% come from certainty gain does not change the fact that θ_e was responsible for 35.8% of that gain. It just means that θ_e is only $45.5\% \times 0.358 = 16.3\%$ of the full story. And overall, the above discussion is only about 45.5% of the full story, hinting that there is still much more physics to be had.

5.4 Conclusions and Moving Forward

Despite having only four azimuthally averaged drivers and a noisy target, the drivers explained about 45.5% of $W(\Delta v_{max}|\mathbf{Y})$. The framework yielded $2^4-1=15$ direct and coupled influences, together yielding explanations of Patricia's RI that were rich in physics. How $I(\Delta v_{max};\mathbf{Y})$ decomposed appears to agree with most of the existing literature. Namely, the solo action of outflow drives intensification mechanically, while high θ_e air in the boundary layer at RMW drives the hurricane thermally.

Specifically, the influence of θ_e by and large comes from its inseparability. Without the secondary circulation, the hot, humid air did little to directly act. But, with the secondary circulation, the air was then in a position to drive intensification by warming the core. Such interactions showed that θ_e is actually ~~the~~ responsible for most of the certainty gain.

Some of the possible physical explanations were new to this analysis. For one, given the nature of the simulations, the concept of self-driven intensification was plausible. For another, the hypothetical view that a TC not only acts like a Carnot engine, but also serves as a physical bridge between the boundary and outflow layers, seems plausible given the importance of the direct, coupled, and total influences of ΔT.

There are many ways to improve upon the current study. For example, azimuthal averages, like all averages, are great summaries that smooth over many of the features which contain vital information. To increase the extractibility and availability of information content, we could perform principle component analysis and use the resulting principle components instead. This is still a summary, but it is more of a summary of information-containing features rather than smoothing over features. It would also implicitly include more drivers in the same number of time series.

Another possibility for improvement could come from using v_{max} instead of Δv_{max} as a target. This would require including v_{max} as a potential driver, which might result in excluding one of the drivers in the present study in order to preserve the accuracy of information estimation. But, the noise-to-signal ratio will be decreased potentially 14 times, thereby yielding a much higher availability of information content. This might also serve to increase the reliability of the information estimators. The only concern with this would be that the target is not as direct a proxy to intensification as Δv_{max} is.

On top of this, we are still discussing if 45.5% is an acceptable amount of explanation. Currently, we are of the opinion that the study is not diminished. In fact, we are hopeful that this drives more research to reduce the 55.5% of unexplained certainty!

CHAPTER 6

Generalizing the Certainty Framework to Precipitation-like Targets

To motivate generalizing the certainty framework to noncontinuous targets, consider precipitation. Either there is no precipitation, or there is some continuously distributed value of precipitation. In other words, precipitation is partially discrete and partially continuous, or quasidiscrete. Precipitation is not the only example of a noncontinuous variable. Discrete variables, too, are a type of noncontinuous variable, examples of which include some implementations of transfer coefficients in convection-permitting models and categorizing phenomena like El Niño Southern Oscillation as strong, weak, neutral, or La Niña. The rest of this section will be devoted to treating precipitation.

The distribution (pdf) of precipitation can be represented continuously as

$$p_R(r) = a\,\delta(r) + (1-a)\,p_{R>0}(r), \tag{6.1}$$

where r is the amount of precipitation, $0 \leq a \leq 1$ is the proportion of the observations without precipitation, $p_{R>0}$ is a true probability density, i.e. integrates to 1 and is nonnegative everywhere, *which is zero when* $r \leq 0$, and δ is the Dirac delta. The Dirac delta is the continuous equivalent of the Kronecker delta, with the property

$$\int_{-\infty}^{\infty} f(r)\delta(r-r_0)\,dr = f(r_0), \tag{6.2}$$

where r_0 is some value in the domain of R. In other words, the value of the entire integral is the value of whatever is multiplied by the Dirac delta whenever the Dirac delta's argument is zero. This property means that the probability density of precipitation (Eq. (6.1)) integrates to 1, making it a true distribution.

The joint pdf of precipitation and a driver, Y, becomes

$$p_{R,Y}(r,y) = a\,\delta(r)\,p_Y(y|r=0) + (1-a)\,p_{R>0,Y}(r,y). \tag{6.3}$$

The mutual information between R and Y is then

$$I(R;Y) = \int p_{R,Y}(r,y)\log\frac{p_{R,Y}(r,y)}{p_R(r)p_Y(y)}\,dr\,dy \tag{6.4}$$

$$= \int \left[a\,\delta(r)\,p_Y(y|r=0) + (1-a)\,p_{R>0,Y}(r,y)\right]\log\frac{a\,\delta(r)\,p_Y(y|r=0) + (1-a)\,p_{R>0,Y}(r,y)}{[a\,\delta(r) + (1-a)\,p_{R>0}(r)]\,p_Y(y)}\,dr\,dy. \tag{6.5}$$

This integral splits into two parts: one where $r = 0$, and another where $r > 0$. In the former, only the Dirac delta part of p_R remains because the continuous portion of the distribution is zero. Similarly, in the latter integral, only the continuous portion because the Dirac delta is zero. Thus,

$$I(R;Y) = \int (a\delta(r)p_Y(y|r=0))\log\frac{a\delta(r)p_Y(y|r=0)}{[a\delta(r)]p_Y(y)}\,dr\,dy + \int (1-a)p_{R>0,Y}(r,y)\log\frac{(1-a)p_{R>0,Y}(r,y)}{[(1-a)p_{R>0}(r)]p_Y(y)}\,dr\,dy. \tag{6.6}$$

In the logarithm of the first integral, the $a\delta(r)$ terms cancel in the limit as $r \to 0$, while in the logarithm of the second integral, only the $(1-a)$ terms cancel. This leaves

$$I(R;Y) = \int (a\delta(r)p_Y(y|r=0))\log\frac{p_Y(y|r=0)}{p_Y(y)}\,dr\,dy + \int (1-a)p_{R>0,Y}(r,y)\log\frac{p_{R>0,Y}(r,y)}{p_{R>0}(r)p_Y(y)}\,dr\,dy \tag{6.7}$$

$$= aD_{KL}(p_{Y|r=0}||p_Y) + (1-a)I_{R>0}(R;Y). \tag{6.8}$$

In the first integral, the Dirac delta returns $p_Y(y|r=0)$ whenever $r = 0$. This means the first integral is the KL divergence of the conditional pdf of Y when there is no precipitation from the marginal pdf of Y, multiplied by the proportion of observations lacking precipitation. The second integral is clearly a mutual information between continuous variables, multiplied by the proportion of observations with precipitation.

The KL divergence comes from when precipitation is discrete, while the mutual information comes from when it is continuous. This shows that the framework itself, when applied to precipitation, can be split into a discrete portion and a continuous portion! To combine the two parts, the results from the discrete part are simply multiplied by the probability of a dry observation, and the results from the continuous part are multiplied by the probability of a wet observation.

To calculate relative influence for the discrete results, the framework uses Shannon entropy like in McGill (1954), but only the part of the Shannon entropy coming from dry observations. That is, instead of thinking of dry observations themselves as a separate unary variable, i.e. having only one possible value, the dry observations are part of a binary variable. More plainly, even though the dry portion of precipitation is the target of the discrete part of the framework, the dry observations are not the only observations in the discrete portion of the framework. While this may seem incorrect, note that the marginal pdf of Y is used in the KL divergence. Since the marginal pdf still contains wet observations,

wet observations are considered in the dry part of the framework. Furthermore, the Shannon entropy of a unary variable is always zero, while the KL divergence coming from the discrete portion of the framework may be positive, meaning normalization is impossible in this case. For these two reasons, I argue that partial Shannon entropy is the correct normalization for the discrete part of the framework.

To calculate partial Shannon entropy, first treat precipitation as a binary (wet or dry) variable, and use only the portion of the Shannon entropy that comes from when it is dry. In other words,

$$H_{R=0}(R) = -a \log a = a D_{KL}(p_{Y|r=0} || p_Y) + H_{R=0}(R|Y), \tag{6.9}$$

where $H_{R=0}(R)$ is the partial Shannon entropy. Thus, the physics *of the dry observations* that is captured by the drivers is

$$\frac{a D_{KL}(p_{Y|r=0} || p_Y)}{H_{R=0}(R)} = -\frac{1}{\log a} D_{KL}(p_{Y|r=0} || p_Y), \tag{6.10}$$

while the unexplained forcing is

$$\frac{H_{R=0}(R|Y)}{H_{R=0}(R)} = \frac{1}{\log a}(\log a + D_{KL}(p_{Y|r=0} || p_Y)). \tag{6.11}$$

The generalized certainty framework differs from the original only in calculating relative influence. The framework can now be applied to quasidiscrete targets with only one value that is considered discrete. More work is needed to generalize the framework further. For example, if a quasidiscrete variable has more than one value at which it is discrete, the question is whether the discrete values should be analyzed together or individually. If the discrete modes are analyzed together, then the results for a completely discrete target follow the framework from McGill (1954) exactly. But, if the discrete modes are analyzed individually, then each mode has a separate analysis that is normalized by its partial Shannon entropy. Which form of analysis is better requires future investigation.

CHAPTER 7

Concluding Remarks and Moving Forward

This thesis expanded upon the certainty framework proposed by van Leeuwen et al. (2021). A thermodynamic interpretation of the differences between entropy and certainty was introduced. The concept of separability of effects was introduced, and two-driver coupled influences were shown to evidence the separability of the effects of two drivers, with negative numbers implying separability and positive numbers implying inseparability. Furthermore, a new reference density was introduced specifically for targets with high noise. Different expressions for coupled influences, $m$links, and total influences are proven in the Appendix. Of great importance was that the theoretical value of total influence is nonnegative, making the decomposition of mutual information into total influences nonnegative.

Implementing the framework was detailed. The choice of computer language was discussed, as was the choice of algorithm for mutual information estimation. To make the estimators more accurate, the k-nearest neighbors (kNN) algorithms of Kraskov et al. (2004) and Vejmelka and Paluš (2008) were refined, the data in the time series were made marginally Gaussian, and new criteria for determining the parameter k were developed. Combining the information estimations to calculate the coupled influences, $m$links, and total influences was detailed. Calculating self-certainty without explicitly implementing a KL divergence estimator was described.

The framework was applied to study the rapid intensification of Hurricane Patricia (2015). To reduce numerical error in the implementation due to the curse of dimensionality, the study was limited to four drivers. These were radial wind in the outflow layer (u_u), radial wind in the boundary layer (BL) at the radius of maximum wind (RMW) (u_l), the equivalent potential temperature at BL RMW (θ_e), and the difference in temperature between the BL and outflow layer (ΔT). The target was the hourly change in maximum tangential wind between the heights 600m and 1km. The drivers explained 45.5% of the target's total certainty.

Decomposing the certainty gain into coupled influences evidenced rich physics based on the separability and inseparability of the drivers. The largest direct contribution was from u_u, which together with relatively weak $m$links was interpreted as u_u decreasing the central pressure and thereby allowing the storm to intensify. The only large coupled influence of u_u was with θ_e, which was interpreted as u_u allowing high θ_e air to convect in the eyewall. Other direct contributions were not nearly as large.

By considering more than one coupled influence, more complex physics was evidenced. For example, the three-driver coupled influence from $\{u_u, u_l, \theta_e\}$ evidenced separability of the effects of the three constituent drivers, but the coupled influence of any two of these three drivers evidenced inseparability. This was interpreted as the two-driver effects themselves being separable, specifically that the pathway of u_u and θ_e acting together is separable from the pathway of u_l and θ_e acting together. For another example of this kind of reasoning, there was another three-driver coupled influence, from $\{u_u, u_l, \Delta T\}$, that instead suggested inseparability of the three drivers, even though the two-driver coupled influences ΔT had with either u_u or u_l suggested separability. This was interpreted that all three drivers together might represent one process, namely the secondary circulation, which was not evidenced by considering the above two-driver coupled influences alone. For a final example, it was suggested that the coupled influences that were negative evidence that the mode of Patricia's RI changed from a dynamic to a thermodynamic origin during the RI.

The framework was generalized to handle variables like precipitation. Precipitation time series in this generalization were either zero or some continuously distributed positive value. The framework was split into studying the one discrete value separate from the continuous part of the distribution. In doing so, partial Shannon entropy was introduced to normalize the discrete portion of the framework. Further generalization to variables with more than one discrete value was mentioned but not evaluated further.

7.1 Future Work

7.1.1 Hurricane Patricia

With the study of Hurricane Patricia, other drivers or targets should be considered for network discovery. For example, studying v_{max} is different from studying Δv_{max}. Since the time series for v_{max} is less affected by uncertainty, however, recovering the analysis of Δv_{max} from the analysis of v_{max} would be convenient. Other drivers many include principal components from principal component analysis, which could potentially capture the structure of the hurricane and determine what structures lead to intensification.

At the same time, the results from the current study already provide interesting leads for future, more focused studies. For instance, whether of not the simulated u_u came from synoptic forcing needs to be determined to verify or reject the hypothesis of self-driven outflow. Also, whether or not ΔT was

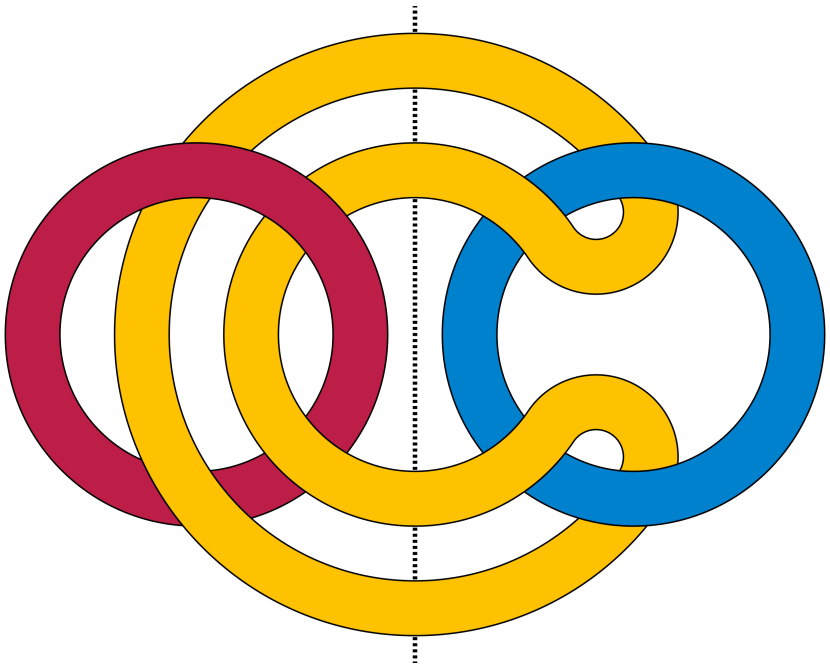

FIG. 7.1. Algebraic link diagram of Borromean rings. Note that cutting any one ring leaves the other two rings unlinked. Image made public domain by David Eppstein via Creative Commons. Image URL https://commons.wikimedia.org/wiki/File:Algebraic_Borromean_link_diagram.svg.

active or passive, or even just a proxy for another process, needs to be determined. These future studies should be designed to determine the physical mechanisms which led to these results, which will in turn provide feedback on the abilities of the framework to detect physical phenomena.

7.1.2 Interpretations and Results in the Framework

As stated in Section 3.2.1, the interpretations of two-driver coupled influences may not hold for coupled influences from more than two drivers. Two potential interpretations need to be studied further. The results in the TCRI application hint at an interpretation related to Brunnian links, a concept from knot theory. A famous example is a set of Borromean rings, which are three rings that are three-wise linked but not pairwise linked. (See Fig. 7.1.) This interpretation is evidenced by showing that the coupled influence of three drivers Y, Z, W on X is

$$ {}^{X}_{YZW}M_{\mathbf{Y}} = II(X,Y,Z,W) = II(X,Y,Z) - II(X,Y,Z|W) = {}^{X}_{YZ}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{YZ}M_{\mathbf{Y}}, \tag{7.1} $$

where ${}^{X}_{YZ}M_{\mathbf{Y}\setminus\{W\}}$ is the coupled influence of Y and Z if W was not included in the analysis. When ${}^{X}_{YZW}M_{\mathbf{Y}} > 0$, this means ${}^{X}_{YZ}M_{\mathbf{Y}\setminus\{W\}} > {}^{X}_{YZ}M_{\mathbf{Y}}$, which does not show if either ${}^{X}_{YZ}M_{\mathbf{Y}\setminus\{W\}}$ or ${}^{X}_{YZ}M_{\mathbf{Y}}$ are positive or negative. That is, the effects of Y and Z become more inseparable when W is excluded

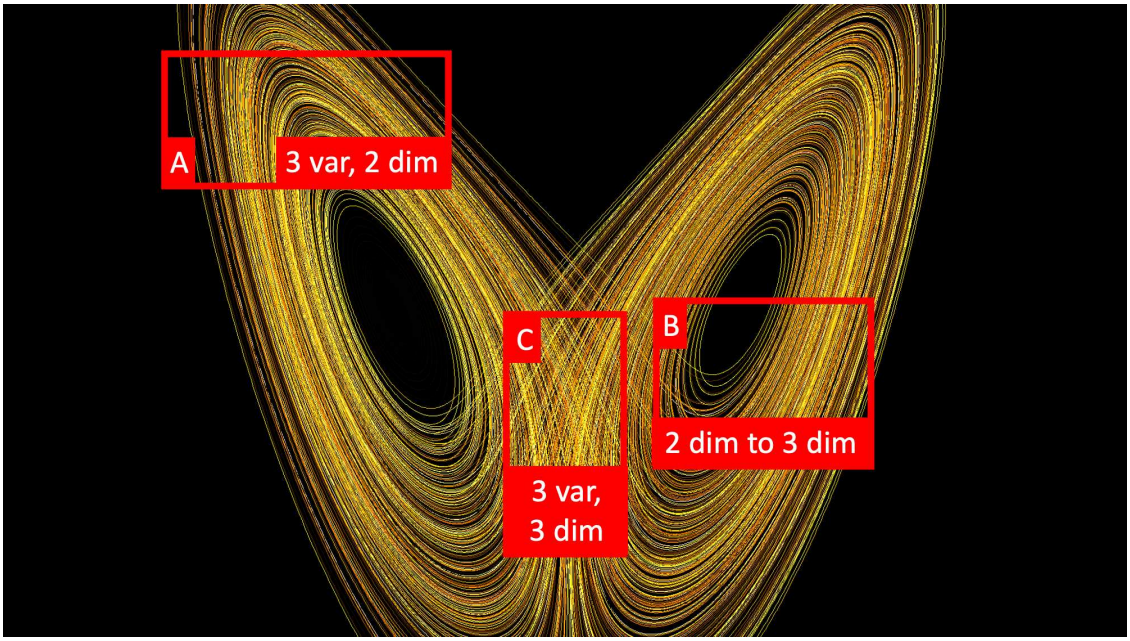

FIG. 7.2. An annotated evaluation of the Lorenz 1963 system. In region A, the system clearly exhibits two dimensional behavior despite being represented by three variables. In region B, the system transitions between being two and three dimensional. In region C, the system is clearly three dimensional. The dimension of the system overall is 2.4013 (Kuznetsov et al. 2020). Image annotations by me. Original image made public domain by Wikimol via Creative Commons. Original image URL https://en.wikipedia.org/wiki/File:Lorenz_system_r28_s10_b2-6666.png.

than when W is included, whether or not the effects of Y and Z are separable. But, because this value is symmetric for Y, Z, W, excluding any one makes the other two more inseparable, suggesting the three variables are three-ways linked so that the effects of one is marginalized into the others. Inversely, a negative three-driver coupled influence implies that excluding any one driver makes the other two more separable, suggesting that the drivers are not three-ways linked, so the effects of one is not marginalized into the others. As evidence for this interpretation in the physical world, I cite the interpretations used for the TCRI study. In general, an m-driver coupled influence may suggest whether or not the drivers are m-linked in the topological sense.

Another possible interpretation involves joint action by multiple drivers. I have yet to clearly define action, let alone joint action and how it relates to coupled influence. But, one of the properties of joint action is a reduction in overall dimension of the system. For example, consider the classical Lorenz 1963 system using standard parameter values (Fig. 7.2). While it is represented using three dimensions, the Hausdorf dimension of the global attractor, i.e. the attractor on the region with stable oscillations, is 2.4013 (Kuznetsov et al. 2020). A reduction of dimension may evidence joint action, and how this

joint action appears in coupled influence should be investigated. Whether or not joint action relates to sufficient-component causation should also be investigated.

Another topic to study is how results change when a driver is removed from the analysis. If a driver's direct influence is zero, removing it from a study does not change the certainty gain. Furthermore, a zero direct influence means the driver is d-separated by the other drivers. But, when a driver's 1link is nonzero, how the results change if the driver is removed is unknown. Early evidence (not shown) suggests that, when a driver has a small relative total influence, removing it did not change the order of the total relative strength of the remaining drivers. Inversely, removing a driver with a large relative total influence dramatically changed the final order of importance. Furthermore, how the causal web changes upon removal of any driver would need to be studied.

7.1.3 Implementation

Making the implementation faster is always a positive result, so long as accuracy is not compromised. Speed gains could be realized by implementing the framework in FORTRAN, as the language performs vector and matrix operations faster than any other language. Based on the same reasoning, designing a graphical processing unit-friendly version could offer an even faster implementation.

The information estimators themselves could be improved, both for speed and accuracy gains. Instead of recombining (conditional) mutual informations to calculate the coupled influences, designing an estimator for coupled influences directly would be both faster and potentially more accurate. The information estimator from Kraskov et al. (2004) is extendable to interaction information by another formulation they introduce, but there seems to be no existing estimator for conditional interaction information, which is what the coupled influences are. The norm used for kNN distance may also affect the results. While the maximum norm is simple and efficient, perhaps using another norm would make the estimators more efficient.

Speed and accuracy gains could also come from transforming the joint distribution of the target and drivers to be multivariate Gaussian. There are analytic equations for information theoretic terms for variables that are multivariately Gaussian distributed. The Shannon entropy of a set X, assumed to be multivariate Gaussian system, or a subset thereof, is

$$H(X) = \frac{1}{2} \ln \det(2\pi e \boldsymbol{\Sigma}_X), \tag{7.2}$$

where $\mathbf{\Sigma}_X$ is the covariance matrix of X. We can use this directly in information calculations, as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \text{ and } I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z). \quad (7.3)$$

These equations were used to rigorously test the information estimators, and being able to use these equations would also greatly speed up the framework. This shows the need to develop a fast algorithm for multivariate Gaussianification such that the theoretical value of mutual information is preserved, if such a transformation exists.

Needless to say, KL divergence estimators will be implemented. The self-certainty estimation is as accurate as the Shannon entropy estimator, but a KL divergence estimator may still be better. Furthermore, in order to implement the generalized framework, a KL divergence estimator is absolutely needed. Previous implementation attempts had large systematic errors.

REFERENCES

Aupetit, M., 2009: Nearly homogeneous multi-partitioning with a deterministic generator. *Neurocomputing*, **72 (7-9)**, 1379–1389, doi: 10.1016/j.neucom.2008.12.024.

Barnes, E. A., S. M. Samarasinghe, I. Ebert-Uphoff, and J. C. Furtado, 2019: Tropospheric and stratospheric causal pathways between the mjo and nao. *J. Geophys. Res.*, **124 (16)**, 9356–9371, doi: 10.1029/2019JD031024.

Barrett, A. B., 2015: Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Phys. Rev. E*, **91 (5)**, 052 802, doi: 10.1103/PhysRevE.91.052802.

Chickering, D. M., 2002: Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res*, **2**, 445–498, URL https://www.jmlr.org/papers/volume2/chickering02a/chickering02a.pdf.

Davis, C. A. and L. F. Bosart, 2004: The tt problem. *Bull. Amer. Meteor. Soc.*, **11**, 1657–1662, doi: 10.1175/BAMS-85-11-1657.

Ebert-Uphoff, I. and Y. Deng, 2012: Causal discovery for climate research using graphical models. *J. Climate*, **25 (17)**, 5648–5665, doi: 10.1175/JCLI-D-11-00387.1.

Emanuel, K., 1986: An air-sea interaction theory for tropical cyclones. part i: Steady-state maintenance. *J. Atmos. Sci.*, **43 (6)**, 585 – 605, doi: 10.1175/1520-0469(1986)043<0585:AASITF>2.0.CO;2.

Emanuel, K., 1991: The theory of hurricanes. *Annu. Rev. Fluid Mech.*, **23 (1)**, 179–196, doi: 10.1146/annurev.fl.23.010191.001143.

Emanuel, K., 2012: Self-stratification of tropical cyclone outflow. part ii: Implications for storm intensification. *J. Atmos. Sci.*, **69 (3)**, 988 – 996, doi: 10.1175/JAS-D-11-0177.1.

Emanuel, K. and R. Rotunno, 2011: Self-stratification of tropical cyclone outflow. part i: Implications for storm structure. *J. Atmos. Sci.*, **68 (10)**, 2236 – 2249, doi: 10.1175/JAS-D-10-05024.1.

Granger, C., 1963: Economic processes involving feedback. *Information and Control*, **6 (1)**, 28–48, doi: 10.1016/S0019-9958(63)90092-5.

Granger, C., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37 (3)**, 424–438, doi: 10.2307/1912791.

Gray, W. M., 1968: Global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.*, **96 (10)**, 669–700, doi: 10.1175/1520-0493(1968)096<0669:GVOTOO>2.0.CO;2.

Hu, C.-C. and C.-C. Wu, 2020: Ensemble sensitivity analysis of tropical cyclone intensification rate during the development stage. *J. Atmos. Sci.*, **77 (10)**, 3387 – 3405, doi: 10.1175/JAS-D-19-0196.1.

Kimberlain, T. B., E. S. Slake, and J. P. Cangialosi, 2016: Hurricane patricia. Tech. Rep. EP202015, National Hurricane Center. URL https://www.nhc.noaa.gov/data/tcr/EP202015_Patricia.pdf.

Koopman, J. S., 1981: Interaction between discrete causes. *Amer. J. Epidemiol.*, **113 (6)**, 716–724, doi: 10.1093/oxfordjournals.aje.a113153.

Kraskov, A., H. Stögbauer, and P. Grassberger, 2004: Estimating mutual information. *Phys. Rev. E*, **69 (6)**, 066 138, doi: 10.1103/PhysRevE.69.066138.

Kretschmer, M., D. Coumou, J. F. Donges, and J. Runge, 2016: Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *J. Climate*, **29 (11)**, 4069–4081, doi: 10.1175/JCLI-D-15-0654.1.

Kuznetsov, N. V., T. N. Mokaev, O. A. Kuznetsova, and E. V. Kudryashova, 2020: The lorenz system: Hidden boundary of practical stability and the lyapunov dimension. *Nonlinear Dyn.*, **102**, 713–732, doi: 10.1007/s11071-020-05856-4.

Matthewman, N. J. and G. Magnusdottir, 2011: Observed interaction between pacific sea ice and the western pacific pattern on intraseasonal time scales. *J. Climate*, **24 (19)**, 5031–5042, doi: 10.1175/2011JCLI4216.1.

McBride, J. L. and R. Zehr, 1981: Observational analysis of tropical cyclone formation. part ii: Comparison of non-developing versus developing systems. *J. Atmos. Sci.*, **38 (6)**, 1132–1151, doi: 10.1175/1520-0469(1981)038<1132:OAOTCF>2.0.CO;2.

McGill, W., 1954: Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, **4 (4)**, 93–111, doi: 10.1109/TIT.1954.1057469.

McGraw, M. and E. Barnes, 2018: Memory matters: A case for granger causality in climate variability studies. *J. Climate*, **31 (8)**, 3289–3300, doi: 10.1175/JCLI-D-17-0334.1.

McTaggart-Cowan, R., G. D. Deane, L. F. Bosart, C. A. Davis, and T. J. G. Jr., 2008: Climatology of tropical cyclogenesis in north atlantic (1948-2004). *Mon. Wea. Rev.*, **136 (4)**, 1284–1304, doi: 10.1175/2007MWR2245.1.

Merrill, R. T., 1988a: Characteristics of the upper-tropospheric environmental flow around hurricanes. *J. Atmos. Sci.*, **45 (11)**, 1665 – 1677, doi: 10.1175/1520-0469(1988)045<1665:COTUTE>2.0.CO;2.

Merrill, R. T., 1988b: Environmental influences on hurricane intensification. *J. Atmos. Sci.*, **45 (11)**, 1678 – 1687, doi: 10.1175/1520-0469(1988)045<1678:EIOHI>2.0.CO;2.

Ooyama, K., 1969: Numerical simulation of the life cycle of tropical cyclones. *J. Atmos. Sci.*, **26 (1)**, 3–40, doi: 10.1175/1520-0469(1969)026<0003:NSOTLC>2.0.CO;2.

Pearl, J., 1995: Causal diagrams for empirical research. *Biometrika*, **82 (4)**, 669–688, doi: 10.1093/biomet/82.4.669.

Pearl, J., 2000: *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Rothman, K. J., 2017: Causes. *Amer. J. Epidemiol.*, **185 (11)**, 1035–1040, doi: 10.1093/aje/kwx099.

Runge, J., 2015: Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E*, **92 (6)**, 062 829, doi: 10.1103/PhysRevE.92.062829.

Runge, J., 2018: Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos*, **28 (7)**, 075 310, doi: 10.1063/1.5025050.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, 2019a: Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, **5 (11)**, eaau4996, doi: 10.1126/sciadv.aau4996.

Runge, J., et al., 2019b: Inferring causation from time series in earth system sciences. *Nat. Commun.*, **10**, doi: 10.1038/s41467-019-10105-3.

Samarasinghe, S., E. Barnes, C. Connolly, I. Ebert-Uphoff, and L. Sun, 2021: Strengthened causal connections between the mjo and the north atlantic with climate warming. *Geophys. Res. Lett.*, **48 (5)**, e2020GL091 168, doi: 10.1029/2020GL091168.

Schreiber, T., 2000: Measuring information transfer. *Phys. Rev. Let.*, **85 (2)**, 461–464, doi: 10.1103/PhysRevLett.85.461.

Spirtes, P. and C. Glymour, 1991: An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.*, **9**, 62–72, doi: 10.1177/089443939100900106.

Spirtes, P., C. Glymour, and R. Scheines, 2000: *Causation, prediction and search (2nd ed.)*. Springer.

Strong, C., G. Magnusdottir, and H. Stern, 2009: Observed feedback between winter sea ice and the north atlantic oscillation. *Journal of Climate*, **22 (22)**, 6021 – 6032, doi: 10.1175/2009JCLI3100.1.

Sugihara, G., R. May, H. Ye, C. Hsieh, E. Deyle, M. Fogarty, and S. Munch, 2012: Detecting causality in complex ecosystems. *Science*, **338**, 496–500, doi: 10.1126/science.1227079.

Sun, J., D. Taylor, and E. Bollt, 2014: Causal network inference by optimal causation entropy. *SIAM J. Appl. Dyn. Syst.*, **14**, 73–106, doi: 10.1137/140956166.

Tao, D., M. Bell, R. Rotunno, and P. J. van Leeuwen, 2020: Why do the maximum intensities in modeled tropical cyclones vary under the same environmental conditions? *Geophys. Res. Lett.*, **47 (3)**, e2019GL085 980, doi: 10.1029/2019GL085980.

van Leeuwen, P. J., M. DeCaria, N. Chakaborty, and M. Pulido, 2021: A framework for causal discovery in non-intervenable systems. URL `https://arxiv.org/abs/2010.02247`.

Vejmelka, M. and M. Paluš, 2008: Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E*, **77 (2)**, 026 214, doi: 10.1103/PhysRevE.77.026214.

Vigh, J. L. and W. H. Schubert, 2009: Rapid development of the tropical cyclone warm core. *J. Atmos. Sci.*, **66 (11)**, 3335–3350, doi: 10.1175/2009JAS3092.1.

Wiener, N., 1956: Theory of prediction. *Modern Mathematics for the Engineer: First Series*, E. F. Beckenbach, Ed., McGraw-Hill, New York, chap. 8, 165–190.

Williams, P. L. and R. D. Beer, 2010: Nonnegative decomposition of multivariate information. *CoRR*, **abs/1004.2515**, URL `http://arxiv.org/abs/1004.2515`.

Winitzki, S., 2008: A handy approximation for the error function and its inverse. URL https://www.academia.edu/9730974/A_handy_approximation_for_the_error_function_and_its_inverse.

APPENDIX A

A.1 D-Separation Example

D-separation is a concept for directed acyclic graphs (DAGs) with statistically measurable consequences. Specifically, if two variables are d-separated by a set of variables on the same DAG, then the two variables are conditionally independent given that set. The set that is conditioned on may be empty, implying the two variables have no common ancestors.

To illustrate a more complex example for d-separation, consider the DAG in Figure A.1. Variables Z and W are d-separated by the empty set, meaning they are unconditionally independent. Similarly, because U is a child only of W, Z and U are d-separated by the empty set. The variable Y d-separates X from the rest of the graph. Variables U and V d-separate Y from W, and U and V are d-separated by W. Meanwhile, V with either W or U d-separates Y from Z.

It is tempting to say V, as the only child of Z, d-separates Z from the rest of the graph. Indeed, conditioning on V blocks the forward path from Z. But, because V is a collider for Z and W, conditioning on V opens a path between Z and W. This path is blocked by adding either W or U to the conditioning.

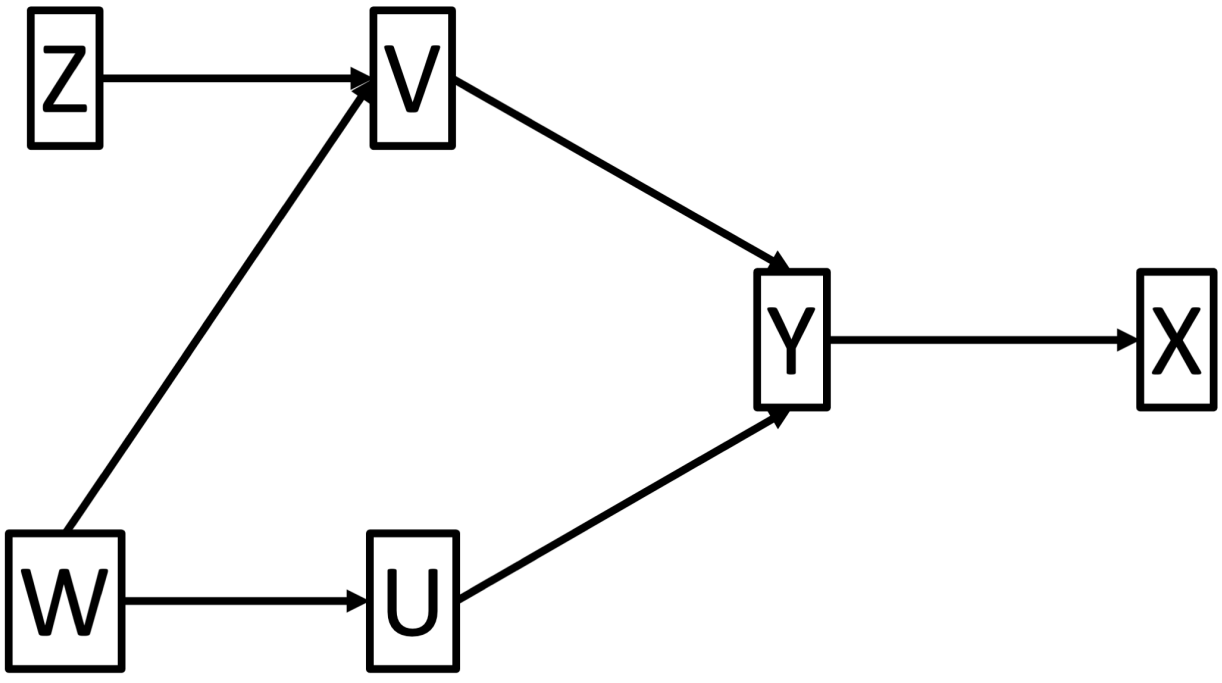

FIG. A.1. Standard directed acyclic graph with six nodes, two of which are root nodes.

A.2 Rewriting Influence as Sum of Mutual Informations

The development below is what the current implementation uses to calculate the coupled influences, as discussed in section 4.2. Calculating coupled influences is actually a newer addition than calculating the $m$links, even though this result is used to prove the next result. The reason is that the influences come from a combination of drivers, while the $m$links revolve around one driver acting with a few others. Before developing the representation of a combination by an integer (Section 4.2), it was difficult to loop through the combinations. This made calculating and storing the values in memory conceptually easier for $m$links.

Another product of this result is to simplify the dependencies that arise from the recursive definition, shown in equation (2.7). These dependencies create difficulty from constantly checking whether or not a coupled influence has been calculated in order to use it. And, coupled influences from large sets naturally have lots of overlapping dependencies. This result pushes the burden off of the recursive definition and onto already calculated mutual informations. Section 4.2 finishes discussing how this result is ultimately used.

Suppose $\mathbf{Y}$ is the set of considered variables. The coupled influence of a set $J \subseteq \mathbf{Y}$ on a target X as

$$ {}^{X}_{J}M_{\mathbf{Y}} = \left({}^{X}_{J}I_{\mathbf{Y}}\right) - \sum_{j \subset J}^{|j|>0} {}^{X}_{j}M_{\mathbf{Y}} \tag{A.1} $$

where $\left({}^{X}_{J}I_{\mathbf{Y}}\right) = I(X; J|\mathbf{Y} \setminus J)$ is introduced as a convenient shorthand. Note that the notation leaves conditioning on the remaining drivers in $\mathbf{Y}$ implied, which will be absent in the case of $J = \mathbf{Y}$. Below, I prove by induction that

$$ {}^{X}_{J}M_{\mathbf{Y}} = \sum_{j \subseteq J}^{|j|>0} (-1)^{|J|-|j|} \left({}^{X}_{j}I_{\mathbf{Y}}\right). \tag{A.2} $$

To start, for $J \subseteq \mathbf{Y}$ such that $|J| = 1$,

$$ {}^{X}_{J}M_{\mathbf{Y}} = \left({}^{X}_{J}I_{\mathbf{Y}}\right) = \sum_{j \subseteq J}^{|j|>0} (-1)^{|J|-|j|} \left({}^{X}_{j}I_{\mathbf{Y}}\right). \tag{A.3} $$

For $J \subseteq \mathbf{Y}$ such that $|J| = 2$, then

$$ {}^{X}_{J}M_{\mathbf{Y}} = \left({}^{X}_{J}I_{\mathbf{Y}}\right) - \sum_{j \subset J}^{|j|>0} {}^{X}_{j}M_{\mathbf{Y}} = \left({}^{X}_{J}I_{\mathbf{Y}}\right) - \sum_{j \subset J}^{|j|>0} \left({}^{X}_{j}I_{\mathbf{Y}}\right) = \sum_{j \subseteq J}^{|j|>0} (-1)^{|J|-|j|} \left({}^{X}_{j}I_{\mathbf{Y}}\right). \tag{A.4} $$

Now, suppose that, for all $K \subset \mathbf{Y}$ such that $|K| \leq \kappa$,

$$ {}^{X}_{K}M_{\mathbf{Y}} = \sum_{k \subseteq K}^{|k|>0} (-1)^{|K|-|k|} \left({}^{X}_{k}I_{\mathbf{Y}}\right). \tag{A.5} $$

Then, for $K \subseteq \mathbf{Y}$ with $|K| = \kappa + 1$,

$$ {}^{X}_{K}M_{\mathbf{Y}} = \left({}^{X}_{K}I_{\mathbf{Y}}\right) - \sum_{k \subset K}^{|k|>0} {}^{X}_{k}M_{\mathbf{Y}} \tag{A.6} $$

$$ = \left({}^{X}_{K}I_{\mathbf{Y}}\right) - \sum_{k \subset K}^{|k|>0} \sum_{k' \subseteq k}^{|k'|>0} (-1)^{|k|-|k'|} \left({}^{X}_{k'}I_{\mathbf{Y}}\right). \tag{A.7} $$

I will rewrite the double summation. First, consider the summation of the influences of all sets $k \subset K$ such that $|k| = |K| - 1 = \kappa$. For each k, there is $1 = \binom{|K|-|k|}{(|K|-1)-|k|}$ term which yields a positive $\left({}^{X}_{k}I_{\mathbf{Y}}\right)$. But, for each $k' \subset K$ such that $|k'| = |K| - 2 = \kappa - 1$, there are $2 = \binom{|K|-|k'|}{(|K|-1)-|k'|}$ sets k that contain k', meaning that summing all ${}^{X}_{k}M_{\mathbf{Y}}$ yields $-2\left({}^{X}_{k'}I_{\mathbf{Y}}\right)$. It should be clear that summing all ${}^{X}_{k}M_{\mathbf{Y}}$ where $k \subset K$ and $|k| = |K| - 1 = \kappa$ will yield $(-1)^{(|K|-1)-|k'|}\binom{|K|-|k'|}{(|K|-1)-|k'|}\left({}^{X}_{k'}I_{\mathbf{Y}}\right)$ for all $k' \subset K$.

Next, consider the summation of the influences of all sets $k \subset K$ such that $|k| = |K| - 2 = \kappa - 1$. There is $1 = \binom{|K|-|k|}{(|K|-2)-|k|}$ term which yields a positive $\left({}^{X}_{k}I_{\mathbf{Y}}\right)$. For each $k' \subset K$ such that $|k'| = |K| - 3 = \kappa - 2$, there are $3 = \binom{|K|-|k'|}{(|K|-2)-|k'|}$ sets k that contain k'. In general, there are $\binom{|K|-|k'|}{(|K|-2)-|k'|}$ sets $k \subset K$ that contain each set $k' \subset K$ with $|k'| \leq |k| = |K| - 2 = \kappa - 1$.

Finally, by continuing this process, we have that, for some $0 < i < |K| = \kappa + 1$, summing all ${}^{X}_{k}M_{\mathbf{Y}}$ where $k \subset K$ and $|k| = i$ yields $(-1)^{i-|k'|}\binom{|K|-|k'|}{i-|k'|}\left({}^{X}_{k'}I_{\mathbf{Y}}\right)$ for each $k' \subset K$ such that $|k'| \leq i$. Then, by collecting similar terms across all values of i,

$$ \sum_{k \subset K}^{|k|>0} \sum_{k' \subseteq k}^{|k'|>0} (-1)^{|k|-|k'|} \left({}^{X}_{k'}I_{\mathbf{Y}}\right) = \sum_{k \subset K}^{|k|>0} \left({}^{X}_{k}I_{\mathbf{Y}}\right) \left(\sum_{i=|k|}^{|K|-1} (-1)^{i-|k|} \binom{|K|-|k|}{i-|k|} \right). \tag{A.8} $$

We can rewrite the inner summation as $\sum_{i=0}^{(|K|-1)-|k|} (-1)^{i}\binom{|K|-|k|}{i}$. Lemma 1 in Aupetit (2009) proved that $\sum_{i=0}^{n}(-1)^{i}\binom{n}{i} = 0$. Instead, we have

$$ \sum_{i=0}^{(|K|-1)-|k|} (-1)^{i} \binom{|K|-|k|}{i} = \sum_{i=0}^{|K|-|k|} (-1)^{i} \binom{|K|-|k|}{i} - (-1)^{|K|-|k|} \binom{|K|-|k|}{|K|-|k|} = 0 - (-1)^{|K|-|k|}. \tag{A.9} $$

Substituting this into equation (A.8) yields

$$\sum_{k\subset K}^{|k|>0}\binom{X}{k}I_{\mathbf{Y}}\left(\sum_{i=|k|}^{|K|-1}(-1)^{i-|k|}\binom{|K|-|k|}{i-|k|}\right)=-\sum_{k\subset K}^{|k|>0}\binom{X}{k}I_{\mathbf{Y}}(-1)^{|K|-|k|}. \tag{A.10}$$

Finally, equation (A.6) becomes

$$\underset{K}{{}^{X}M}_{\mathbf{Y}}=\binom{X}{K}I_{\mathbf{Y}}+\sum_{k\subset K}^{|k|>0}(-1)^{|K|-|k|}\binom{X}{k}I_{\mathbf{Y}} \tag{A.11}$$

$$=\sum_{k\subseteq K}^{|k|>0}(-1)^{|K|-|k|}\binom{X}{k}I_{\mathbf{Y}}, \tag{A.12}$$

meaning that the supposition for all sets with κ processes implies the supposition for a set with $\kappa+1$ processes.

Since the supposition was true for all sets with 1 or 2 processes, by the principal of mathematical induction, for all $J\subseteq\mathbf{Y}$,

$$\underset{J}{{}^{X}M}_{\mathbf{Y}}=\sum_{j\subseteq J}^{|j|>0}(-1)^{|J|-|j|}\binom{X}{j}I_{\mathbf{Y}}. \tag{A.13}$$

A.3 Rewriting $m$link Influence as Summation of Mutual Informations

Writing $m$link influences in terms of mutual informations allowed for the original implementation for calculating them, as stated in section 4.3. This was before the implementation calculated the coupled influence terms, which are now used to calculate $m$link influences. The result at the end, however, is used in the next result, which is critical to the framework as a whole.

Let $\mathbf{Y}$ be the set of all considered drivers, with $|\mathbf{Y}|=n$. The $m$link influence of driver $Y\in\mathbf{Y}$ on target X is

$$(Y\to X)_m=\frac{1}{m}\sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{|J|=m-1}\underset{JY}{{}^{X}M}_{\mathbf{Y}}. \tag{A.14}$$

Using the previous result, this is

$$m\,(Y\to X)_m=\sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{|J|=m-1}\left[\sum_{j\subseteq(J\cup\{Y\})}^{|j|>0}(-1)^{(|J|+1)-|j|}\binom{X}{j}I_{\mathbf{Y}}\right]. \tag{A.15}$$

To find J, note Y is already in the active set. So, $m-1$ other variables are chosen from $n-1$ other variables. Thus, in the inner summation of equation (A.15), there is $1=\binom{(n-1)-(m-1)}{(m-1)-(m-1)}=\binom{n-(|J|+1)}{m-(|J|+1)}$ set that can contribute a positive $\binom{X}{JY}I_{\mathbf{Y}}$ and a negative $\binom{X}{J}I_{\mathbf{Y}}$. For each $j\subset(\mathbf{Y}\setminus\{Y\})$ such that $|j|=m-2$,

there are $\binom{(n-1)-(m-2)}{(m-1)-(m-2)} = \binom{n-(|j|+1)}{m-(|j|+1)}$ sets J that contain j, so the inner summation yields $\binom{n-(|j|+1)}{m-(|j|+1)}$ copies of negative $\left(\begin{smallmatrix}X\\ j\cup\{Y\}\end{smallmatrix} I_{\mathbf{Y}}\right)$ and $\binom{n-|j|-1}{m-|j|-1}$ copies of positive $\left(\begin{smallmatrix}X\\ j\end{smallmatrix} I_{\mathbf{Y}}\right)$. This pattern continues until $|j|=0$, i.e. j is empty. In this case, the inner summation yields $\binom{n-1}{m-1}$ copies of $(-1)^{m-1}\left(\begin{smallmatrix}X\\ Y\end{smallmatrix} I_{\mathbf{Y}}\right)$.

Therefore,

$$m(Y\to X)_m = (-1)^{m-1}\binom{n-1}{m-1}\left(\begin{smallmatrix}X\\ Y\end{smallmatrix} I_{\mathbf{Y}}\right) + \sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{0<|J|<m} (-1)^{m-(|J|+1)}\binom{n-|J|-1}{m-|J|-1}\left[\left(\begin{smallmatrix}X\\ JY\end{smallmatrix} I_{\mathbf{Y}}\right) - \left(\begin{smallmatrix}X\\ J\end{smallmatrix} I_{\mathbf{Y}}\right)\right]. \tag{A.16}$$

A.4 Total Influence is Nonnegative

The need for total influence to be nonnegative is discussed briefly in section 4.3. One of the first applications, before even explicitly storing the coupled influence terms, was studying the time series data from Christman field. It was a really long time series, so we were certain that the estimators would be accurate. The results came the day before I presented them, but they contained negative total influences. Unfortunately, this seemed to be what generated the most discussion after my presentation. At that time, I could only intuit that total influence was nonnegative.

To prove this, start from result from the $m$links. Then, I develop a coefficient for each mutual information in the summation that is not expressed as a summation. This is then combined with an inequality of mutual informations at the end to prove the measure is indeed nonnegative.

The total influence of a driver $Y\in\mathbf{Y}$ on target X is the sum of all $m$links from Y to X. That is,

$$(Y\to X)_{tot} = \sum_{m=1}^{n}(Y\to X)_m\,, \tag{A.17}$$

where the $m$links have already been normalized by m. Using the previous expression and then collecting terms, this becomes

$$\begin{aligned}(Y\to X)_{tot} = &\sum_{m=1}^{n}\frac{1}{m}\sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{|J|<m}(-1)^{m-(|J|+1)}\binom{n-(|J|+1)}{m-(|J|+1)}\left(\begin{smallmatrix}X\\ JY\end{smallmatrix} I_{\mathbf{Y}}\right)\\ &+\sum_{m=1}^{n}\frac{1}{m}\sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{0<|J|<m}(-1)^{m-|J|}\binom{n-|J|-1}{m-|J|-1}\left(\begin{smallmatrix}X\\ J\end{smallmatrix} I_{\mathbf{Y}}\right) &\text{(A.18a)}\\ = &\sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})} A_{|J|}^{|\mathbf{Y}|}\left(\begin{smallmatrix}X\\ JY\end{smallmatrix} I_{\mathbf{Y}}\right) + \sum_{J\subseteq(\mathbf{Y}\setminus\{Y\})}^{|J|\neq 0} B_{|J|}^{|\mathbf{Y}|}\left(\begin{smallmatrix}X\\ J\end{smallmatrix} I_{\mathbf{Y}}\right), &\text{(A.18b)}\end{aligned}$$

where $A^{|\mathbf{Y}|}_{|J|} = \sum_{m=|J|+1}^{|\mathbf{Y}|} \frac{(-1)^{m-(|J|+1)}}{m}\binom{|\mathbf{Y}|-(|J|+1)}{m-(|J|+1)}$ and $B^{|\mathbf{Y}|}_{|J|} = -A^{|\mathbf{Y}|}_{|J|}$. We need to show $A^{|\mathbf{Y}|}_{|J|} \geq 0$.

For the sake of brevity, I use $k = |J| \geq 0$ and $n = |\mathbf{Y}|$, so

$$A^n_k = \sum_{m=k+1}^{n} \frac{(-1)^{m-(k+1)}}{m}\binom{n-(k+1)}{m-(k+1)}. \tag{A.19}$$

A well known result is

$$\sum_{m=1}^{n}\frac{1}{m} = \sum_{m=1}^{n}\frac{(-1)^{m-1}}{m}\binom{n}{m}. \tag{A.20}$$

Thus, for $k = 0$ and any $n \in \mathbb{N}$,

$$\begin{aligned}
A^n_k =& \sum_{m=1}^{n}\frac{(-1)^{m-1}}{m}\binom{n-1}{m-1} && \text{(A.21)}\\
=& \frac{(-1)^{n-1}}{n} + \sum_{m=1}^{n-1}\frac{(-1)^{m-1}}{m}\binom{n-1}{m-1} && \text{(A.22)}\\
=& \frac{(-1)^{n-1}}{n} + \sum_{m=1}^{n-1}\frac{(-1)^{m-1}}{m}\left[\binom{n}{m}-\binom{n-1}{m}\right] && \text{(A.23)}\\
=& \sum_{m=1}^{n}\frac{(-1)^{m-1}}{m}\binom{n}{m} - \sum_{m=1}^{n-1}\frac{(-1)^{m-1}}{m}\binom{n-1}{m} && \text{(A.24)}\\
=& \sum_{m=1}^{n}\frac{1}{m} - \sum_{m=1}^{n-1}\frac{1}{m} && \text{(A.25)}\\
=& \frac{1}{n} && \text{(A.26)}\\
=& \left[(k+1)\binom{n}{k+1}\right]^{-1}. && \text{(A.27)}
\end{aligned}$$

Now, assume $A^n_{k'} = \sum_{m=k'+1}^{n} \frac{(-1)^{m-(k'+1)}}{m}\binom{n-(k'+1)}{m-(k'+1)} = \left[(k'+1)\binom{n}{k'+1}\right]^{-1}$ for $0 \leq k' \leq k < n-1$. Again, this is for any $n \geq 1$. Then,

$$\begin{aligned}
A^n_{k+1} =& \sum_{m=k+2}^{n}\frac{(-1)^{m-(k+2)}}{m}\binom{n-(k+2)}{m-(k+2)} && \text{(A.28)}\\
=& \frac{(-1)^{n-(k+2)}}{n} + \sum_{m=k+2}^{n-1}\frac{(-1)^{m-(k+2)}}{m}\binom{n-(k+2)}{m-(k+2)} && \text{(A.29)}\\
=& -\frac{(-1)^{n-(k+1)}}{n} - \sum_{m=k+2}^{n-1}\frac{(-1)^{m-(k+1)}}{m}\binom{n-(k+2)}{m-(k+2)} && \text{(A.30)}\\
=& -\frac{(-1)^{n-(k+1)}}{n} - \sum_{m=k+2}^{n-1}\frac{(-1)^{m-(k+1)}}{m}\left[\binom{n-(k+1)}{m-(k+1)}-\binom{n-(k+2)}{m-(k+1)}\right] && \text{(A.31)}
\end{aligned}$$

From here, the lower bound in the summation is decreased by 1 to $m = k+1$. When this value is used in the summation, the bottom part of each combination term will be zero. This makes the value of each combination term 1, so their difference is 0. Thus, the value of the summation does not change. So,

$$A^n_{k+1} = -\frac{(-1)^{n-(k+1)}}{n} - \sum_{m=k+1}^{n-1} \frac{(-1)^{m-(k+1)}}{m}\left[\binom{n-(k+1)}{m-(k+1)} - \binom{(n-1)-(k-1)}{m-(k+1)}\right] \tag{A.32}$$

$$= -\sum_{m=k+1}^{n} \frac{(-1)^{m-(k+1)}}{m}\binom{n-(k+1)}{m-(k+1)} + \sum_{m=k+1}^{n-1} \frac{(-1)^{m-(k+1)}}{m}\binom{(n-1)-(k+1)}{m-(k+1)} \tag{A.33}$$

$$= -A^n_k + A^{n-1}_k \tag{A.34}$$

$$= -\left[(k+1)\binom{n}{k+1}\right]^{-1} + \left[(k+1)\binom{n-1}{k+1}\right]^{-1} \tag{A.35}$$

$$= \left[(k+2)\binom{n}{k+2}\right]^{-1}, \tag{A.36}$$

where the final equality comes from a bit of algebra. By the principal of mathematical induction,

$$A^n_k = \sum_{m=k+1}^{n} \frac{(-1)^{m-(k+1)}}{m}\binom{n-(k+1)}{m-(k+1)} = \left[(k+1)\binom{n}{k+1}\right]^{-1} > 0, \text{ for all } n \in \mathbb{N} \text{ and } 0 \le k < n. \tag{A.37}$$

To prove the total influence is nonnegative, an important relationship is that, for any processes X, Y, Z, W (where W may be empty),

$$I(X;Y,Z|W) = I(X;Y|Z,W) + I(X;Z|W) \ge I(X;Y|Z,W). \tag{A.38}$$

So, for process $Y \in \mathbf{Y}$ and any nonempty subset $J \subseteq \mathbf{Y} \setminus \{Y\}$, $\binom{X}{JY}I_{\mathbf{Y}}\big) \ge \binom{X}{J}I_{\mathbf{Y}}\big)$. Rewriting (A.18) shows

$$(Y \to X)_{tot} = \frac{1}{n}\binom{X}{Y}I_{\mathbf{Y}}\big) + \sum_{J \subseteq (\mathbf{Y}\setminus\{Y\})}^{|J|\neq 0}\left[(|J|+1)\binom{n}{|J|+1}\right]^{-1}\left[\binom{X}{JY}I_{\mathbf{Y}}\big) - \binom{X}{J}I_{\mathbf{Y}}\big)\right] \ge 0, \tag{A.39}$$

as desired.

A.5 Coupled Influence is Interaction Information

The development in this section shows that coupled influence is a version of interaction information. Interaction information has been researched extensively since McGill (1954) introduced it. Thus, with this proof, we are able to use and even add to this already rich body of research, as demonstrated in section 3.2.1.

Assume that we have a target X and a set containing at least 2 drivers $\mathbf{Y} = \{Y, Z, \dots\}$. Then, following equation (2.7),

$$
\begin{aligned}
{}^{X}_{YZ}M_{\mathbf{Y}} &= I(X; Y, Z|\mathbf{Y}\setminus\{Y,Z\}) - I(X; Z|\mathbf{Y}\setminus\{Z\}) - I(X; Y|\mathbf{Y}\setminus\{Y\}) && \text{(A.40)}\\
&= I(X; Y|\mathbf{Y}\setminus\{Y,Z\}) - I(X; Y|\mathbf{Y}\setminus\{Z\}) && \text{(A.41)}\\
&= II(X, Y, Z|\mathbf{Y}\setminus\{Y,Z\}), && \text{(A.42)}
\end{aligned}
$$

which is the interaction information between X, Y, Z given all other drivers. Note that $I(X; Y|\mathbf{Y}\setminus\{Y,Z\})$ excludes both Y and Z from the condition, while Z is not active. We can rewrite this term as ${}^{X}_{Y}M_{\mathbf{Y}\setminus\{Z\}}$, so

$$
{}^{X}_{YZ}M_{\mathbf{Y}} = {}^{X}_{Y}M_{\mathbf{Y}\setminus\{Z\}} - {}^{X}_{Y}M_{\mathbf{Y}}. \tag{A.43}
$$

I will prove a similar relation for all coupled influences.

Now, for some $K \subseteq \mathbf{Y}$ such that $|K| \geq 2$, suppose that for any nonempty $k \subset K$ and any process $W \in \mathbf{Y}\setminus k$,

$$
{}^{X}_{kW}M_{\mathbf{Y}} = {}^{X}_{k}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{k}M_{\mathbf{Y}}. \tag{A.44}
$$

This implies also that, for any partition of $K = k \cup \{W\}$, ${}^{X}_{K}M_{\mathbf{Y}} = {}^{X}_{k}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{k}M_{\mathbf{Y}}$. Then,

$$
\begin{aligned}
{}^{X}_{KW}M_{\mathbf{Y}} &= \left({}^{X}_{KW}I_{\mathbf{Y}}\right) - \sum_{k\subset(K\cup\{W\})}^{|k|>0} \left({}^{X}_{k}M_{\mathbf{Y}}\right) && \text{(A.45)}\\
&= \left({}^{X}_{KW}I_{\mathbf{Y}}\right) - \left({}^{X}_{W}I_{\mathbf{Y}}\right) - \sum_{k\subset K}^{|k|>0} \left({}^{X}_{kW}M_{\mathbf{Y}} + {}^{X}_{k}M_{\mathbf{Y}}\right) - {}^{X}_{K}M_{\mathbf{Y}} && \text{(A.46)}\\
&= \left({}^{X}_{K}I_{\mathbf{Y}\setminus\{W\}}\right) - \sum_{k\subset K}^{|k|>0} \left({}^{X}_{k}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{k}M_{\mathbf{Y}} + {}^{X}_{k}M_{\mathbf{Y}}\right) - {}^{X}_{K}M_{\mathbf{Y}} && \text{(A.47)}\\
&= \left({}^{X}_{K}I_{\mathbf{Y}\setminus\{W\}}\right) - \sum_{k\subset K}^{|k|>0} \left({}^{X}_{k}M_{\mathbf{Y}\setminus\{W\}}\right) - {}^{X}_{K}M_{\mathbf{Y}} && \text{(A.48)}\\
&= {}^{X}_{K}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{K}M_{\mathbf{Y}}. && \text{(A.49)}
\end{aligned}
$$

Thus, the supposition holds when joining one process $W \in \mathbf{Y}\setminus K$ to K.

By the principle of mathematical induction, for all $J \subset \mathbf{Y}$ and any $Z \in \mathbf{Y}\setminus J$,

$$
{}^{X}_{JZ}M_{\mathbf{Y}} = {}^{X}_{J}M_{\mathbf{Y}\setminus\{Z\}} - {}^{X}_{J}M_{\mathbf{Y}}. \tag{A.50}
$$

The definition of n-variable interaction information follows a similar recursive definition, e.g. for 4 variables $I(X,Y,Z,W) = I(X,Y,Z) - I(X,Y,Z|W)$. We may arbitrarily add conditions as we add drivers to the study. In our case, we now require at least 3 drivers $\mathbf{Y} = \{Y,Z,W,\dots\}$. Since we already showed the two-driver coupled influence is a three-variable interaction information, it is easy to show that

$$
\begin{aligned}
{}^{X}_{YZW}M_{\mathbf{Y}} &= {}^{X}_{YZ}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{YZ}M_{\mathbf{Y}} && \text{(A.51)}\\
&= II(X,Y,Z|\mathbf{Y}\setminus\{Y,Z,W\}) - II(X,Y,Z|\mathbf{Y}\setminus\{Y,Z\}) && \text{(A.52)}\\
&= II(X,Y,Z,W|\mathbf{Y}\setminus\{Y,Z,W\}). && \text{(A.53)}
\end{aligned}
$$

Now, assume that, for some nonempty $K \subset \mathbf{Y}$,

$$
{}^{X}_{K}M_{\mathbf{Y}} = II(X,K|\mathbf{Y}\setminus K). \tag{A.54}
$$

Note that, as in the 2link case, if any driver not in K is removed from $\mathbf{Y}$, this merely redefines the universal set of drivers, so the above equation holds. Then, for any process $W \in \mathbf{Y}\setminus K$,

$$
\begin{aligned}
{}^{X}_{KW}M_{\mathbf{Y}} &= {}^{X}_{K}M_{\mathbf{Y}\setminus\{W\}} - {}^{X}_{K}M_{\mathbf{Y}} && \text{(A.55)}\\
&= II(X,K|\mathbf{Y}\setminus(K\cup\{W\})) - II(X,K|\mathbf{Y}\setminus K) && \text{(A.56)}\\
&= II(X,K,W|\mathbf{Y}\setminus(K\cup\{W\})). && \text{(A.57)}
\end{aligned}
$$

Thus, even by joining one process W to K, the assumption still holds.

By the principle of mathematical induction, for any $J \subseteq \mathbf{Y}$,

$$
{}^{X}_{J}M_{\mathbf{Y}} = II(X,J|\mathbf{Y}\setminus J). \tag{A.58}
$$