THESIS

HEATWAVES BEEN FAKING ME OUT: EVALUATING 2-M TEMPERATURE FORECAST ERRORS WHEN AI WEATHER PREDICTION MODELS CAN'T CATCH THE HEAT

Submitted by

Kelsey E. Ennis

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2025

Master's Committee:

Advisor: Elizabeth Barnes
Co-Advisor: Eric Maloney

Brooke Anderson

ABSTRACT

HEATWAVES BEEN FAKING ME OUT: EVALUATING 2-M TEMPERATURE FORECAST ERRORS WHEN AI WEATHER PREDICTION MODELS CAN'T CATCH THE HEAT

Extreme heat is the deadliest weather-related hazard in the United States. Furthermore, it is also increasing in intensity, frequency, and duration, making skillful forecasts vital to protecting life and property. Traditional numerical weather prediction (NWP) models struggle with extreme heat for medium-range and subseasonal-to-seasonal (S2S) timescales. Meanwhile, artificial intelligence-based weather prediction (AIWP) models are progressing rapidly. However, it is largely unknown how well AIWP models forecast extremes, especially for medium-range and S2S timescales. This study investigates 2-m temperature forecasts for 60 heat waves across the four boreal seasons and over four CONUS regions at lead times up to 20 days, using two AIWP models (Google GraphCast and Pangu-Weather) and one traditional NWP model (NOAA United Forecast System Global Ensemble Forecast System (UFS GEFS)). First, case study analyses show that both AIWP models and the UFS GEFS exhibit consistent cold biases on regional scales in the 5–10 days of lead time before heat wave onset. GraphCast is the more skillful AIWP model, outperforming UFS GEFS and Pangu-Weather in most locations. Next, the two AIWP models are isolated and analyzed across all heat waves and seasons, with events split between models' testing (2018–2023) and training (1979–2017) periods. There are cold biases before and during the heat waves in both models and all seasons, except Pangu-Weather in winter, which exhibits a mean warm bias before heat wave onset. Overall, results

offer encouragement that AIWP models may be useful for medium-range and S2S prediction of extreme heat.

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to my advisors Dr. Elizabeth Barnes and Dr. Eric Maloney for their unwavering support and outstanding mentorship throughout my process of the Master's degree program at Colorado State University. I am thankful to Dr. Brooke Anderson for serving on my graduate committee as well. I would also like to thank my co-authors Dr. Marybeth Arcodia and Dr. Martin Fernandez for their insightful feedback and continued guidance at every stage of this project.

I also would like to express my sincere appreciation towards the members of the Barnes Research Group. The collaborative spirit and supportive work environment that the group promoted made each day at work productive and enjoyable. Thank you to my former research advisor Dr. Shawn Milrad for providing valuable feedback throughout the writing process of this project. Finally thank you to all of my friends, family, and mentors who have supported me throughout my graduate career at Colorado State.

To my parents, thank you for always encouraging my dreams and for being the most influential role models in my life. To my two younger sisters, thanks for being my biggest cheerleaders. This research was supported by a NOAA Office of Atmospheric Research (OAR) grant (NA22OAR4310621) and a Heising-Simons Foundation grant (#2023-4720).

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Extreme Heat Impacts

Extreme heat is the deadliest weather-related hazard in the US. The National Oceanic and Atmospheric Administration (NOAA) reports that in 2023 (the most recent year of complete data), 555 deaths could be directly attributed to extreme heat (Fig. 1). Furthermore, the 10-year and 30-year averages in heat fatalities are more than double the next highest weather-related cause (floods; Fig. 1). However, directly attributed fatalities often do not tell the full story, as most heat-related deaths go unreported or misattributed. Research shows that the average number of excess heat-related deaths per year in the US may be up to ten times the directly attributable value in 2023 (e.g., Weinberger et al. 2021). Other estimates suggest six heat-related deaths per 100,000 North American residents each year (Zhao et al. 2021). Beyond excess mortality, extreme heat and heat stress are associated with many adverse health impacts such as dehydration, heat exhaustion, and heat stroke. Furthermore, heat stress exacerbates underlying conditions like cardiovascular disease, respiratory ailments, and diabetes (e.g., WHO 2024).

Excessive heat can also damage property and industry, particularly heat-sensitive infrastructure, energy, water supply, agriculture, and forestry. The energy grid is often stressed to its limit during extreme heat events, with elevated humidity levels exacerbating the issue (e.g., Rastogi et al. 2021). During excessive heat in the Western US, temporary power shut offs and brownouts are often necessary to prevent sparking wildfires (e.g., USGCRP 2023). In regions

without much heat-resistant infrastructure, extreme heat can damage roadways, railways, and airport runways. For example, during the 2021 US Pacific Northwest heat wave, glass shattered, and roadways buckled in Portland, Oregon with temperatures reaching around 45 °C (Thompson et al. 2022).

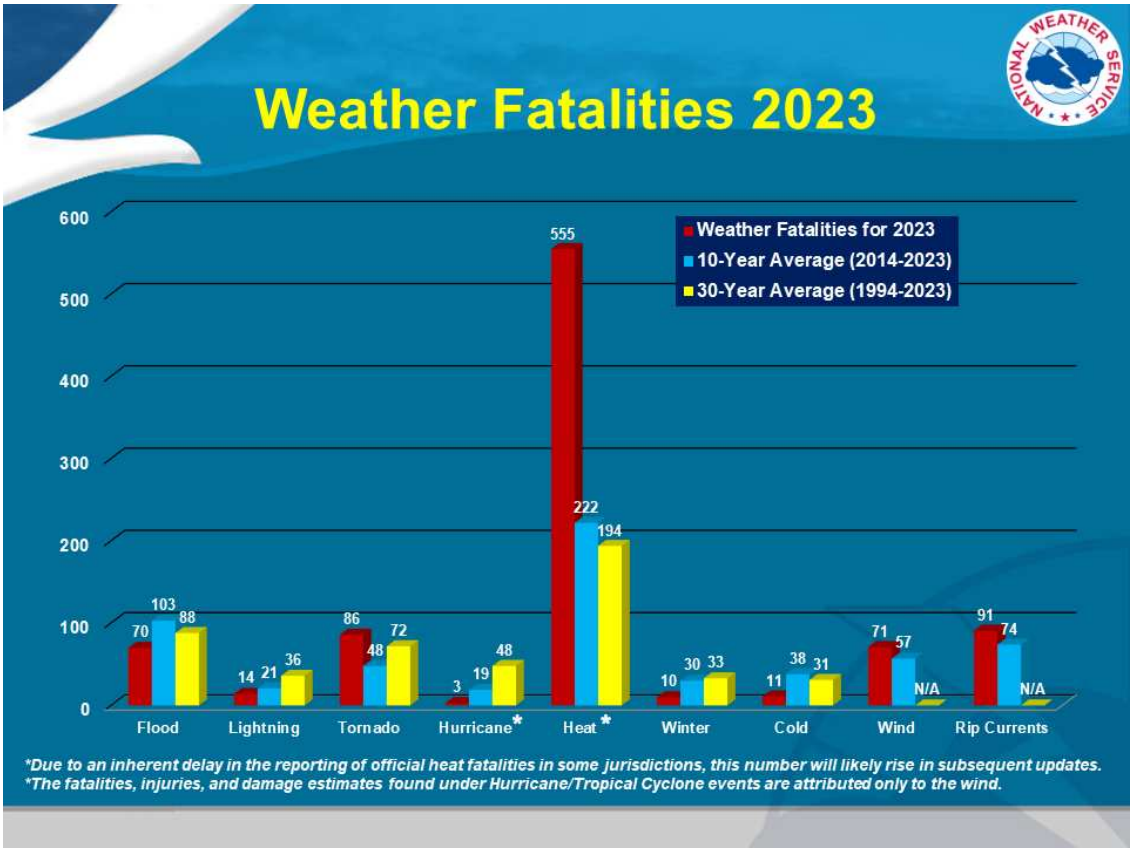

Figure 1.1: From NWS (2025): 2023 weather-related fatalities (red bars) directly attributed to specific hazards; 10-year (blue bars) and 30-year (yellow bars) averages are also shown.

Extreme heat disproportionately affects vulnerable individuals (Alizadeh et al. 2022), primarily because of their smaller adaptive capacity. Here, adaptive capacity refers to the ability of people to adjust to the health impacts of extreme heat, find mitigation outlets, and/or cope with extreme heat's consequences (Wu et al. 2022). For example, groups with small adaptive capacities often lack consistent access to cooling, reliable health care, and public transportation.

Other groups with small adaptive capacities may include older people who live in hot and/or humid climates such as the Southeast US (Ennis and Milrad 2024). Extreme heat also impacts industries such as agriculture and construction, where outdoor work is required and protective regulations are often sparse (e.g., Bitencourt et al. 2021; McAllister et al. 2022). In turn, regional and national economies can be greatly impacted. For example, García-León et al. (2021) report an average GDP loss of 0.3 to 0.5% for five European countries over several extreme heat summers since 2010, while the Fifth US National Climate Assessment finds that heat waves are one of the consistently costliest weather-related hazards to the US economy (USGCRP 2023).

Heat waves can also exacerbate drought conditions and water shortages, particularly in arid regions (IPCC 2021; USGCRP 2023). A recent study for Mexico (Sutanto et al. 2024) finds that agricultural yields decrease by 25% during extreme drought events, but by 44% during compound and/or cascading droughts and heat waves. Extreme heat can also increase heat stress on livestock (e.g., Thornton et al. 2022), as well as damage plants and forests by reducing photosynthesis and stunting growth (e.g., Teskey et al. 2015). Overall, the increases in extreme heat and associated heat waves pose enormous risks to life, health, ecosystems, property, industry, and economies. As such, improved forecasts and alert lead times of these events should be a priority.

1.2 Extreme Heat Trends

Extreme heat and associated heat waves are becoming more frequent, intense, and longer (e.g., Perkins 2015; Mora et al. 2017; Raymond et al. 2020; Bekris et al. 2023), largely due to climate change (IPCC 2021). To that end, Vicedo-Cabrera et al. (2021) attribute 37% of heat-related deaths to anthropogenic warming during the period 1991-2018. Furthermore, some studies (e.g., Raymond et al. 2020) report that a wet bulb temperature of 35°C, which they cite as

the limit for human tolerance, is occasionally exceeded in certain tropical and subtropical areas such as India and the Middle East. In addition, climate projections show that the 35 °C threshold may be exceeded more frequently and for longer durations in the future (Raymond et al. 2020; Matthews et al. 2025).

Heat waves are also becoming more common in locations not acclimated to them, particularly mid- and high-latitude continental locations (e.g., Patterson 2023; Milrad and Ennis 2025). For example, during the extreme summer of 2022, temperatures hit an unprecedented 40 °C in the UK (Zachariah et al. 2022). The epitome of these trends is the aforementioned June 2021 heat wave in the US Pacific Northwest and southwestern Canada. During that event, temperatures reached 50 °C at 50 °N latitude, making it one of the most anomalous weather events on record (Schumacher et al. 2022; Thompson et al. 2022; White et al. 2023).

Highly anomalous heat waves in places not used to them can exacerbate adverse health impacts and excess mortality, in large part due to the lack of cooling infrastructure within those regions. For example, Ballester et al. (2023) estimate more than 60,000 excess deaths in Europe during the extreme heat summer of 2022, in no small part due to limited cooling infrastructure throughout many countries including the UK. Vogel et al. (2023) attribute 441 deaths to the June 2021 Pacific Northwest heatwave in Washington State alone, with most occurring in the Seattle metropolitan area. In greater Seattle, two-thirds of households with incomes less than $50,000 per year and 70% of all rented homes have no air conditioning, which decreases adaptive capacity and directly contributes to adverse health and mortality impacts during extreme events such as the 2021 heat wave (Vogel et al. 2023).

1.3 Extreme Heat Mechanisms

Heat waves are typically caused by a combination of several mechanisms, including atmospheric blocking, land-atmosphere processes, ocean-atmosphere coupling, and topographical forcing (Jiménez-Esteve and Domeisen 2022; Barriopedro et al. 2023; Domeisen et al. 2023). Atmospheric blocking, in part characterized by persistent and anomalous mid-upper tropospheric ridging, is particularly important for mid- and high-latitude heat waves (Jiménez-Esteve and Domeisen 2022). In such regions, blocking patterns as well as the associated horizontal warm advection and subsidence are the primary cause of heat extremes (Jiménez-Esteve and Domeisen 2022). Furthermore, because blocking patterns are so persistent, they can dramatically affect the duration of heat extremes, which in turn exacerbates impacts.

Although we understand the role that atmospheric blocking plays in forcing extreme heat and weather events, the onset, decay, and predictability of blocks are not as well understood (Nakamura and Huang 2018; Barriopedro et al. 2023). Blocking patterns are often associated with anomalous and persistent meandering of the jet stream (Nakamura and Huang 2018). Some research proposes an analogy of blocking patterns to “traffic jams”, where blocking is the traffic jam, and the highway is the jet stream (Nakamura and Huang 2018; Paradise et al. 2019). These traffic jams (i.e., blocks) consistently occur in regions of minimum jet stream flow capacity, thereby causing a slowdown in the jet stream and persistent mid-upper tropospheric ridging when that capacity is exceeded (Nakamura and Huang 2018; Paradise et al. 2019; Yan et al. 2024).

Numerical weather prediction (NWP) models traditionally struggle predicting atmospheric blocking, especially on longer timescales (e.g., Davini et al. 2021). Furthermore, there is evidence that blocking patterns may become more common in a warming world (Kornhuber et

al. 2020), resulting in more frequent temperature and precipitation extremes. In sum, there is a need for improved predictability of atmospheric blocks and associated heat extremes (e.g., Paradise et al. 2019; Barriopedro et al. 2023), using both traditional NWP and novel approaches.

1.4 Artificial Intelligence Based Weather Prediction (AIWP) Models

Overall, medium-range and subseasonal-to-seasonal (S2S) model forecasts of extreme heat events need improvement. Recently, artificial intelligence-based weather prediction (AIWP) models have been rapidly gaining in usage and popularity across the weather-climate enterprise. AIWP models are already beneficial to weather and climate forecasts; for example, the European Centre for Medium-range Weather Forecasts (ECMWF) currently has an operational AIWP (ECMWF 2025). These models are advantageous largely because they run exponentially faster than traditional physics-based NWP models and at a fraction of the computational cost (e.g., Bouallègue et al. 2024; Radford et al. 2025). In other words, AIWP developers and end-users take advantage of the small computational cost, based on the assumption that the AIWP algorithms can effectively learn complex and nonlinear patterns from large training datasets (e.g., ECMWF ERA5 reanalysis; Hersbach et al. 2020; ECMWF 2025).

Despite their rapid development and promise, AIWP models still have clear limitations, as shown by a recent study comparing Pangu-Weather, Google GraphCast, and FourCastNet models (Bonavita 2024). First, they are not able to represent physical processes such as atmospheric dynamics and thermodynamics (Selz and Craig 2023; Bonavita 2024). Second, their grid spacing is typically coarse compared to traditional NWP models, and therefore they struggle to reproduce sub-synoptic and mesoscale atmospheric phenomena such as convection and localized extremes in e.g., temperature (Bonavita 2024). Third, they lack the physical consistency of traditional NWP models, which impacts model interpretation and skill (Bonavita 2024).

Another limitation of AIWP models is that more verification studies are needed, particularly with respect to forecasts of extreme events such as heat waves, heavy precipitation, and tropical cyclones (Xie et al. 2024; Camps-Valls et al. 2025; Pasche et al. 2025; Radford et al. 2025). Recent work by Pasche et al. (2025) shows that AIWP models are competitive with traditional NWP models for a set of extreme heat events in the US and Asia but cannot simulate the mesoscale details and localized extremeness of the heat waves, particularly in regions of complex terrain. Additionally, AIWP models do not include enough surface variables relevant to extremes, as most (e.g., GraphCast, Pangu-Weather) output only 2-m temperature, 10-m wind, and mean sea-level pressure (Camps-Valls et al. 2025; Pasche et al. 2025). Meanwhile, Sun et al. (2024) investigate the ability of FourCastNet to predict "gray swan" tropical cyclones, which they define as strong and primarily unseen extreme events. To do this, they remove Category 3–5 tropical cyclones from their training dataset across various Northern Hemisphere basins and find that FourCastNet is unable to extrapolate stronger tropical cyclones from weaker ones. In addition, the model is unable to satisfy gradient wind balance, suggesting a lack of consistency with respect to the representation of incumbent physical processes.

1.5 Subseasonal to Seasonal (S2S) Forecasting

Improving S2S forecasts, particularly of extreme events, has been a research priority (e.g., Vitart and Robertson 2018). For example, Stan et al. (2022) examine the ability of S2S forecast systems to predict the Madden-Julian Oscillation (MJO) and its associated impacts, including on tropical cyclones and mid-latitude weather extremes in North America. Other examples include work done on S2S forecasts of atmospheric rivers, large-scale moisture flows that greatly impact precipitation extremes in many regions (e.g., DeFlorio et al. 2019; Zhang et al. 2024). Through better S2S forecasts of teleconnections such as the MJO and ENSO, long-range model forecasts

of temperature (e.g., Baker et al. 2023) and precipitation (e.g., Zheng et al. 2025) can also become more skillful.

S2S forecasts of extreme heat events exhibit some skill (e.g., Vitart and Robertson 2018), but are dependent on a host of interconnected physical processes and mechanisms, such as large-scale teleconnections and land-atmosphere interactions. For example, Zhou et al. (2024) find that the coupling of soil moisture conditions to temperature has a large impact on the population impacted by heat waves. Furthermore, Tak et al. (2024) show that the skill of a leading S2S forecast model is hypersensitive to soil moisture drying in forecasts of a 2018 heat wave in Europe.

As extreme events become more common in a warming world (e.g., Vitart and Robertson 2018; Kornhuber et al. 2020), we must make giant leaps in S2S forecast skill of these events. Heat waves, which are in large part dependent on certain large-scale flow patterns (i.e., blocks), offer an opportunity to do so. While traditional NWP approaches to S2S exhibit steady improvement in forecast skill, AIWP models have the potential to accelerate skill gains, with an added benefit of requiring much less computational power.

1.6 Study Motivation

In summary, extreme heat is the deadliest weather-related hazard and is becoming more intense, frequent, and longer duration with climate change. This necessitates better forecasts, particularly on medium-range and S2S timescales. Traditional dynamical NWP approaches to S2S forecasts of extreme temperatures exhibit progress, but dynamical models have numerous limitations that result in a potential predictability plateau. AIWP models have the potential to revolutionize weather and climate forecasting but currently exhibit mixed results when applied to extreme events, and additional verification studies are needed.

As heat wave prediction remains an ongoing challenge, AIWP models offer an opportunity to improve forecasts. However, not much is known about their skill with respect to extreme heat, particularly at longer lead times. As such, the primary objective of this work is to systematically evaluate one traditional NWP model and two AIWP models with respect to heat wave forecasts across the US, at lead times of up to 20 days. While we already have broad insight as to how dynamical NWP models handle extreme temperature forecasts, there is no such knowledge suite for AIWP models. The verification comparison in this study elucidates whether current AIWP models can compete with and potentially surpass the skill of a leading traditional S2S NWP model. Through the evaluation in this study, we expect to make broad conclusions about whether AIWP models can add value to medium- and long-range forecasts of extreme heat events.

Chapter 2, alongside the supplemental material in Appendix A, is submitted as the following paper.

- Ennis, K.E., Barnes, E.A., Arcodia, M.C., Fernandez, M.A., Maloney, E.D. Turning Up the Heat: Assessing 2-m Temperature Forecast Errors in AI Weather Prediction Models During Heat waves. *Submitted to American Meteorological Society Journal of Weather and Forecasting.*

Chapter 3 describes key takeaways from this study, project limitations, future work, and broader impacts.

Chapter 2

Turning Up the Heat: Assessing 2-m Temperature Forecast Errors in AI Weather Prediction Models During Heat Waves

2.1 Introduction

Extreme heat is one of the deadliest weather-related hazards across the globe (e.g., Limaye et al. 2018; Rennie et al. 2021; Ballester et al. 2023), resulting in as many as 5,000 excess deaths per year in the U.S. alone (Weinberger et al. 2021). Excess mortality can be particularly amplified in regions with less modern infrastructure, adaptive capacity, and/or cooling access (e.g., Wu et al. 2022). For example, during a series of heat waves in Europe in 2022, Ballester et al. (2023) estimates that there were more than 60,000 excess deaths due to the cascading impacts of extreme heat. According to the IPCC Sixth Assessment (AR6) and the Fifth US National Climate Assessment (NCA), extreme heat causes adverse impacts to socioeconomic and environmental systems, including but not limited to health, ecosystems, agriculture, energy, as well as national and regional economies (IPCC 2022; USGCRP 2023). Furthermore, the Fifth NCA emphasizes that extreme heat in the US can be associated with drought (i.e., reduced water supply and agricultural yields), increased wildfire activity, and stress on the energy grid (USGCRP 2023). Power outages and blackouts during extreme heat events can also greatly increase mortality and morbidity (e.g., Stone Jr. et al. 2023). Furthermore, energy demand during

extreme heat events is increasing with time and the changing climate (e.g., Rastogi et al. 2021), stressing the electrical grid and in turn making power outages and blackouts more likely.

Extreme heat intensity, frequency, and duration are increasing and projected to continue to do so, both across the U.S. and globally (e.g., Perkins 2015; Mora et al. 2017; Keellings and Moradkhani 2020; Clarke et al. 2022; Barriopedro et al. 2023; Domeisen et al. 2023). To this end, Khatana et al. (2024) projects that deaths related to heat waves and extremely hot temperatures will increase substantially by 2050, particularly impacting older individuals and those with less access to cooling.

Although there is no universal definition for heat waves, they are generally characterized as periods of highly anomalous surface warmth that lasts for at least three days and can occur with or without elevated humidity levels (e.g., Perkins 2015; Barriopedro et al. 2023). Heat waves can be caused by a variety of physical mechanisms, particularly atmospheric blocking, land-atmosphere processes (e.g., anomalously low soil moisture), ocean-atmosphere coupling (e.g., anomalously warm sea surface temperatures), and topography that can set the location of stationary waves and therefore impact the shape of the jet stream (Jiménez-Esteve and Domeisen 2022; Barriopedro et al. 2023; Domeisen et al. 2023).

Traditional numerical weather prediction (NWP) model forecast skill is constantly improving, with useful forecasts (defined as anomaly correlation coefficient > 0.6; Bauer et al. 2015) being regularly produced 10–14 days into the future (Alley et al. 2019; Cahill et al. 2024). However, the predictability of heat waves and other extreme weather events is still a challenge for traditional NWP models, especially for medium-range and subseasonal-to-seasonal (S2S) timescales (e.g., Vitart and Robertson 2018; Lin et al. 2022; Barriopedro et al. 2023; Xie et al. 2024). Improving predictability of extremes on S2S timescales is crucial to establish early

warning systems and increase societal readiness (Vitart and Roberston 2018), including to sectors such as health, agriculture, energy, water resources, and emergency management (Klemm and McPherson 2017; White et al. 2017; White et al. 2022). S2S forecasts have improved over time and can predict the evolution of large-scale and long-duration weather events (e.g., Vitart and Robertson 2018). However, for extreme heat events results from traditional NWP efforts on S2S timescales are decidedly mixed. For example, Ford et al. (2018) found that a NOAA coupled model failed to capture the full duration of heat waves over the US. In addition, the model inaccurately represents land-atmosphere feedbacks in certain regions, degrading forecast skill, which is similarly found by Seo et al. (2024) over the Western US. Using the extended ensemble forecast system from the European Centre for Medium-range Weather Forecasts (ECMWF), Lavaysse et al. (2019) showed that cold spells are more skillfully predicted than heat waves on S2S timescales, although both phenomena become much less predictable at lead times of greater than two weeks. Barriopedro et al. (2023) stated that misrepresentation of coupled land-atmosphere, diabatic, and/or convective processes, as well as model biases with respect to large-scale circulation patterns, can adversely impact S2S forecast skill of extreme heat. Overall, there is much room for improvement to S2S forecasts of extreme heat, and emerging hope that newer artificial intelligence-based weather prediction (AIWP) models can offer substantial advances in predictive skill (Pasche et al. 2025).

AIWP models are increasingly being used in atmospheric science research and operational weather forecasting, as they require a fraction of the required computational power compared to traditional NWP models (Bouallègue et al. 2024; Radford et al. 2025). AIWP models are proving to be skillful forecast tools, especially on larger spatial scales and in some cases longer lead times (e.g., Bouallègue et al. 2024; Waqas et al. 2024; Xie et al. 2024). For example, Xie et al.

(2024) diagnose the evolution of heat waves in China multiple weeks in advance, by training a convolutional neural network (CNN) model to utilize extreme heat precursors. Along those lines, Lopez-Gomez et al. (2023) use a set of neural weather models to forecast global surface temperature anomalies, finding significant skill improvement compared to traditional S2S NWP models up to 28 days in advance. Meanwhile, Li et al. (2023) train a graph neural network model on surface station data across the CONUS and use it to make skillful forecasts of regional heatwaves. Overall, AIWP models can produce forecasts that compete with traditional operational NWP models (Hakim and Masanam 2024), offering promise for their future in predicting key weather elements such as temperature and precipitation (e.g., Waqas et al. 2024).

Despite their potential, AIWP models have numerous limitations, including their inability to represent fundamental dynamic and thermodynamic processes (Selz and Craig 2023; Bonavita 2024), struggles to accurately simulate mesoscale weather features (Bonavita 2024), producing overly smooth forecasts and increasing biases with time (Bouallègue et al. 2024), and failure to reproduce the butterfly effect associated with atmospheric chaos (i.e., the model forecasts are not sensitive enough to initial conditions; Selz and Craig 2023). AIWP models also require a more intense verification by the scientific community, particularly for forecasts of extreme events (Pasche et al. 2025; Radford et al. 2025; Ullrich et al. 2025). As an example, Pasche et al. (2025) examine three recent extreme events (two heat waves) using four AIWP models. Specifically for the 2021 Pacific Northwest and 2023 South Asian heat waves, they find similar accuracy between the AIWP models and a leading traditional NWP model. However, the AIWP models lack small scale details and the appropriate variables to conclusively diagnose the extreme heat events.

This study aims to systematically evaluate the ability of two AIWP models (Google GraphCast and Pangu-Weather, hereafter GraphCast and Pangu) to predict heat waves in all four boreal seasons across the CONUS and compare their skill to a traditional S2S NWP model, the NOAA United Forecast System Global Ensemble Forecast System (UFS GEFS). We focus our evaluation on the medium-range and subseasonal prediction timescales (out to 20 days). First, surface temperature forecasts from the two AIWP models and UFS GEFS are evaluated and compared for two heat wave case studies across disparate CONUS regions, allowing us to investigate model skill and sensitivity at regional levels. Subsequently, the regional and seasonal temperature forecast skill of the two AIWP models is explored across a larger group of heat waves, during and outside of the model training period, allowing for a robust evaluation of model performance during extreme heat events. Our primary objective is to assess—specifically for heat waves across the CONUS—if, where, and when the two AIWP models show promise or offer advantages over a traditional NWP system.

2.2 Data and Methods

2.2.1 Traditional NWP Model and Verification Dataset

For the traditional NWP model, we use the NOAA GEFSv12, an uncoupled version of the UFS (hereafter UFS GEFS) produced in September 2020 (Guan et al. 2022). Variables from the control run are available on a six-hourly basis at 0000, 0600, 1200, and 1800 UTC. The reforecasts are run out to lead times of 1–35 days (inclusive), leading to a total of 1042 samples initialized every seven days from 5 January 2000 to 18 December 2019. To create daily mean temperatures, we average these 6-h instantaneous forecasts for each day over the entire dataset. NOAA provides UFS GEFS data for lead times of Days 11–35, which is produced at half the

grid spacing (0.5° x 0.5°) of the data for lead times 1–10 (0.25° x 0.25°). Therefore, prior to merging the datasets for calculations, we regrid output fields for lead times 11–35 using the python module xESMF's bilinear interpolation (Zhuang et al. 2024), such that all lead times have a 0.25° x 0.25° grid spacing. This ensures that all forecast lead times share the same spatial coordinates, allowing for comparisons and calculations without spatial mismatch.

To calculate errors in model forecasts and identify large scale heat waves (section 2.2.3), we use the ECMWF ERA5 reanalysis (Hersbach et al. 2020) as our ground truth. ERA5 is produced on a reduced Gaussian grid, with a quasi-uniform grid spacing of approximately 0.25° x 0.25° and has hourly data from 1940–present. To obtain daily 2-m temperature fields, we aggregate the 6-h 2-m temperature output to 24-h resolution. While the full ERA5 dataset goes back to 1940, the two AIWP models (section 2.2.2) are trained on ERA5 data from 1979–2018.

2.2.2 AIWP Models

GraphCast, an AIWP deep-learning model, is based on a graph neural network (GNN) architecture following an "encode-process-decode" configuration (Lam et al. 2023). GNN-based learned simulators are successful at capturing complex fluid dynamics and other governing partial differential equations, making them well equipped for weather modeling (e.g., Lam et al. 2023). GraphCast is autoregressive, meaning it can be "rolled out" by feeding its own predictions back in as input, to generate longer trajectories of weather states (Lam et al. 2023). The model includes four surface variables (2-m temperature, mean sea-level pressure, 10-m u-component of the wind, 10-m v-component of the wind) and five additional variables (temperature, geopotential height, specific humidity, u-component of the wind, v-component of the wind) on 37 pressure levels, outputting variables on a six-hourly basis.

Pangu Weather is also an AIWP deep learning model and incorporates a 3D Earth-specific transformer (3DEST) deep network architecture, which allows the model to represent spatial dependencies through self-attention mechanisms (Bi et al. 2023). Pangu outputs five variables (temperature, geopotential height, specific humidity, u-component of the wind, v-component of the wind) on 13 pressure levels; at the surface, there are four output variables (2-m temperature, mean sea-level pressure, 10-m u-component of the wind, 10-m v-component of the wind). Pangu has four trained networks, each optimized for different forecast intervals: 1-h, 3-h, 6-h, and 24-h. For medium-range forecasting, the Pangu model developers introduced hierarchical temporal aggregation. As a part of this approach, developers suggest using the 24-h forecast model when running forecasts out to longer lead times to minimize the number of iterations required (Bi et al. 2023). However, we use the Pangu 6-h model to remain consistent with UFS GEFS and GraphCast output. This ensures that we can compute consistent daily averages for each lead time across all the models for accurate comparison. We discuss this choice in more detail in section 2.4.

2.2.3 Heat Wave Identification

For a thorough investigation of the 2-m temperature prediction skill of the UFS GEFS and AIWP models, we investigate a total of 60 heat waves (15 per season) during 2000–2023. These heat waves span the AIWP model training period (1979–2017) and extend past it into the model testing years (2018–2023). We select 60 heat waves to ensure that we have a large enough sample size per boreal season to be able to divide events into two groups: heat waves in the model testing years and heat waves in the training years. There are seven total heat waves in each season that are in Pangu and GraphCast's testing years and eight that occur during the training years. Seven events in each season provides a focused dataset for evaluating the predictive skill

of the models in the testing years. The heat waves in the six testing years offer insight into their prediction skill post-training, although we acknowledge that seven heat waves per season is still a limited sample size.

To produce our heat wave event database, we identify heat waves during boreal summer (JJA), autumn (SON), winter (DJF), and spring (MAM) for 2000–2023. Heat waves are identified using ERA5 daily mean 2-m temperature data. There is no universal definition of heat waves (Perkins 2015); however, many studies stipulate a percentile-based threshold approach with a minimum duration of three days (e.g., Perkins 2015; Cloutier-Bisbee et al. 2019), which we apply here. We choose the 95th percentile of 2-m temperatures in each season as our threshold. A heat wave event at each grid point is identified when the 95th percentile is exceeded for three or more days. Using the 98th and 99th percentiles result in some similar heat waves but provides us with too few events per season.

To identify large-scale heat waves, we use the NCA regions (USGCRP 2023, Fig. 1). We focus on heat waves within four disparate NCA regions: the Northwest, Southeast, Midwest, and Northeast (Fig. 2.1). We require each heat wave to have at least 100 contiguous grid points (a 2.5° x 2.5° box if a square) exceeding the 95th percentile of seasonal 2-m temperatures on a given day. This ensures that each heat wave affects a significant portion of the CONUS. As an example, Fig. 2.2 shows results for the record-shattering June 2021 Pacific Northwest heat wave (Schumacher et al. 2022; White et al. 2023). In Fig. 2.2a, only a few grid points exceed the 95th percentile threshold on 27 June 2021. One day after (28 June 2021, Fig. 2.2b), there is a large area that exceeds our percentile and spatial thresholds over much of the Northwest NCA region; as such, we define the start day of the heat wave as 28 June 2021, consistent with other work (e.g., Schumacher et al. 2022). We allow a heat wave event to exist in multiple NCA regions,

simultaneously and/or over the course of the event. For example, a large heat wave from 30 June to 2 July 2012 affected parts of the Midwest and Southeast regions (Fig. A.1).

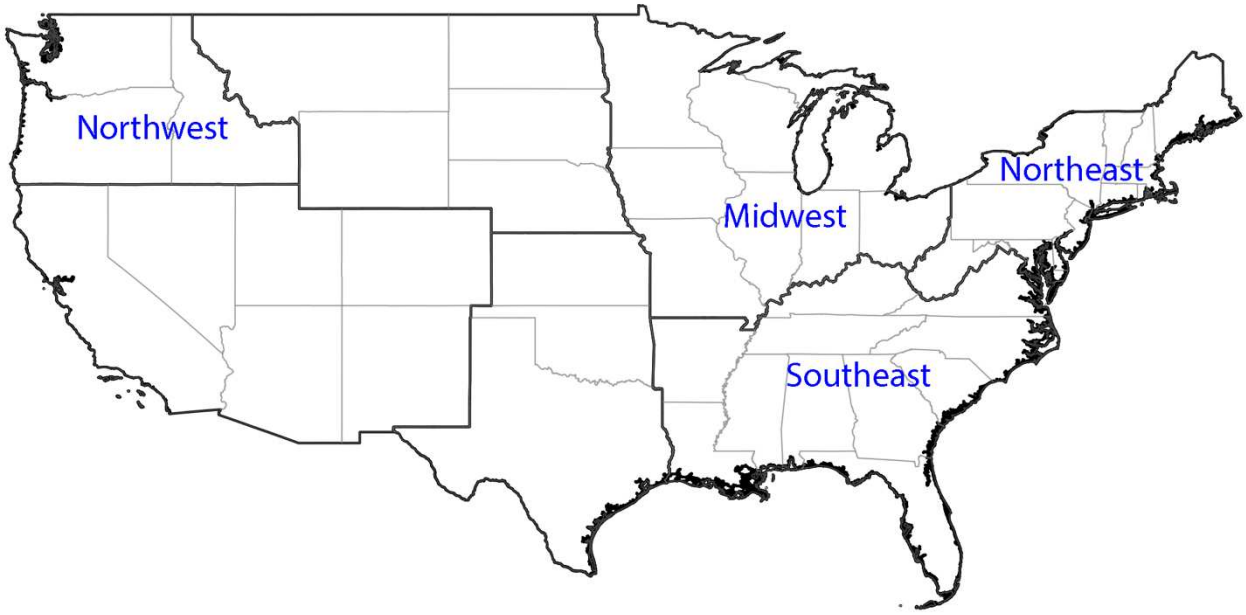

Figure 2.1: Map of the seven US National Climate Assessment regions (black bold borders) over the CONUS, with the four regions used in this study (Northwest, Midwest, Northeast, Southeast) labeled in blue. Adapted from USGCRP (2023).

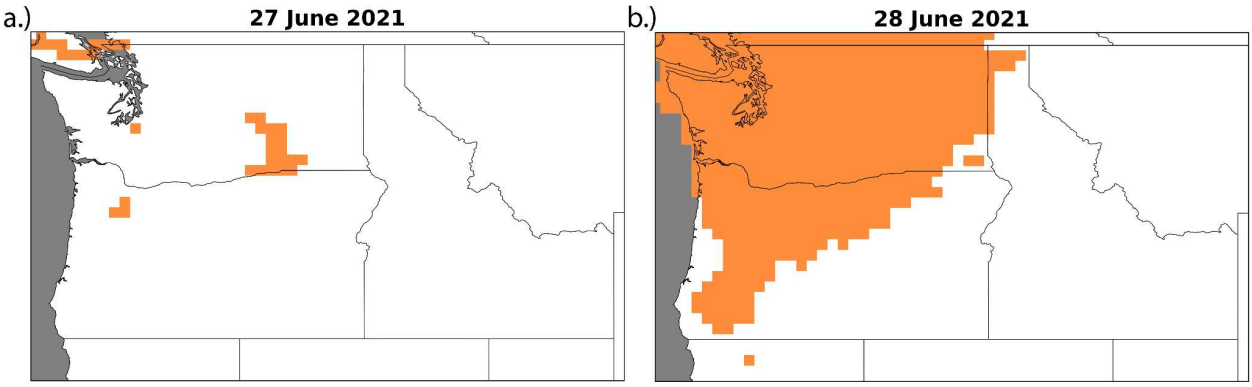

Figure 2.2: The June 2021 record-shattering heat wave in the Northwest NCA region (Washington, Oregon, Idaho), where the orange shading indicates areas where daily mean 2-m temperatures exceeded the 95th percentile on (a) 27 June and (b) 28 June.

Finally, we emphasize that anomalous warm events outside of boreal summer have no standard name. Some studies (e.g., Schwarz et al. 2020) refer to them as “extreme heat”, while

other sources such as ECMWF term them “warm spells” (ECMWF 2020). Regardless of specific terminology, our definition of heat wave events (exceeding the 95th percentile of 2-m temperature with a minimum duration of three days over at least 100 adjacent grid points) remains consistent across all four seasons and our four regions. Therefore, for the remainder of this paper we use the term “heat wave” regardless of region and season.

2.2.4 Model Comparison

In section 2.3.1, we investigate how 2-m temperature forecast error for both AIWP models, GraphCast and Pangu, compares to that of the UFS GEFS during heat waves. We examine two case studies to demonstrate the performance of each model; specifically, we select the August 2011 Southeast summer and 2019 Pacific Northwest autumn heat waves. These heat waves are chosen for our case studies because 1) they occur in two different seasons, and 2) the 2011 Southeast heat wave falls within the AIWP model training years, while the 2019 heat wave occurs during the testing years. Furthermore, because it is only initialized every seven days, UFS GEFS has specific forecast dates from which to choose. Thus, we must be intentional when choosing UFS GEFS forecast initialization dates to ensure that the reforecast period covers the majority of days during each heat wave. For the August 2011 heat wave, we choose the reforecast file with an initialization date of 27 July 2011, seven days prior to the onset of the heat wave event. For the September 2019 event, we choose the reforecast file with an initialization date of 28 August 2019, four days prior to the onset of the heat wave. We then generate 20-day GraphCast and Pangu hindcasts, initializing them on the same date as the UFS GEFS reforecasts. ERA5 is used as the relevant “ground truth” baseline throughout each event’s forecast period to compare performance.

In section 2.3.2, we isolate the two AIWP models and assess their subseasonal 2-m temperature prediction skill systematically. Unlike the UFS GEFS, GraphCast and Pangu do not have specific weekly initialization dates, enabling us to initialize hindcasts at any date. We initialize these hindcasts such that lead time 10 of the model forecast is the first day of the heat wave. Hindcasts are run out to 20 days.

2.3 Results

2.3.1 Case Study Evaluation

For the August 2011 Southeast heat wave, the UFS GEFS maintains low errors throughout the entire forecast, despite a slight warm bias in some locations (Fig. 2.3). UFS GEFS outperforms Pangu at all lead times but exhibits only slightly smaller errors than GraphCast through day 12 (Fig. 2.3). At a lead time of 7 days (3 August 2011; the first day of the heat wave), UFS GEFS (Fig. 2.4b) matches well with ERA5 (Fig. 2.4a), while both GraphCast (Fig. 2.4c) and Pangu (Fig. 2.4d) exhibit a cool bias throughout the Southeast. However, the Pangu cool bias is substantially larger than that of GraphCast. We choose lead time 7 for analysis here (3 August 2011) not only because it is the first day of the heat wave, but also because it is the hottest day in terms of the Southeast regional average temperature (Fig. 2.3a). To that end, of the three models, only UFS GEFS captures the widespread intensity of the heat wave across the region (Fig. 2.4). However, an encouraging aspect of the forecasts is that all three models capture the relatively cooler temperatures in the mountains of western North Carolina and eastern Tennessee, albeit with slightly varying skill (Fig. 2.4).

The hindcasts in Fig. 2.3 illustrate forecast performance evolution leading up to and through the duration of the August 2011 Southeast heat wave. UFS GEFS is the best of the three forecast

models prior to and during the early part of the heat wave, with temperature errors near zero (Fig. 2.3b). GraphCast is the better of the two AIWP models and exhibits smaller errors than UFS GEFS toward the end of the heat wave (Fig. 2.3). In contrast, Pangu exhibits the largest errors and inconsistency throughout much of the heat wave, with a persistent cold bias.

The September 2019 Northwest heat wave (Figs. 2.3c,d) features larger temperature fluctuations and errors across all models compared to the 2011 Southeast event. Large warm biases in UFS GEFS boreal summer 2-m air temperature forecasts are demonstrated over the Northwest by Seo et al. (2024), and agree with our results, especially prior to and during the first half of the heat wave (Fig. 2.3d). In contrast to the UFS GEFS warm bias, GraphCast exhibits a cool bias leading up to the 2019 heat wave, lower error during much of the event, and a prominent warm bias near and just after the end of the heat wave (Fig. 2.3d). Pangu errors are larger than GraphCast but competitive with UFS GEFS; Pangu is particularly inconsistent prior to and during the first half of the heat wave (Fig. 2.3d). Overall, while GraphCast exhibits smaller errors (by 2–3°C) than UFS GEFS during the heatwave period, both GraphCast and Pangu have warm biases near the end of and just after the heat wave that exceed the warm bias in UFS GEFS by 2–4°C (Fig. 2.3d).

We next examine temperature errors for the August 2011 heat wave at three distinct locations within the Southeast NCA region (Fig. 2.1): Asheville, NC; Little Rock, AR; and New Orleans, LA. These locations are chosen based on their differing geographical characteristics (i.e., coastal proximity and elevation). All three models exhibit a cold bias for nearly the entire period at all three cities (Figs. 2.5b,d,f), except for Pangu near the end of the heat wave at Little Rock (Fig. 2.5f). GraphCast consistently outperforms both UFS GEFS and Pangu at the highest elevation location (Asheville; Fig. 2.5b). Meanwhile, the largest errors in all three models occur at Little

Rock (Fig. 2.5f), especially before and during the first half of the heat wave. While Asheville (650 m) is at a much higher elevation than Little Rock (102 m), both western North Carolina and Arkansas have large topography gradients. Therefore, it is plausible that the coarse grid spacing of UFS GEFS limits its ability to accurately predict 2-m temperature extremes in such regions. At New Orleans (Fig. 2.5d), UFS GEFS has a slightly larger cold bias compared to both AIWP models. This is notable because New Orleans is located in a flat area near or below sea level, and adjacent to the Gulf. Therefore, its temperature is regulated by its proximity to the Gulf, is not impacted by orographic effects, and generally has lower variability than locations farther inland.

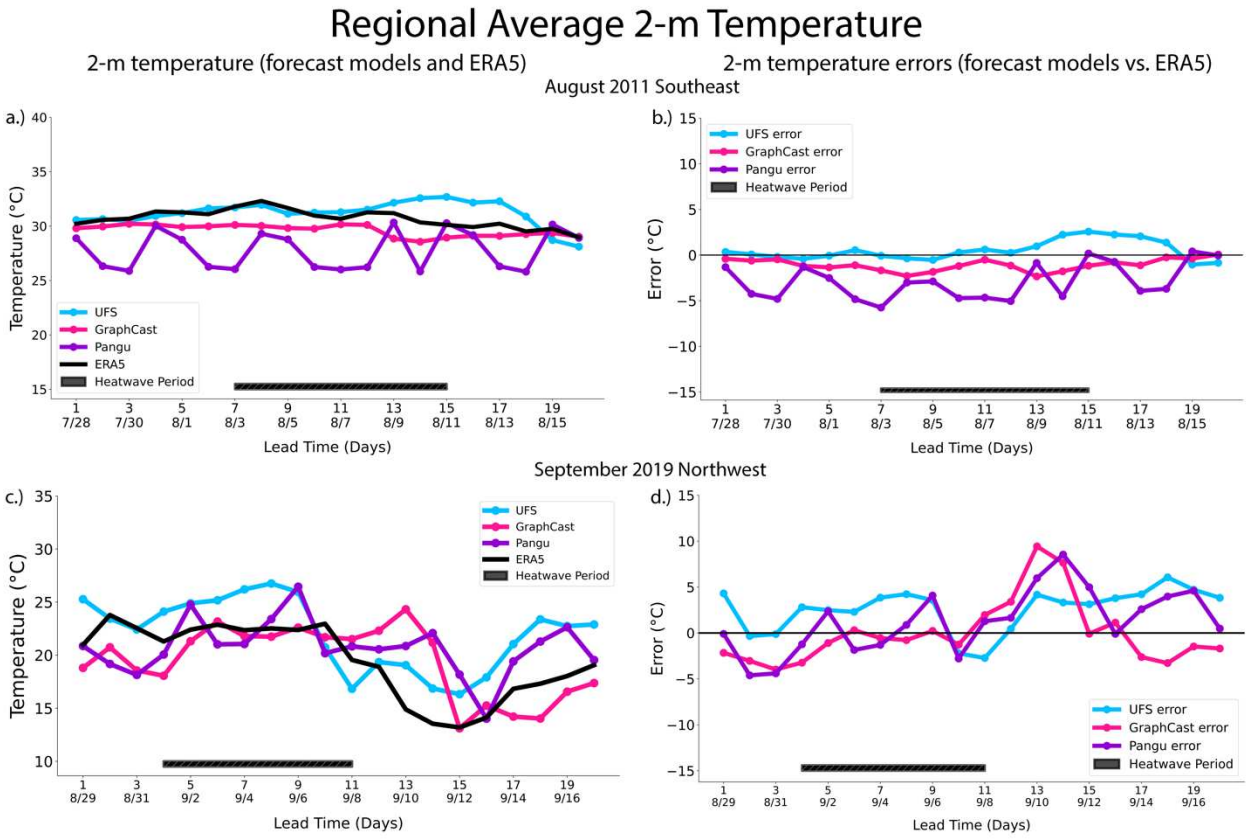

Figure 2.3: Regional average (left) 2-m temperature (°C) and (right) 2-m temperature errors (°C) in UFS GEFS (blue line), GraphCast (magenta line), and Pangu (purple line) for the (a,b) August 2011 Southeast heat wave and (c,d) September 2019 Northwest heat wave. Temperature errors are calculated using ERA5 (black line on left-hand panels) as truth and are shown as a function of time. The heat wave period is illustrated by the black line at the bottom of each panel.

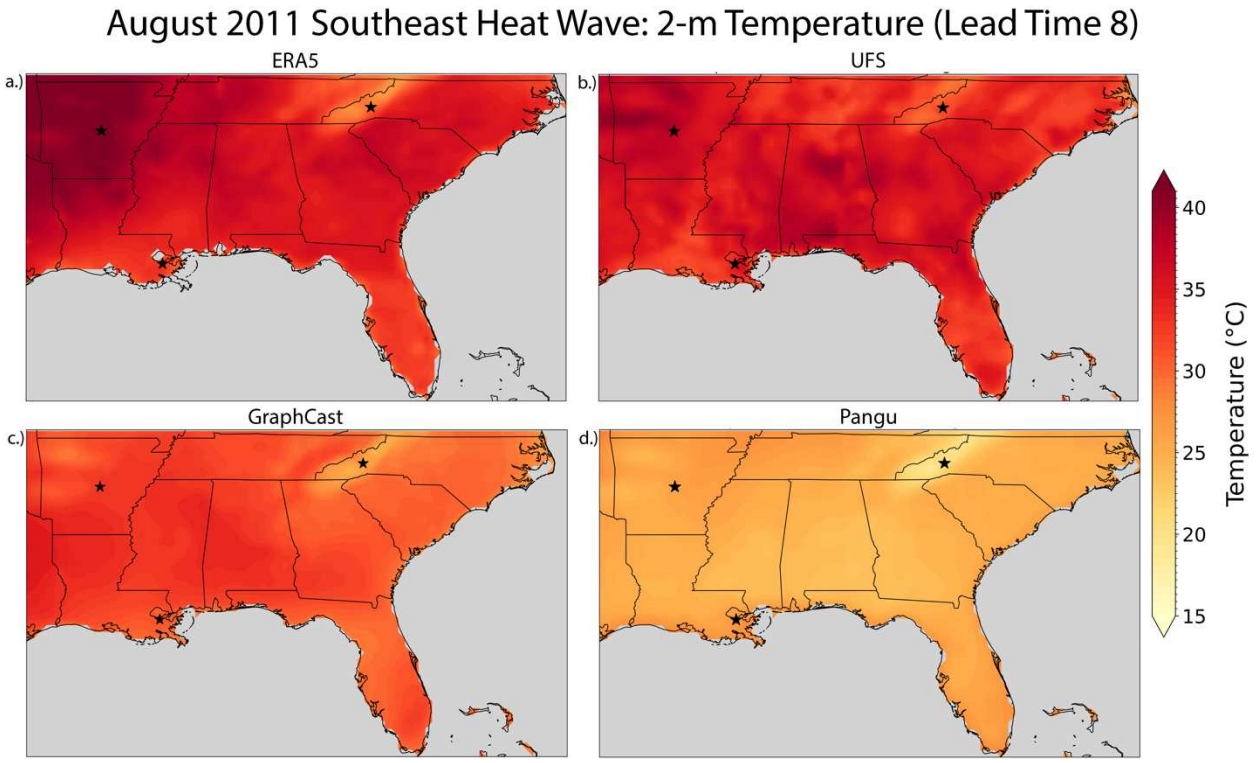

Figure 2.4: For the August 2011 Southeast heat wave valid at lead time 7 (3 August 2011; the first day of the heat wave): 2-m temperature (°C, shaded) for (a) ERA5 verification; and forecasts from (b) UFS GEFS; (c) GraphCast; and (d) Pangu. The black stars (Asheville, NC; New Orleans, LA; Little Rock, AR) denote the three cities selected for analysis in Fig. 2.5.

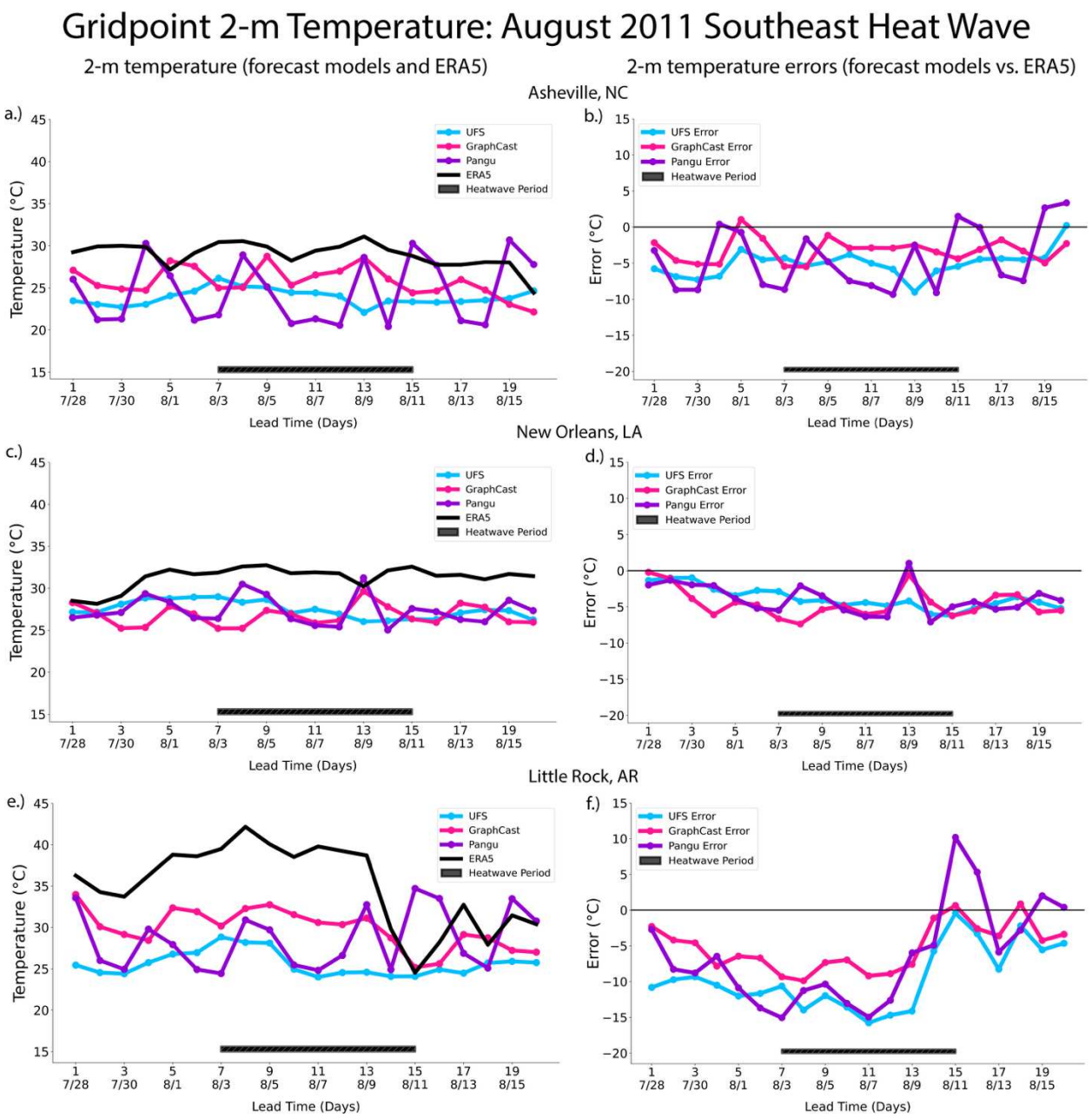

Figure 2.5: Grid point values of (left) 2-m temperature (°C) and (right) 2-m temperature errors (°C) in UFS GEFS (blue line), GraphCast (magenta line), and Pangu (purple line) for three locations during the August 2011 Southeast heat wave: (a,b) Asheville, NC; (c,d) New Orleans, LA; (e,f) Little Rock, AR. Temperature errors are calculated using ERA5 (black line on left-hand panels) as truth, and are shown as a function of time. The heat wave period is illustrated by the black line at the bottom of each panel.

We next investigate temperature errors in Seattle, WA; Bend, OR; and Boise, ID for the September 2019 Northwest heat wave (Fig. 2.6). As for the August 2011 case, we examine grid points with varying geography. UFS GEFS exhibits a consistent cool bias in all three locations (Figs. 2.6b,d,f), mirroring results for the August 2011 Southeast event, and is perhaps also related to the coarse resolution limiting accuracy in regions of complex terrain and topography gradients. GraphCast maintains the lowest errors for all three locations before and during the heat wave period. However, the model exhibits a consistent warm bias and higher error after the heat wave ends (Figs. 2.6b,d,f). Pangu performs well compared to UFS GEFS in all three locations but exhibits much larger error variability in Boise (Fig. 2.6f) than in Seattle (Fig. 2.6b) and Bend (Fig. 2.6d). All three models perform the worst in Boise (Fig. 2.6f), which of our three cities is the location with the highest elevation and most complex terrain. While both AIWP models outperform UFS GEFS at the three specific Northwest locations, this is not the case for the regional averages (Fig. 2.3).

2.3.2 AIWP Model Performance

We next examine the regional average performance of GraphCast and Pangu during four heat waves that affected four different NCA regions (Northwest, Midwest, Northeast, Southeast). Figures 2.7a and 2.7b show results for the June 2021 record-shattering Northwest heat wave (e.g., Schumacher et al. 2022; White et al. 2023). Pangu exhibits a persistent cold bias with large variations (Fig. 2.7b), matching the regional average results for the September 2019 event (Fig. 2.4). The cold bias is particularly pronounced during the early and middle portions of the heat wave (Fig. 2.7b). GraphCast errors are minimal before heat wave onset; however, there is a consistent cold bias during the heat wave, albeit a smaller one than Pangu (Fig. 2.7b). The

average GraphCast absolute error over the 20-day forecast for this event (3.2 °C) is the largest of the four events in Fig. 2.7.

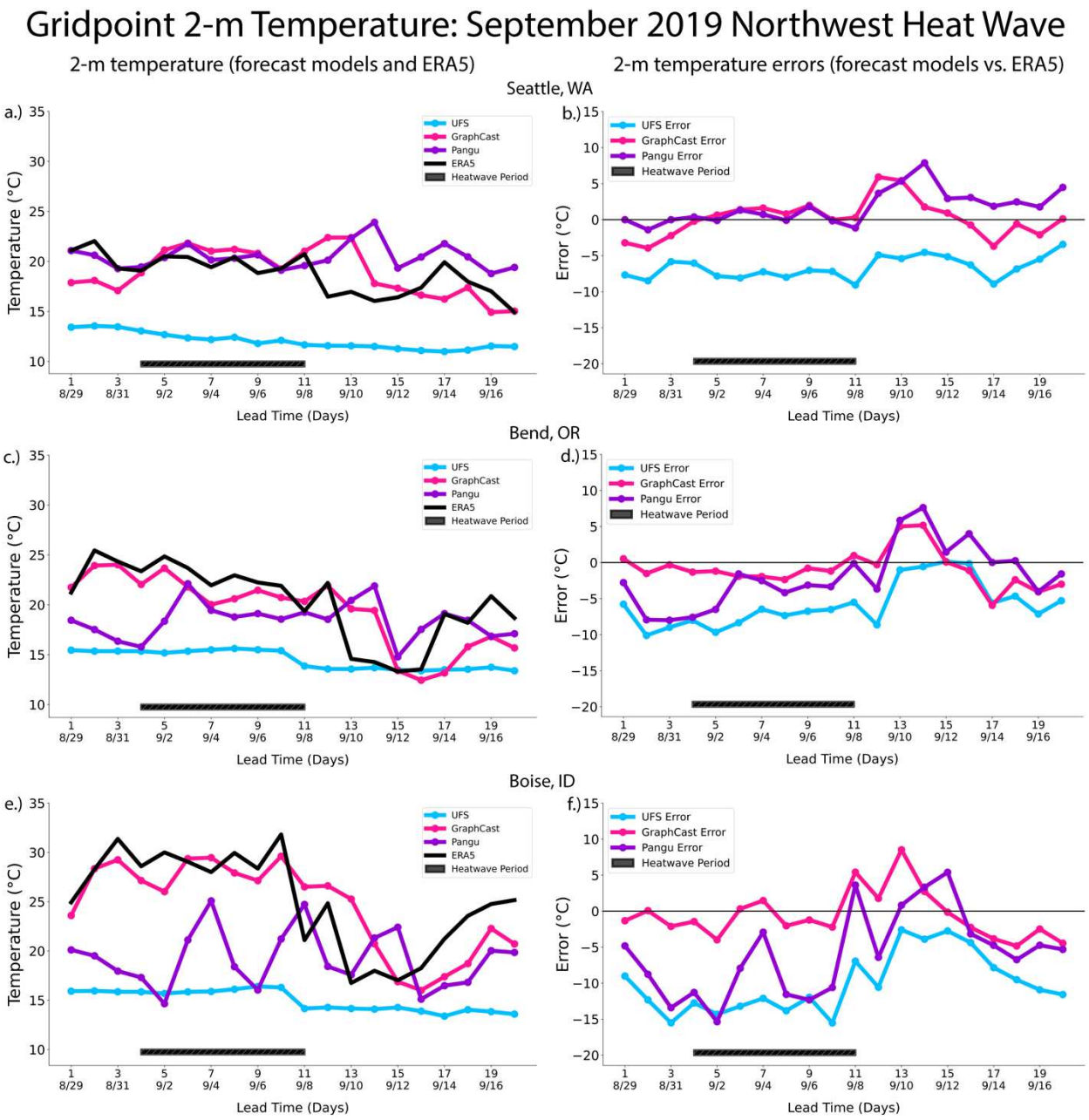

Figure 2.6: As in Fig. 2.5, but for the September 2019 Northwest heat wave.

For the September 2023 Midwest heat wave (Figs. 2.7c,d), GraphCast consistently outperforms Pangu. GraphCast error is nearly zero until heat wave onset, followed by a slight cold bias during the heat wave and a small warm bias after the heat wave (Fig. 2.7d). The average GraphCast absolute error (1.4 °C) is the lowest of the four events in Fig. 2.7. Pangu exhibits the same general error pattern as GraphCast but is more inconsistent and has a much larger cold bias during the heat wave (Fig. 2.7d).

The GraphCast error behavior for the December 2015 Northeast event mirrors the September 2023 Midwest results, but with larger absolute errors during the heat wave period (5–10 °C vs. 2 – 4°C) and in the overall averages (2.7 °C vs. 1.4 °C) (Figs. 2.7d, f). Specifically, GraphCast exhibits small errors before the heat wave, has a cold bias during most of the heat wave period, and exhibits a warm bias near the end of the heat wave (Fig. 2.7f). There is a similar error behavior in Pangu during and after the heat wave period, but Pangu has a large warm bias just prior to heat wave onset (Fig. 2.7f). Pangu for the Northeast event is the only case in which either model has a large warm bias prior to the heat wave, perhaps reflective of it initiating event onset too soon.

GraphCast and Pangu errors for the May 2004 Southeast heat wave (Figs. 2.7g,h) agree well with the August 2011 event in the same region (Fig. 2.4). Both models exhibit a cold bias throughout the entire forecast period, with errors larger after heat wave onset than before it (Fig. 2.7h). As in the August 2011 event, GraphCast consistently outperforms Pangu, especially near heat wave onset. The average GraphCast absolute error (1.9 °C) is the second lowest of the four events (after September 2023 Midwest), and like the cases in Fig. 2.4, GraphCast performs better in the Southeast than the Northwest.

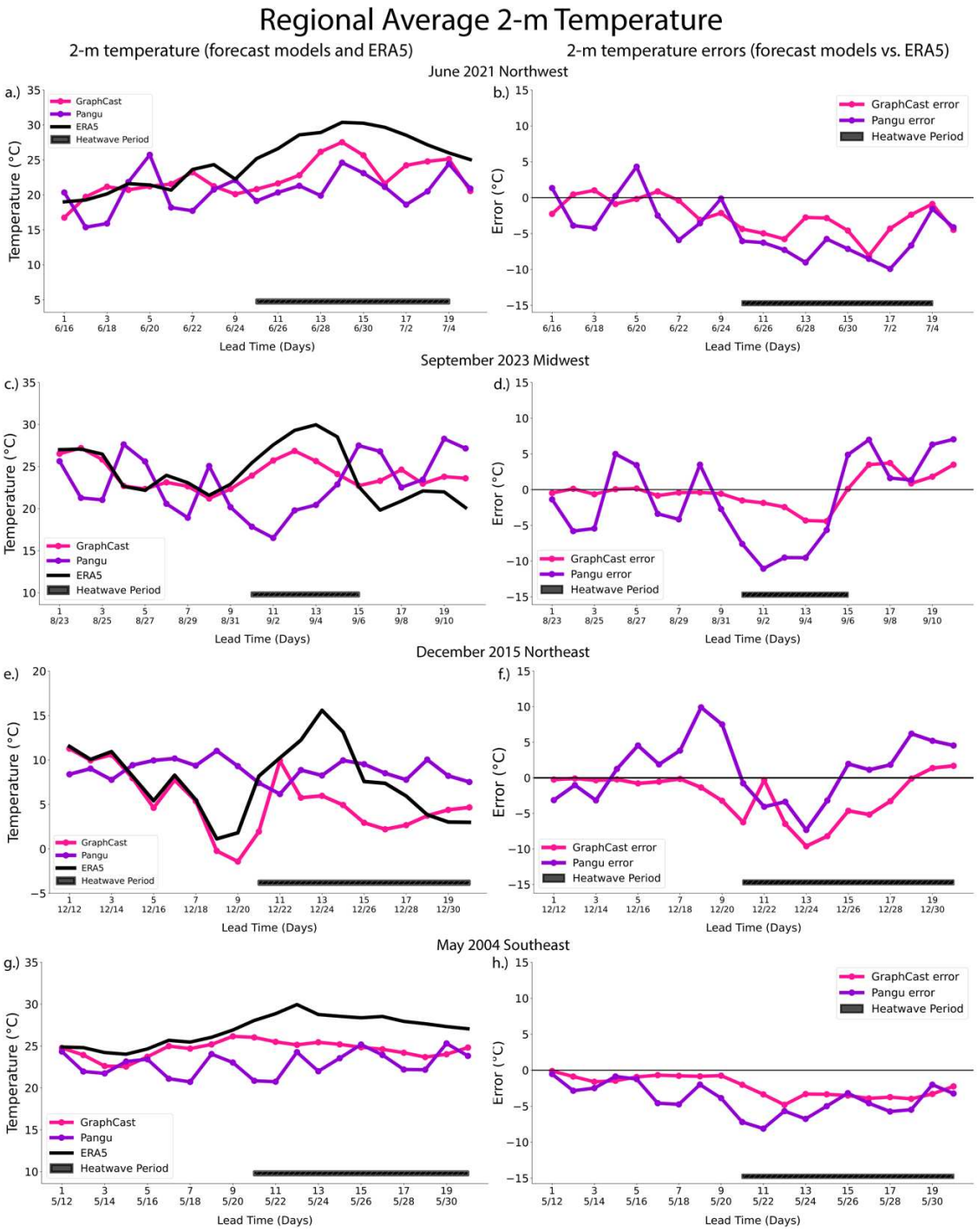

Figure 2.7: Regional average (left) 2-m temperature (°C) and (right) 2-m temperature errors (°C) in GraphCast (magenta line) and Pangu (purple line) for the (a,b) June 2011 Northwest heat wave; (c,d) September 2023 Midwest heat wave; (e,f) December 2015 Northeast heat wave; and (g,h) May 2004 Southeast heat wave. Temperature errors are calculated using ERA5 (black line on left-hand panels) as truth and are shown as a function of time. The heat wave period is illustrated by the black line at the bottom of each panel.

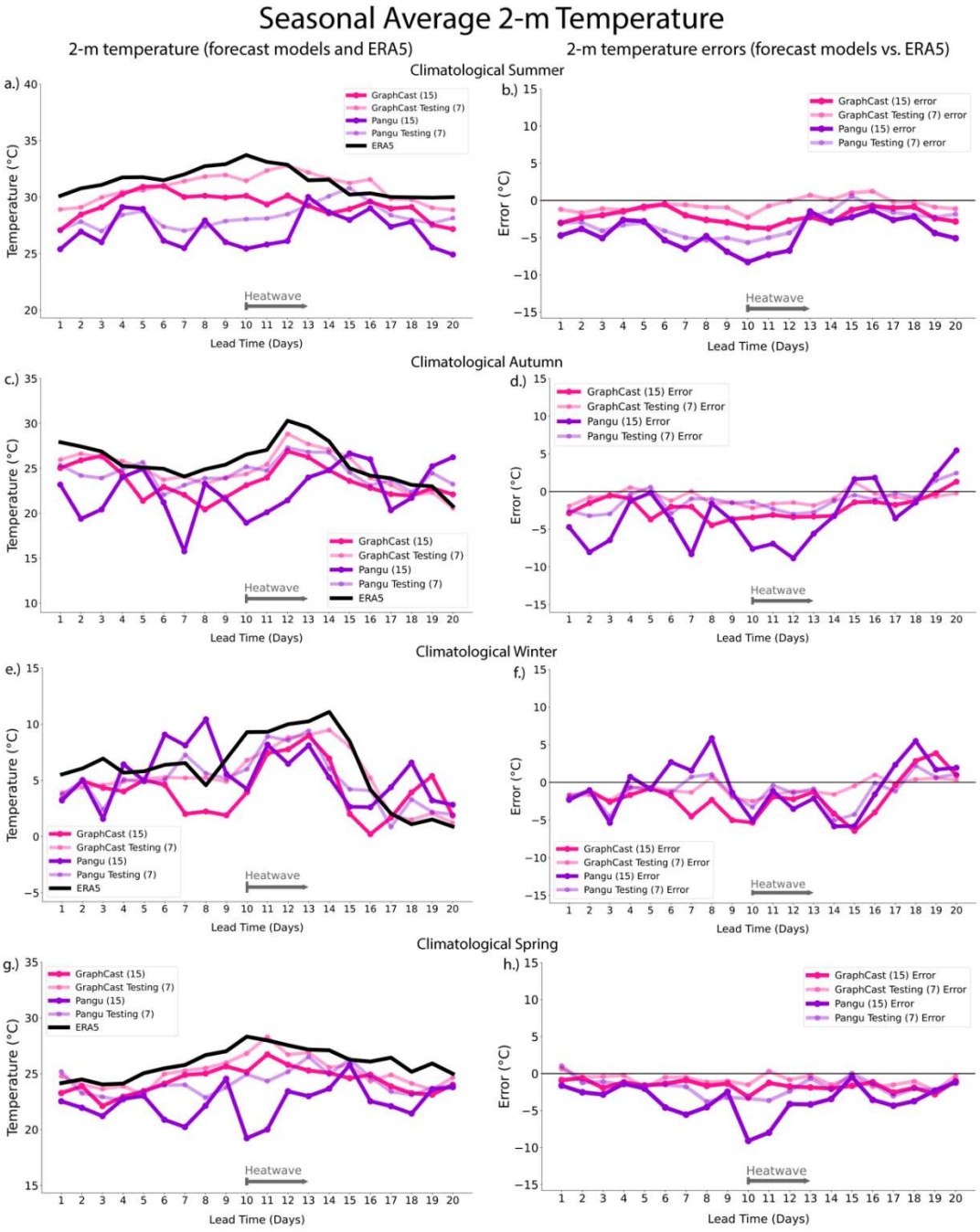

Figure 2.8: Seasonal average (left) 2-m temperature (°C) and (right) 2-m temperature errors (°C) for heat waves during the GraphCast training (1979–2017; magenta line) and testing (2018–2023; light pink line) periods, and Pangu training (1979–2017; purple line) and testing (2018–2023; lavender line) periods. Temperatures and temperature errors are averaged over eight training and seven testing period heat waves during (a,b) boreal summer (June-July-August); (c,d) boreal autumn (September-October-November); (e,f) boreal winter (December-January-February); and (g,h) boreal spring (March-April-May). Temperature errors are calculated using ERA5 (black line on left-hand panels) as truth and are shown as a function of time. The heat wave period starts at lead time 10, as illustrated by the black line at the bottom of each panel.

We next examine the performance of the AIWP models on a seasonal average basis, with an approximately equal number of cases in all four regions. Fifteen heat waves (section 2.2.3) per season are chosen for analysis. In each season, eight heat waves are during the model training years (1979–2017) and seven heat waves during the model testing years (2018–2023).

For boreal summer (June-July-August), GraphCast and Pangu both exhibit consistent cold biases throughout the forecast period, regardless of whether the event is during the training or testing years (Figs. 2.8a,b). The largest errors (cold biases) occur near heat wave onset, consistent with our case study results. Overall, GraphCast substantially outperforms Pangu during boreal summer, especially leading up to and during the heat wave period. Boreal spring (March-April-May) results (Figs. 2.8g,h) mirror our summer results (Figs. 2.8a,b), in that GraphCast substantially outperforms Pangu and there is a consistent cold bias throughout the forecast period in both models, regardless of whether it is the training or testing years.

In boreal winter (December-January-February), Pangu exhibits a warm bias leading up to heat wave onset, which is substantially larger during the training period (Figs. 2.8e,f). GraphCast and Pangu absolute errors are similar for most of the forecast period, although the GraphCast errors are manifested through a cold bias, the opposite of Pangu. Finally, boreal autumn (September-October-November) results (Figs. 2.8c, d) exhibit an early cool bias similar to our summer results (Fig. 2.8b), but a post-heat wave warm bias more like our winter results (Fig. 2.8f).

To provide spatial context to the regional average results shown throughout this section, Fig. 2.9 shows GraphCast (Fig. 2.9a) and Pangu (Fig. 2.9b) 2-m temperature errors at lead time 10 for all heat waves across the four selected NCA regions. To create Fig. 2.9, in each region we first average lead time 10 forecasts over all events; lead time 10 is the first day of the heat wave in

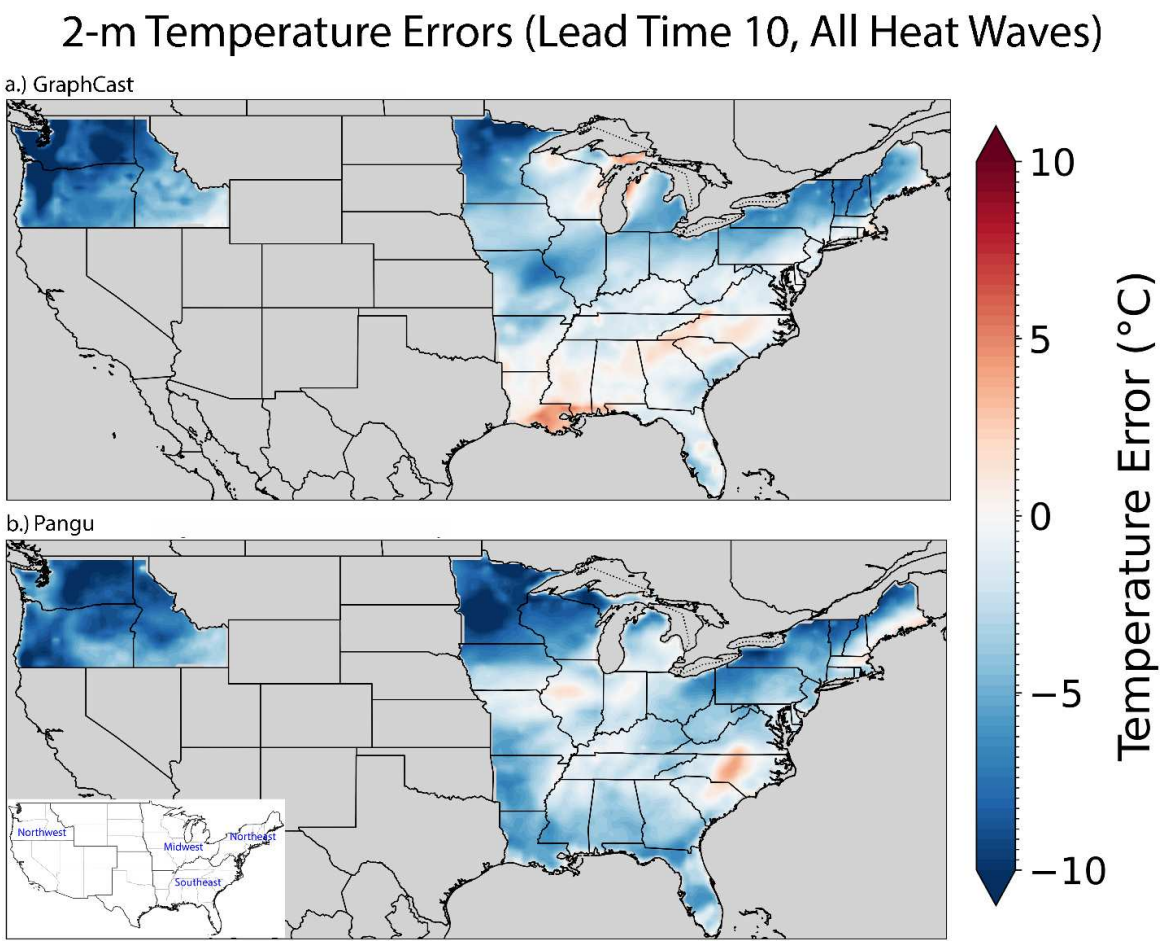

Figure 2.9: For the four NCA regions (Northwest, Midwest, Northeast, Southeast), spatial distribution of 2-m temperature errors (°C, shaded) averaged over all heat wave events across all boreal seasons for (a) GraphCast and (b) Pangu. The lower left inset shows a map of the seven NCA regions (black outlines; USGCRP 2023) with the four regions used in this study labeled in blue, also shown in Fig. 2.1.

each event. Next, we plot the average lead time 10 temperature error for each grid point within a respective region (Fig. 2.9). For example, the 2-m temperature errors shown in the Northwest are lead time 10 errors at each grid point averaged over the 16 total heat waves (i.e., heat waves that fall within both the model testing and training years) in that region. It is evident that both GraphCast and Pangu have the tendency to underestimate 2-m temperatures during heat waves, particularly within the Northwest region (Figs. 2.9a,b). In contrast, the AIWP models predict slightly warmer than observed temperatures within the mountainous areas of the Carolinas in the

Southeast region. This discrepancy may arise from the relatively coarse representation of terrain features in the AIWP models in this study. Furthermore, our analysis demonstrates that Pangu's temperature errors (Fig. 2.9b) generally exceed those of GraphCast. GraphCast exhibits a specific tendency to overpredict temperatures along the Gulf Coast, particularly in Louisiana and Alabama (Fig. 2.9a), but still generally outperforms Pangu.

2.4 Discussion and Conclusions

This study evaluates the 2-m temperature forecast performance of UFS GEFS and two AIWP models (GraphCast and Pangu) for a set of heat waves across four US NCA regions. For our two case studies (Figs. 2.3–2.6), GraphCast and UFS GEFS exhibit similar performance in our regional average results (Figs. 2.3 and 2.4). The UFS GEFS results throughout our case studies are mixed: There are small warm biases in the regional averages, especially in the Northwest (Fig. 2.3), but large cold biases at specific locations throughout the Northwest and Southeast (Figs. 2.5 and 2.6). To provide additional context, in a study of UFS boreal summer forecast skill, Seo et al. (2024) report that UFS has a consistent warm bias in the western CONUS at all lead times. However, for annual mean 2-m temperature, Stefanova et al. (2022) show a large mean cold bias over the entire CONUS at all lead times. In an examination of boreal summer extreme events, Krisnamurthy and Stan (2022) show large UFS errors for 2-m temperature in both the Southeast and Northwest with large sub-regional variability throughout the western CONUS, especially at longer lead times. Previous studies suggest several potential reasons for large UFS 2-m temperature forecast errors, especially for extreme heat events at longer lead times and over the Western CONUS: biases in the slowly varying modes of climate variability in the Pacific Ocean (e.g., ENSO, Krisnamurthy and Stan 2022; Stan et al. 2023), tropical Pacific convection and the Madden-Julian Oscillation (Choi and Stan 2025), as well as land-atmosphere

coupling and soil moisture (Benson and Dirmeyer 2023; Seo et al. 2024). Although it is beyond the scope of this study to investigate the causes of the UFS GEFS errors in our case studies, future work could examine each of these potential mechanisms as a source of forecast error during heat waves.

Overall, GraphCast performs quite well across all seasons, albeit with a persistent small cold bias prior to and during the heat wave period (Fig. 2.8). Pangu does better in the seasonal evaluation (Fig. 2.8) than the case studies (Figs. 2.4–2.7). In all four seasons, both models consistently exhibit smaller errors during the testing period compared to the training years (Fig. 2.8). However, we emphasize that the total number of heat waves in our study is relatively small, and the comparison between the testing and training years would need to be expanded to make broad conclusions. Finally, both GraphCast and Pangu exhibit consistent cold biases throughout all seasons except boreal winter, especially leading up to heat wave onset (Fig. 2.8). To that point, boreal winter has two unique findings among the four seasons: 1) Pangu has a warm bias prior to the heat wave, suggesting too early an onset, and 2) both AIWP models exhibit warm biases after the heat wave ends (Fig. 2.8), suggesting that they do not decay the heat wave fast enough. Future work should investigate the potential causes of warm biases being predominantly unique to winter.

As mentioned in section 2.2, the Pangu developers (Bi et al. 2023) introduce hierarchical temporal aggregation for medium-range forecasts and suggest using the 24-h model for longer lead times. The hierarchical temporal aggregation is intended to reduce the number of iterations required to train a series of models with longer lead times, thereby potentially reducing medium range forecast errors (Bi et al. 2023). There are four initialization times for the 24-h model outputs (00, 06, 12, and 18 UTC) and each forecast is instantaneous, not a daily average. In other

words, one must choose an initialization time and use the 24-h model from that time. Therefore, in this study we use the Pangu 6-h model, because UFS GEFS and GraphCast both output on a 6-h basis and an instantaneous value for a specific forecast hour would not be consistent with the uniform daily averages of UFS GEFS and GraphCast. Furthermore, Pasche et al. (2025) warn that most AIWP models use a large autoregressive time step (i.e., coarse temporal resolution) that can impact forecasts in which the daily maximum and/or minimum values of a forecast parameter are important, such as 2-m temperature during heat waves. Pasche et al. (2025) stipulate that even a 6-h time step can miss such peaks and valleys, lending even further credence to our choice not to focus on the Pangu 24-h model. Alternatively, we could average Pangu 24-h output at each of the four daily initialization times and create daily averages from those. However, doing so would not preserve consistency with how we create the daily averages (using 6-h output) of UFS GEFS and GraphCast, particularly for the 20-day forecast period we evaluate. Our results using the 6-h Pangu output (Figs. 2.3–2.8) show that Pangu errors largely exceed GraphCast errors across most cases, regions, and seasons. In addition, Pangu exhibits a "see-saw" effect, where the errors have large oscillations within a given 24-h period (see example in Fig. 2.3b compared to UFS GEFS and GraphCast. These oscillations are likely directly related to our choice to use the 6-h output. For more insight into the differences between the 6-h and 24-h Pangu output, Fig. A.2 shows results for the September 2019 Northwest heat wave.

In summary, extreme heat is the deadliest weather-related hazard in the United States, and is becoming more frequent, intense, and longer duration over time. While heat wave predictability has improved, forecast skill still lags on medium-range and S2S timescales, necessitating technological advances and novel approaches. This study investigates the abilities of two AIWP

models (GraphCast and Pangu) to forecast 2-m temperature extremes and associated heat waves across four CONUS regions on medium-range and subseasonal timescales. We also compare AIWP model temperature forecast skill to that of one traditional physics-based S2S NWP model (UFS GEFS). As AIWP models are relatively new and are not always able to accurately represent physical processes, investigating their forecast skill along with a comparison to a traditional NWP model is vital to understanding whether AIWP models can offer advantages in the predictability of extremes (e.g., heat waves). Our results demonstrate that GraphCast and Pangu show promise and skill advances by some measures relative to UFS GEFS. While more work is required to better understand the limitations and driving mechanisms of AIWP forecast skill, results such as those for GraphCast in this study offer promise that these new tools can result in a major leap in the medium-range and S2S predictability of heat waves and other extreme weather events.

Chapter 3

Conclusions

3.1 Key Takeaways

Our results show that the UFS GEFS and the two AIWP models frequently struggle to predict the onset and magnitude of heat waves in the four NCA regions that we examine. Cold biases are common throughout our results, especially before and during the heat wave period, suggesting that the models underestimate heat wave intensity. There are also frequent warm biases following the heat wave period, suggesting that the models do not decay the heat waves quickly enough.

Additionally, we emphasize that the two AIWP models do not exhibit the same skill. GraphCast consistently shows better skill than Pangu throughout our evaluations, regardless of region, lead time, or season. In addition, GraphCast often outperforms the dynamical NWP model (UFS GEFS), demonstrating that AIWP models offer promise in extreme heat prediction. The struggles of UFS GEFS are particularly prominent in regions with substantial complex terrain, such as the Northwest and Northeast. GraphCast also exhibits larger errors in the Northeast and Northwest than in the Southeast and Midwest (which have less complex topography). Finally, we note that the three models use vastly different architectures: UFS GEFS is a traditional dynamical NWP model, GraphCast uses a Graphical Neural Network (GNN) approach, and Pangu uses three-dimensional deep networks.

In summary, this study advances understanding of AIWP model utility for extreme heat forecasts. Because AIWP models are so new and yet rapidly gaining in popularity and usage, with some already operational (e.g., ECMWF 2025), it is vital to investigate how well they can

forecast extreme events (Pasche et al. 2025). Our comparison of the two AIWP models to a traditional NWP model is a key step in improving understanding of these models' operational forecast capabilities.

3.2 Study Limitations

Limitations to this study include the relatively small number of heat waves (15 per season) that we examine. This makes it difficult to definitively compare forecast skill between the testing and training periods and make broad conclusions about differences between those periods. In the future, more cases could be added to both the testing and training periods, especially as the testing period becomes larger over time. Another strategy to expand the number of cases would be to compare model forecast skill among short- (i.e., 3–4 days) and long-duration (i.e., > 5 days) heat waves. Additionally, heat waves in other CONUS NCA regions (i.e., Northern Great Plains, Southern Great Plains, Southwest) and/or other countries could be investigated.

Another limitation of this work is the UFS GEFS initialization strategy. While both AIWP models can be initialized manually at any date and time, the UFS GEFS is only run out to 35 days once a week. This makes it difficult to create an apples-to-apples comparison among all heat waves, since there is no guarantee that all heat waves will start at the same lead time relative to the most recent or useful UFS GEFS initialization date. Future work could incorporate ECMWF extended ensemble reforecasts (e.g., Switanek et al. 2023), which are initialized twice per week out to 46 days, and have data starting in 2000.

Finally, the two AIWP models examined in this study contain a limited set of surface variables: 2-m temperature, 10-m wind, and mean sea-level pressure. To fully investigate and understand the ability of AIWP models to predict extreme heat, it would be advantageous if the variable suites could expand to include e.g. humidity and soil moisture. For example, negative

soil moisture anomalies are a well-established precursor to extreme heat events, especially in arid regions (e.g., Benson and Dirmeyer 2023). While the UFS GEFS and ECMWF extended ensemble reforecasts both include soil moisture as a forecast variable, it would be interesting to investigate how well AIWP models can forecast and be trained on soil moisture anomalies once they become equipped with land-atmosphere coupling.

3.3 Future Work

An important avenue of future work is to investigate the ability of other AIWP models to predict heat waves for medium-range and S2S timescales. For example, we could examine FourCastNet (Pathak et al. 2022), NeuralGCM (Kochov et al. 2024), and Google GenCast (Price et al. 2024). Additionally, it would be useful to investigate the ability of AIWP models to predict the large-scale characteristics of heat waves above the surface, using existing model fields such as 500-hPa geopotential height, upper-air temperature, and specific humidity. Insight into large-scale characteristics could elucidate how well AIWP models forecast atmospheric blocking patterns and associated amplified Rossby wave trains. For example, the u-component of the wind near the top of the troposphere (e.g., 300 hPa) could provide insight into forecast skill of blocking patterns, because the jet stream slows substantially during such events (i.e., traffic jam theory; Nakamura and Huang 2018). Specific humidity would be useful to determine how well the AIWP models forecast air mass characteristics associated with heat waves, since specific humidity is not dependent on temperature. While many heat waves are dry, they can be predominantly humid in some warm and moist climates such as the Southeast US (e.g., Ennis and Milrad 2024). Examining specific humidity would allow us to assess how well AIWP models can differentiate between dry and humid heat waves.

Other potential future work could involve examining more heat waves over additional regions. For example, to gain a more robust understanding of warm-season model forecast skill of extreme heat, we could choose to investigate 100 total heat waves during boreal summer over all seven NCA regions (USGCRP 2023). We could then compare those results to a similar evaluation in other seasons, such as boreal winter. While the four NCA regions in this study have large differences among each other, adding the two Great Plains regions would provide additional insight into forecast skill over relatively uniform terrain, while adding the Southwest would provide events over the most arid and hottest region in the CONUS. It would also be valuable to expand the ideas presented in this project and examine heat wave forecast skill on the global scale.

Additional investigations should evaluate the ability of AIWP models to predict extreme heat events in a particular region across multiple timescales: short-range (e.g., 3 days), medium-range (7–10 days), and subseasonal (15–20 days). This would elucidate how AIWP model skill varies temporally and whether it degrades over time at the same rate as traditional NWP model skill.

Finally, future work should robustly examine forecast skill differences of extreme heat between the testing and training periods. Although this study finds slightly better skill for GraphCast and Pangu in the testing periods compared to the training periods, we have too few cases to make broad conclusions. A future investigation should stipulate an equal number of heat waves within the testing and training periods, and focus on one season (e.g., summer). Furthermore, we could train an AIWP model with and without strong heat waves and evaluate if the model can simulate extreme heat without first learning about it. A similar investigation for tropical cyclones is detailed by Sun et al. (2024).

3.4 Broader Impacts

There are still many unknowns regarding how AIWP models forecast extreme weather and climate events (e.g., Pasche et al. 2025). As extreme events occur more frequently, more validation studies of AIWP model forecasts are required. There is a particular lack of validation for AIWP model forecasts of extreme heat, especially systematic studies across more than a handful of cases. This study helps to fill that research gap and demonstrates that AIWP models can compete with traditional NWP models for extreme heat forecasts. However, more work is required to evaluate an even larger set of extreme heat events over wider forecast regions and using additional forecast (both NWP and AIWP) models.

As AIWP models increase in popularity and usage across the weather-climate enterprise, it is also vital that we better understand their strengths and faults. For example, many traditional NWP models have persistent well-known biases, which operational forecasters can account for in making their real-time assessments and forecast grids. However, the biases of AIWP models are not well understood, in large part because these models are so new. Therefore, this study is a first step toward better informing operational forecasters and extreme event stakeholders alike. In addition, our comparison to a traditional physics-based NWP model highlights when and where AIWP models can compete with and exceed traditional NWP model prediction skill for extreme heat events.

Finally, as extreme heat becomes more frequent, intense, and longer in a warming world, better forecasts are needed, particularly for longer lead times. While traditional NWP models exhibit a predictability plateau of around two weeks due to atmospheric chaos, little is known about the ability of AIWP models to make useful forecasts of extreme heat on medium-range and S2S timescales. If these cutting-edge tools can demonstrate more reliable predictive capabilities

than traditional NWP, especially for longer lead times, then society and industry will benefit from improved forecasts that better protect life and property from extreme heat.

REFERENCES

Alizadeh, M. R., J. T. Abatzoglou, J. F. Adamowski, J. P. Prestemon, B. Chittoori, A. A. Asanjan, and M. Sadegh, 2022: Increasing heat-stress inequality in a warming climate. *Earth's Future*, **10**, e2021EF002488, https://doi.org/10.1029/2021EF002488.

Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, https://doi.org/10.1126/science.aav7274.

Baker, L., A. Charlton-Perez, and K. L. Mattu, 2023: Skilful sub-seasonal forecasts of aggregated temperature over Europe. *Meteor. Appl.*, **30**, e2169, https://doi.org/10.1002/met.2169.

Ballester, J., and Coauthors, 2023: Heat-related mortality in Europe during the summer of 2022. *Nat. Med.*, **29**, 1857–1866, https://doi.org/10.1038/s41591-023-02419-z.

Barriopedro, D., R. García-Herrera, C. Ordóñez, D. G. Miralles, and S. Salcedo-Sanz, 2023: Heat waves: physical understanding and scientific challenges. *Rev. Geophys.*, **61**, e2022RG000780, https://doi.org/10.1029/2022RG000780.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bekris, Y., P. C. Loikith, and J. D. Neelin, 2023: Short warm distribution tails accelerate the increase of humid-heat extremes under global warming. *Geophys. Res. Lett.*, **50**, e2022GL102164, https://doi.org/10.1029/2022GL102164.

Benson, D. O., and P. A. Dirmeyer, 2023: The soil moisture–surface flux relationship as a factor for extreme heat predictability in subseasonal to seasonal forecasts. *J. Climate*, **35**, 6375–6392, https://doi.org/10.1175/JCLI-D-22-0447.1.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bitencourt, D. P., L. M. Alves, E. K. Shibuya, I. de Ângelo da Cunha, and J. P. E. de Souza, 2021: Climate change impacts on heat stress in Brazil—Past, present, and future implications for occupational heat exposure. *Int. J. Climatol.*, **41**, E2741–E2756, https://doi.org/10.1002/joc.6877.

Bonavita, M., 2024: On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.*, **51**, e2023GL107377, https://doi.org/10.1029/2023GL107377.

Bouallègue, Z. B., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bull. Amer. Meteor. Soc.*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Cahill, J., E. A. Barnes, E. D. Maloney, S. R. Sain, P. Harr, and L. Madaus, 2024: Errors of opportunity: Using neural networks to predict errors in the Global Ensemble Forecast System (GEFS) on S2S time scales. *Wea. Forecasting*, **39**, 1817–1831, https://doi.org/10.1175/WAF-D-23-0125.1.

Camps-Valls, G., and Coauthors, 2025: Artificial intelligence for modeling and understanding extreme weather and climate events. *Nat. Commun.*, **16**, 1919, https://doi.org/10.1038/s41467-025-56573-8.

Choi, N., and C. Stan, 2025: Large-scale surface air temperature bias in summer over the CONUS and its relationship to Tropical Central Pacific convection in the UFS Prototype 8. *J. Climate,* **38**, 117–129, https://doi.org/10.1175/JCLI-D-24-0078.1.

Clarke, B., F. Otto, R. Stuart-Smith, and L. Harrington, 2022: Extreme weather impacts of climate change: an attribution perspective. *Env. Res. Climate*, **1**, 012001, https://doi.org/10.1088/2752-5295/ac6e7d.

Cloutier-Bisbee, S. R., A. Raghavendra, and S. M. Milrad, 2019: Heatwaves in Florida: Climatology, trends, and related precipitation events. *J. Appl. Meteor. Climatol.*, **58**, 447–466, https://doi.org/10.1175/JAMC-D-18-0165.1.

Davini, P., A. Weisheimer, M. Balmaseda, S. J. Johnson, F. Molteni, C. D. Roberts, R. Senan, and T. N. Stockdale, 2021: The representation of winter Northern Hemisphere atmospheric blocking in ECMWF seasonal prediction systems. *Quart. J. Roy. Meteor. Soc.*, **147**, 1344–1363, https://doi.org/10.1002/qj.3974.

DeFlorio, M. J., and Coauthors, 2019: Experimental subseasonal-to-seasonal (S2S) forecasting of atmospheric rivers over the Western United States. *J. Geophys. Res. Atmos.*, **124**, 11242–11265, https://doi.org/10.1029/2019JD031200.

Domeisen, D. I. V., and Coauthors, 2023: Prediction and projection of heatwaves. *Nat. Rev. Earth Environ.*, **4**, 36–50, https://doi.org/10.1038/s43017-022-00371-z.

ECMWF, 2020: Heatwaves and warm spells. Accessed 12 December 2024, https://climate.copernicus.eu/esotc/2020/heatwaves-and-warm-spells-during-2020.

ECMWF, 2025: ECMWF's AI forecasts become operational. Accessed 10 April 2025, https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational.

Ennis, K. E., and S. M. Milrad, 2024*:* Man, it's a hot one: Trends and extremes in Florida autumn heat stress. *Int. J. Climatol.*, **44**, 1816–1830, https://doi.org/10.1002/joc.8415.

Ford, T. W., P. A. Dirmeyer, and D. O. Benson, 2018: Evaluation of heat wave forecasts seamlessly across subseasonal timescales. *NPJ Climate Atmos. Sci.*, **1**, 20, https://doi.org/10.1038/s41612-018-0027-7.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hakim, G. J., and S. Masanam, 2024: Dynamical tests of a deep learning weather prediction model. *Artif. I. Earth Sys.*, **3**, 1–11, https://doi.org/10.1175/AIES-D-23-0090.1.

IPCC, 2022: AR6 Climatic Change 2021: The Physical Science Basis. IPCC, 82 pp.
Jiménez-Esteve, B., and D. Domeisen, 2022: The role of atmospheric dynamics and large-scale topography in driving heatwaves. *Quart. J. Roy. Meteor. Soc.*, **148**, 2344–2367, https://doi.org/10.1002/qj.4306.

Keellings, D., and H. Moradkhani, 2020: Spatiotemporal evolution of heat wave severity and coverage across the United States. *Geophys. Res. Lett.*, **47**, e2020GL087097, https://doi.org/10.1029/2020GL087097.

Khatana, S. A. M., J. J. Szeto, L. A. Eberly, A. S. Nathan, J. Puvvula, and A. Chen, 2024: Projections of extreme temperature-related deaths in the US. *JAMA Netw. Open.*, **7**, e2434942, https://doi.org/10.1001/jamanetworkopen.2024.34942.

Klemm, T., and R. A. McPherson, 2017: The development of seasonal climate forecasting for agricultural producers. *Agric. For. Meteor.*, **232**, 384–399, https://doi.org/10.1016/j.agrformet.2016.09.005.

Kochov, D., and Coauthors, 2024: Neural general circulation models for weather and climate. *Nature*, **632**, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y.

Kornhuber, K., D. Coumou, E. Vogel, C. Lesk, J. F. Donges, J. Lehmann, and R. M. Horton, 2020: Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nat. Climate Change*, ***10***, 48–53, https://doi.org/10.1038/s41558-019-0637-z.

Krishnamurthy, V., and C. Stan, 2022: Prediction of extreme events in precipitation and temperature over CONUS during boreal summer in the UFS coupled model. *Climate Dyn.*, **59**, 109–125, https://doi.org/10.1007/s00382-021-06120-0.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lavaysse, C., G. Naumann, L. Alfieri, P. Salamon, and J. Vogt, 2019: Predictability of the European heat and cold waves. *Climate Dyn.*, **52**, 2481–2495, https://doi.org/10.1007/s00382-018-4273-5.

Lembo, V., and Coauthors, 2024: Dynamics, statistics, and predictability of Rossby waves, heat waves, and spatially compounding extreme events. *Bull. Amer. Meteor. Soc.*, **105**, E2283–E2293, https://doi.org/10.1175/BAMS-D-24-0145.1.

Li, P., Y. Yu, D. Huang, Z-H. Wang, and A. Sharma, 2023: Regional heatwave prediction using graph neural network and weather station data. *Geophys. Res. Lett.*, **50**, e2023GL103405, https://doi.org/10.1029/2023GL103405.

Limaye, V. S., J. Vargo, M. Harkey, T. Holloway, and J. A. Patz, 2018: Climate change and heat-related excess mortality in the eastern USA. *EcoHealth*, **15**, 485–496, https://doi.org/10.1007/s10393-018-1363-0.

Lin, H., R. Mo, and F. Vitart, 2022: The 2021 Western North American heatwave and its subseasonal predictions. *Geophys. Res. Lett.*, **49**, e2021GL097036, https://doi.org/10.1029/2021GL097036.

Lopez-Gomez, I., A. McGovern, S. Agrawal, and J. Hickey, 2023: Global extreme heat forecasting using neural weather models. *Artif. I. Earth Syst.*, **2**, 1–28, https://doi.org/10.1175/AIES-D-22-0035.1.

Lupo, A. R., 2021: Atmospheric blocking events: a review. *Ann. N.Y. Acad. Sci.*, **1504**, 5–24, https://doi.org/10.1111/nyas.14557.

Matthews, T., C. Raymond, J. Foster, J. W. Baldwin, C. Ivanovich, Q. Kong, P. Kinney, and R. M. Horton, 2025: Mortality impacts of the most extreme heat events. *Nat. Rev. Earth Environ.*, **6**, 193–210, https://doi.org/10.1038/s43017-024-00635-w.

McAllister, C., A. Stephens, and S. M. Milrad, 2022: The heat is on: Observations and trends of heat stress metrics during Florida summers**.** *J. Appl. Meteor. Climatol.,* **61**, 277–296, https://doi.org/10.1175/JAMC-D-21-0113.1.

Milrad, S. M., and K. E. Ennis, 2025: Assessing summer humid heat in Europe: Trends, extremes, and drivers. *Int. J. Climatol.*, in press.

Mora, C., and Coauthors, 2017: Global risk of deadly heat. *Nat. Climate Change*, *7*, 501–506, https://doi.org/10.1038/nclimate3322.

Nakamura, N., and C. S. Y. Huang, 2018: Atmospheric blocking as a traffic jam in the jet stream. *Science*, **361**, 42–47, https://doi.org/10.1126/science.aat0721.

NWS, 2025: Weather related fatality and injury statistics. Accessed 1 April 2025, https://www.weather.gov/hazstat/.

Paradise, A., C. B. Rocha, P. Barpanda, and N. Nakamura, 2019: Blocking statistics in a varying climate: Lessons from a "traffic jam" model with pseudostochastic forcing. *J. Atmos. Sci.*, **76**, 3013–3027, https://doi.org/10.1175/JAS-D-19-0095.1.

Pasche, O. C., J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, 2025: Validating deep learning weather forecast models on recent high-impact extreme events. *Artif. I. Earth Syst.*, **4**, 1–24, https://doi.org/10.1175/AIES-D-24-0033.1.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint*, https://doi.org/10.48550/arXiv.2202.11214.

Patterson, M., 2023: North-west Europe hottest days are warming twice as fast as mean summer days. *Geophys. Res. Lett.*, **50**, e2023GL102757, https://doi.org/10.1029/2023GL102757.

Perkins, S. E., 2015: A review on the scientific understanding of heat waves—Their measurement, driving mechanisms, and changes at the global scale. *Atmos. Res.*, **164–165**, 242–267, https://doi.org/10.1016/j.atmosres.2015.05.014.

Price, I., and Coauthors, 2024: GenCast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint*, https://doi.org/10.48550/arXiv.2312.1579.

Radford, J. T., I. Emme-Uphoff, J. Q. Stewart, K. D. Musgrave, R. DeMaria, N. Tourville, and K. Hilburn, 2025: Accelerating community-wide evaluation of AI models for global weather prediction by facilitating access to model output. *Bull. Amer. Meteor. Soc.*, **106**, E68–E76, https://doi.org/10.1175/BAMS-D-24-0057.1.

Rastogi, D., F. Lehner, T. Kuruganti, K. J. Evan, K. Kurte, and J. Sanyal, 2021: The role of humidity in determining future electricity demand in the southeastern United States. *Environ. Res. Lett.*, **16**, 114017, https://doi.org/10.1088/1748-9326/ac2fdf.

Raymond, C., T. Matthews, and R. M. Horton, 2020: The emergence of heat and humidity too severe for human tolerance. *Sci. Adv.*, **6**, eaaw1838, https://doi.org/ 10.1126/sciadv.aaw1838.

Rennie, J. J., M. A. Palecki, S. P. Heuser, and H. J. Diamond, 2021: Developing and validating

heat exposure products using the US climate reference network. *J. Appl. Meteor. Climatol.*, **60**, 543–558, https://doi.org/10.1175/JAMC-D-20-0282.1.

Schumacher, D. L., M. Hauser, and S. I. Seneviratne, 2022: Drivers and mechanisms of the 2021 Pacific Northwest heatwave. *Earth's Futur.*, **10**, e2022EF002967, https://doi.org/10.1029/2022EF002967.

Schwarz, L., and Coauthors, 2020: The health burden of fall, winter and spring extreme heat events in Southern California and contribution of Santa Ana Winds. *Environ. Res. Lett.*, **15**, 054017, https://doi.org/10.1088/1748-9326/ab7f0e.

Selz, T., and G. C. Craig, 2023: Can artificial intelligence-based weather prediction models simulate the butterfly effect. *Geophys. Res. Lett.*, **50**, e2023GL105747, https://doi.org/10.1029/2023GL105747.

Seo, E., P. A. Dirmeyer, M. Barlage, H. Wei, and M. Ek, 2024: Evaluation of land-atmosphere coupling processes and climatological bias in the UFS global coupled model. *J. Hydrometeor.*, **25**, 161–175, https://doi.org/10.1175/JHM-D-23-0097.1.

Stan, C., and Coauthors, 2022: Advances in the prediction of MJO teleconnections in the S2S forecast systems. *Bull. Amer. Meteor. Soc.*, **103**, E1426–E1447, https://doi.org/10.1175/BAMS-D-21-0130.1.

Stan, C., and Coauthors, 2023: The impact of Tropical Pacific SST biases on the S2S forecast skill over North America in the UFS global coupled model. *J. Climate*, **36**, 2439–2456, https://doi.org/10.1175/JCLI-D-22-0196.1.

Stefanova, L., and Coauthors, 2022: Description and results from UFS coupled prototypes for future global, ensemble and seasonal forecasts at NCEP. NCEP Office Note 510, 201 pp, https://doi.org/10.25923/knxm-kz26.

Stone Jr., B., and Coauthors, 2023: How blackouts during heat waves amplify mortality and morbidity risk. *Environ. Sci. Technol.*, **57**, 8245–8255, https://doi.org/10.1021/acs.est.2c09588.

Sun, Y. Q., P. Hassanzadeh, M. Zand, A. Chattopadhyay, J. Weare, and D. S. Abbott, 2024: Can AI weather models predict out-of-distribution gray swan tropical cyclones? *arXiv preprint*, https://arxiv.org/html/2410.14932.

Sutanto, S. J., S. B. Zarzoza Mora, I. Supit, and M. Wang, 2024: Compound and cascading droughts and heatwaves decrease maize yields by nearly half in Sinaloa, Mexico. *NPJ Nat. Hazards*, **1**, 26, https://doi.org/10.1038/s44304-024-00026-7.

Switanek, M. B., T. M. Hamill, L. N. Long, and M. Scheuerer, 2023: Predicting subseasonal tropical cyclone activity using NOAA and ECMWF reforecasts. *Wea. Forecasting*, **38**, 357–370, https://doi.org/10.1175/WAF-D-22-0124.1.

Tak, S., E. Seo, P. A. Dirmeyer, and M-I. Lee, 2024: The role of soil moisture-temperature coupling for the 2018 Northern European heatwave in a subseasonal forecast. *Wea. Climate Extrem.*, **44**, 100670, https://doi.org/10.1016/j.wace.2024.100670.

Teskey, R., T. Wertin, I. Bauweraerts, M. Ameye, M. A. McGuire, and K. Steppe, 2015: Responses of tree species to heat waves and extreme heat events. *Plant Cell Environ.*, **38**, 1699–1712, https://doi.org/10.1111/pce.12417.

Thompson, V., and Coauthors, 2022: The 2021 western North America heat wave among the most extreme events ever recorded globally. *Sci. Adv.*, **8**, eabm6860, https://doi.org/10.1126/sciadv.abm6860.

Thornton, P., G. Nelson, D. Mayberry, and M. Herrero, 2021: Increases in extreme heat stress in domesticated livestock species during the twenty-first century. *Glob. Chang. Biol.*, **27**, 5762–5772, https://doi.org/10.1111/gcb.15825.

Ullrich, P. A., and Coauthors, 2025: Recommendations for comprehensive and independent evaluation of machine learning-based earth system models. *arXiv preprint*, https://arxiv.org/abs/2410.19882v2.

USGCRP, 2023: Fifth National Climate Assessment. U.S. Global Change Research Program, 1834 pp., https://doi.org/10.7930/NCA5.2023.

Vicedo-Cabrera, A.M., and Coauthors, 2021: The burden of heat-related mortality attributable to recent human-induced climate change. *Nat. Climate Change*, **11**, 492–500, https://doi.org/10.1038/s41558-021-01058-x.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *NPJ Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

Vogel, J., J. Hess, Z. Kearl, K. Naismith, K. Bumbaco, B. G. Henning, R. Cunningham, and N. Bond, 2023: In the hot seat: saving lives from extreme heat in Washington State. University of Washington Climate Impacts Group, 24 pp., https://cig.uw.edu/wp-content/uploads/sites/2/2023/06/CIG-Report-Heat-202-pages.pdf.

Waqas, M., U. W. Humphries, B. Chueasa, and A. Wangwongchai, 2024: Artificial intelligence and numerical weather prediction models: A technical survey. *Nat. Hazards Res.*, in press, https://doi.org/10.1016/j.nhres.2024.11.004.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

White, C. J., and Coauthors, 2022: Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Amer. Meteor. Soc.*, **103**, E1448–E1472, https://doi.org/10.1175/BAMS-D-20-0224.1.

White, R. C., and Coauthors, 2023: The unprecedented Pacific Northwest heatwave of June 2021. *Nat. Commun.*, **14**, 727, https://doi.org/10.1038/s41467-023-36289-3.

World Health Organization (WHO), 2024: Heat and health. Accessed 1 April 2025, https://www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health.

Wu, C., and Coauthors, 2022: Heat adaptive capacity: What causes the differences between residents of Xiamen Island and other areas? *Front. Public Health*, **10**, 799365, https://doi.org/10.3389/fpubh.2022.799365.

Xie, J. P-C. Hsu, Y. Hu, H. Zhang, and M. Ye, 2024: Advancing subseasonal surface air temperature and heat wave prediction skill in China by incorporating scale interaction in a deep learning model. *Geophys. Res. Lett.*, **51**, e2024GL111076, https://doi.org/10.1029/2024GL111076.

Yan, X., L. Wang, E. P. Gerber, V. Castañeda, and K. Y. Ho, 2024: Traffic bottlenecks: Predicting atmospheric blocking with a diminishing flow capacity. *Geophys. Res. Lett.*, **51**, e2024GL111035, https://doi.org/10.1029/2024GL111035.

Zachariah, M., and Coauthors, 2022: Without human-caused climate change temperatures of 40 °C in the UK would have been extremely unlikely. World Weather Attribution, 26 pp., https://www.worldweatherattribution.org/wp-content/uploads/UK-heat-scientific-report.pdf

Zhang, W., B. Xiang, K-C. Tsena, N. C. Johnson, L. Harris, T. Delworth, and B. Kirtman, 2024: Subseasonal-to-seasonal (S2S) prediction of atmospheric rivers in the Northern Winter. *npj Climate Atmos. Sci.*, **7**, 275, https://doi.org/10.1038/s41612-024-00827-7.

Zhao, Q., and Coauthors, 2021: Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: A three-stage modelling study. *Lancet Planet. Health*, **5**, e415–e425.

Zheng, C., H. Kim, E. LaJoie, S. He., and E. K. M. Chang, 2025: Improving statistical prediction of subseasonal CONUS precipitation based on ENSO and the MJO by training with large ensemble climate simulations. *Geophys. Res. Lett.*, **52**, e2024GL110925, https://doi.org/10.1029/2024GL110925.

Zhou, J., A. J. Teuling, S. I. Seneviratne, and A. L. Hirsch, 2024: Soil moisture-temperature coupling increases population exposure to future heatwaves. *Earth's Future*, **12**, e2024EF004697, https://doi.org/10.1029/2024EF004697.

Zhuang, J., and Coauthors, 2024: xESMF: Universal regridder for geospatial data. Zenodo, https://doi.org/10.5281/zenodo.4294774.

Appendix A

Supplementary Figures

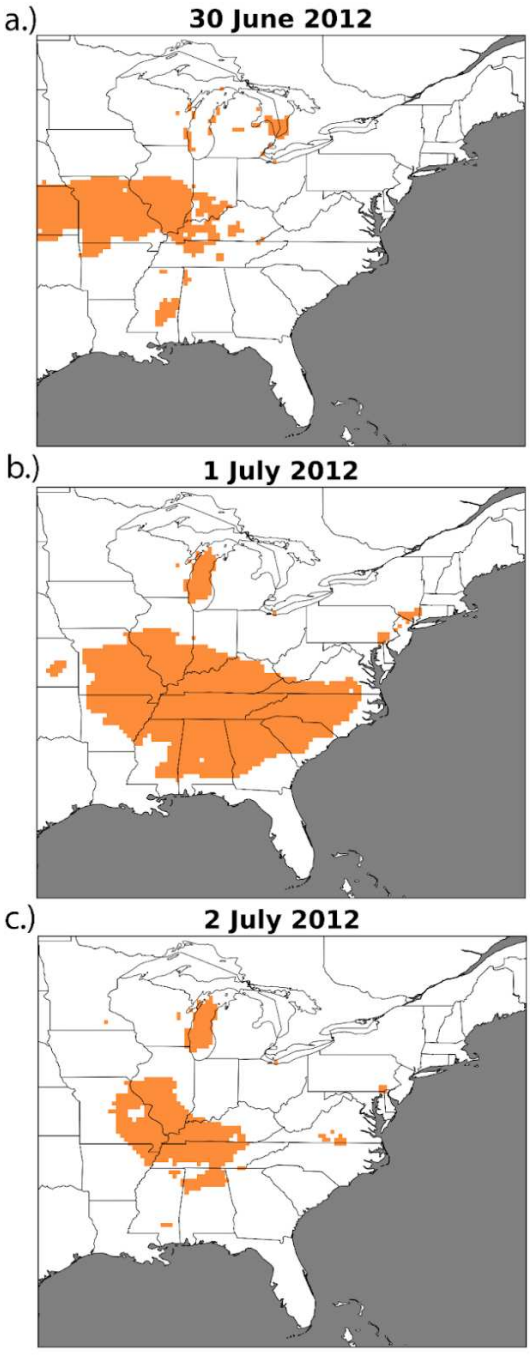

Figure A.1: For the 2012 heat wave that impacted both the Midwest and Southeast NCA regions, the orange shading indicates areas where daily mean 2-m temperatures exceeded the 95th percentile on (a) 30 June and (b) 1 July, and (c) 2 July.

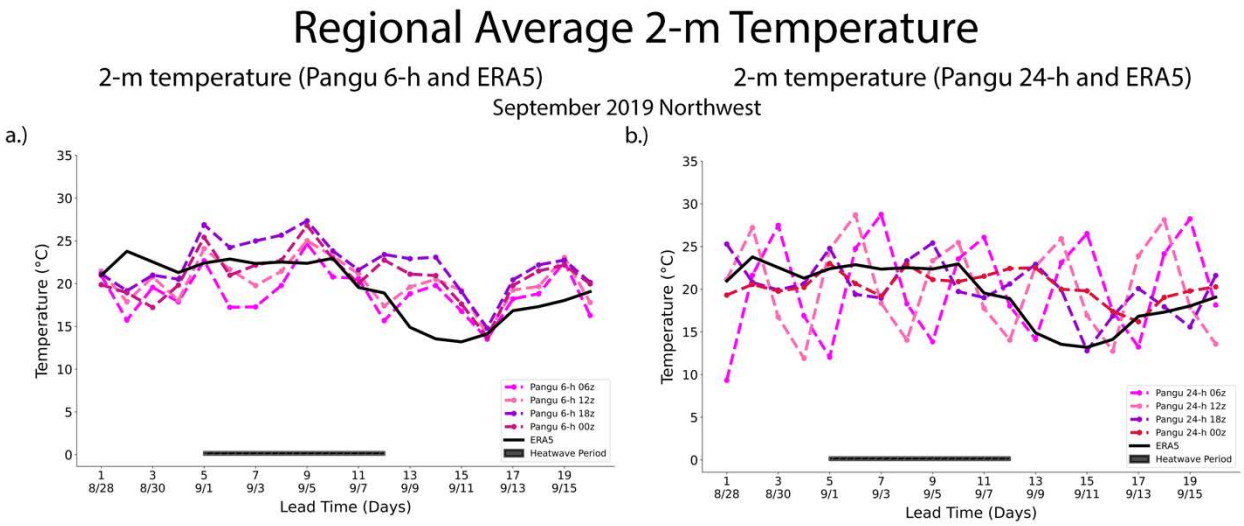

Figure A.2: For the September 2019 Northwest heat wave, regional average 2-m temperature (°C) for (a) four runs (each six hours apart) of the Pangu 6-h model and (b) four runs (each six hours apart) of the Pangu 24-h model. In each panel, the 00 UTC run is plotted in dashed red, the 06 UTC run in dashed magenta, the 12 UTC run in dashed pink, and the 18 UTC run in dashed purple. In both panels, ERA5 2-m temperatures (°C, solid black lines) are plotted as truth, and the heat wave period is illustrated by the black line at the bottom.