The 2021 Pacific Northwest Heat Wave: meteorological interpretation of forecast uncertainties in data-driven and physics-based ensembles

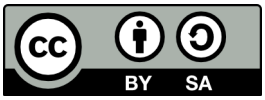
Master's Thesis in
Meteorology and Climate Physics
by

Yangfan Zhou

June 2024

INSTITUTE OF METEOROLOGY AND CLIMATE RESEARCH
KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Supervisor:	Prof. Dr. Peter Knippertz
Co-supervisor:	Dr. Julian Quinting



Abstract

Heat waves are highly impactful extreme events with significant health, social, and economic effects. Accurate prediction of the timing and intensity of heat waves is crucial for effective preparedness. Recently, data-driven weather prediction models based on machine learning techniques have demonstrated performance comparable to physics-based Numerical Weather Prediction (NWP) models in medium-range weather forecasting. However, their ability to predict extreme events remains uncertain due to the rare occurrence of such events in training data. Moreover, it remains uncertain whether data-driven models can learn the underlying physical mechanisms during extreme events. Based on this, the thesis aims to evaluate the predictive skill of data-driven models in heat wave prediction and interpret their predictions in the context of the underlying physical processes they represent, using the record-breaking 2021 Pacific Northwest Heat Wave as a case study.

This study compares three data-driven weather prediction models (FourCastNet1, FourCastNet2, and Pangu-Weather) with the state-of-the-art NWP model of Integrated Forecasting System (IFS). Ensemble forecasts are generated for the two versions of FourCastNet with perturbations to initial conditions and compared with the IFS ensemble forecast (ENS). Additionally, deterministic forecasts from Pangu-Weather and FourCastNet2 are included for comparison with IFS high-resolution forecast (HRES).

The performance of these models in predicting peak magnitude during the 2021 Pacific Northwest Heat Wave and associated anomalous large-scale circulation patterns was evaluated. FourCastNet2 and IFS ensembles skillfully predicted the peak magnitude of the heat wave at a lead time of 7 days. Pangu-Weather showed skillful prediction 8 days ahead but had the largest errors at a lead time of 10 days. FourCastNet1 only captured the peak 5 days ahead. Regarding anomalous large-scale circulation patterns, IFS ensembles showed skill 8 days ahead, while FourCastNet2 had skillful predictions 7 days ahead, and Pangu-Weather demonstrated skill as early as 9 days but almost had no skill at a lead time of 10 days. FourCastNet1 struggled to accurately represent the circulation pattern even at short lead times.

Analyzing the different member groups of FourCastNet2 ensembles based on their predictive skill of the heat wave suggests that FourCastNet2 might effectively learn the link between the anomalous large-scale circulation patterns and high-temperature anomalies during the heat wave. Vertical temperature and humidity profiles indirectly investigate local processes involving land-atmosphere feedback and upper-tropospheric heat. Compared to the data-driven models, IFS exhibits more robust forecasts across lead times, showing earlier emergence of upper-tropospheric heat and subsequent lower-level heat development. Notably, both Pangu-Weather and FourCastNet2 experienced an overestimation of low-level atmospheric moisture, while IFS did not exhibit any. This was further indicated by the dampened diurnal evolution of near-surface air temperature of data-driven models compared to IFS.

The results highlight that while data-driven models showed promising performance in predicting heat wave magnitude over larger regions and representing large-scale circulation patterns, they exhibit uncertainty in representing local thermodynamical processes. Since this study is based on a case study level, a more detailed examination of systematic errors in data-driven models is needed for future evaluation. Additionally, sensitivity tests are necessary to understand whether data-driven models can learn the complex dynamics and feedbacks.

Contents

1 Introduction

Heat waves are one of the most impactful natural hazards, described as events with excessively high temperatures that continue for several consecutive days. As extreme heat typically affects human health, ecosystems, and infrastructure, it garners significant public interest. Except for its direct high impact, the associated risk of wildfire and drought also captures significant attention (Domeisen et al., 2022b). With higher future global warming levels, changes in hot temperature extremes will be more significant in frequency and intensity (IPCC, 2021). Given these huge impacts and the increasing frequency and magnitude of heat waves, there is a growing societal and political demand for accurate forecasting of the timing and intensity of such events across a wide range of lead time (Barriopedro et al., 2023).

A recent record-shattering extreme heat event was the 2021 Pacific Northwest Heat Wave. From 25 June to 1 July 2021, the Pacific Northwest region of Canada and the United States experienced an unprecedented and extremely severe heat wave. Compared to the 1981-2021 climatology, near-surface air temperature anomalies soared to extreme highs of 16-20°C (White et al., 2023). On 29 June, the village of Lytton in Canada set a new national temperature record of 49.6°C, surpassing the previous record by an extraordinary 4.6°C. This temperature was reportedly the highest ever recorded north of 45° latitude worldwide (Environment and Climate Change Canada, 2022). This heat wave resulted in an estimated 740 excess deaths in the province of British Columbia due to heat-related issues (Henderson et al., 2022). While several subseasonal-to-seasonal prediction models captured this above-normal temperature signal with a lead time of 2–3 weeks (Lin et al., 2022; Emerton et al., 2022), the unprecedented magnitude of the heat wave was only accurately predicted by state-of-the-art numerical weather prediction models within a week (Lin et al., 2022; Emerton et al., 2022; Oertel et al., 2023). This relatively short lead time is due to complex interactions between a chain of synoptic events, causing a predictability barrier (Oertel et al., 2023).

The high surface temperature anomaly during the 2021 Pacific Northwest (PNW) heat wave is strongly linked to a high-amplitude upper-level ridge, also known as a "heat dome," causing high surface temperature anomaly by advection, subsidence, and clear-sky conditions (Zschenderlein et al., 2020; Neal et al., 2022). Previous studies have highlighted the significant role of upwind latent heating, concentrated in two separate Warm Conveyor Belts (WCBs) from the tropical West Pacific, in initiating and maintaining this upper-level ridge during the event (Oertel et al., 2023; Neal et al., 2022). This process was also a key driver of mid- to upper-tropospheric heat, with positive temperature anomalies aloft being crucial for surface heat accumulation by suppressing moist convection (Zhang and Boos, 2023; Neal et al., 2022). Additionally, desiccated soils and mountainous terrain in this region facilitated the establishment of deep atmospheric boundary

layers, allowing this warm aloft in the mid-to-upper-troposphere above the heat wave region to directly contribute to surface temperatures through mixing into the deep atmospheric boundary layer. Further, the dry soil conditions also acted to amplify surface heating through reducing evaporative cooling.(Schumacher et al., 2022).

NWP models is the dominant method for medium-range weather forecasting, relying on solving governing equations to predict future weather states. Over the past four decades, the performance of NWP models has improved due to advancements in science and technology, including better representation of unresolved processes, ensemble methods, objective analysis techniques, and increased supercomputing power (Bauer et al., 2015). However, these advancements have also introduced challenges for future NWP models. Specifically, the tremendous computational resources required for further improvements present new constraints and challenges for the development of future NWP models.

Recently, data-driven weather forecasting models based on machine learning methods have demonstrated significant improvement in weather prediction. Since 2022, a series of data-driven models have been developed, representing significant advancements in the field (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2022; Nguyen et al., 2023; Chen et al., 2023a; Chen et al., 2023c; Chen et al., 2023b; Lessig et al., 2023; Price et al., Price et al.). Contrary to NWP models based on physical laws governing atmospheric processes, these recently developed data-driven weather forecasting models trained with ERA5 reanalysis (Hersbach, 2023) directly learn patterns and relationships statistically. While these models often require significant computational resources during training, trained models can deliver rapid predictions during inference, thus reducing computational costs significantly (de Burgh-Day and Leeuwenburg, 2023). These models focus on weather forecasting tasks from the medium to sub-seasonal range and have demonstrated impressive and comparable forecast scores to outperform deterministic state-of-the-art numerical weather prediction model (Ben Bouallègue et al., 2024; Rasp et al., 2023). All these features make data-driven weather forecasting models a compelling alternative to traditional numerical methods.

However, these data-driven models also bring potential risks, particularly concerning extreme events. The data-driven models are most trained with ERA5 reanalysis data from 1979, and a limited training period may result in undersampling extreme events (Ebert-Uphoff and Hilburn, 2024). Additionally, these models are often optimized using local error measurements averaged over large regions, potentially compromising the prediction of specific meteorological features related to extreme events and introducing validation bias (McGovern et al., 2024).

Given the significant potential benefits and risks associated with data-driven weather prediction models, it is crucial to urgently evaluate these models, especially for forecasting extreme events. Proper interpretation is essential to ensure they can safely and effectively meet public needs. (Ebert-Uphoff and Hilburn, 2024). Several attempts have been made to evaluate the performance of data-driven models in predicting extreme events, including extreme temperatures, wind speed extremes, tropical cyclones, and storm systems (Ben Bouallègue et al., 2024; Lam et al., 2022; Charlton-Perez et al., 2024; Pasche et al., 2024; Olivetti and Messori, 2024). Most evaluation studies focus on comparing threshold exceedances between model predictions and validation across

large regions (Ben Bouallègue et al., 2024; Lam et al., 2022; Olivetti and Messori, 2024). Although some studies have compared the predictive skill of deterministic forecasts for standard variables on a case-study level (Charlton-Perez et al., 2024; Pasche et al., 2024), little attention has been given to the link between model performance and specific mechanisms. Consequently, it remains unclear whether a model's predictive skill for extreme events is associated with accurately representing relevant processes. A concern with data-driven models is that they may not include all relevant variables or adequately capture dependencies between predicted variables, leading to forecast uncertainty (Ebert-Uphoff and Hilburn, 2024).

To address the overarching questions, this thesis aims to evaluate the predictive skill of data-driven models for heat wave prediction and examine the connection between their predictions and the underlying physical processes they represent. Given its exceptional nature, the 2021 PNW Heat Wave is chosen as the case study, which presents a significant test for data-driven models. A detailed case study of this event will be conducted to address the following research questions:

1. **At what lead time do data-driven models start providing skillful predictions of the intensity for the peak of the heat wave and the associated anomalous atmospheric circulation pattern, and how do they compare with numerical weather prediction models?**

2. **To what extent can data-driven models capture the relationship between extreme temperature anomalies and the associated anomalous large-scale atmospheric circulation patterns?**

3. **How do data-driven models represent the local thermodynamical processes during the heat wave?**

The outline of this Master thesis is as follows: Chapter 2 begins by laying out the theoretical background of this thesis. The data and method used are introduced in Chapter 3. Chapter 4 presents the evaluation of the forecast of data-driven models and compares it with the conventional numerical weather prediction model at the peak of the selected heat wave. Chapter 5 discusses and presents the large-scale circulation patterns and processes during the heat wave and their representation in data-driven models. The final chapter (Chapter 6) summarizes the results, answers the research questions, and discusses the implications of the findings.

2 Theoretical Background

This chapter begins by providing background information on NWP and data-driven weather prediction. Section 2.1.1 introduces deterministic numerical weather prediction. Since understanding and quantifying forecast uncertainty is crucial for the extreme weather forecasting explored in this thesis, a significant focus is placed on ensemble forecasts. Therefore, the concept of ensemble forecasts will be introduced in the subsequent section (Section 2.1.2). Section 2.2 covers the development and outlook of data-driven weather prediction. Additionally, gaining insights into the important drivers and feedbacks influencing heatwave development is addressed in Section 2.3.1. Finally, the chapter concludes with an overview of the 2021 Pacific Northwest Heat Wave in Section 2.3.2.

2.1 Numerical Weather Prediction

2.1.1 Deterministic numerical weather prediction

The development of NWP models can be traced back to the early 1900s when Bjerknes (1904) first stated that predicting future atmospheric state is an initial-value problem. To predict the future state of the atmosphere, a full set of partial differential equations that govern atmospheric motion and evolution must be integrated into the next time step. These equations include the Navier-Stokes equations for fluid motion, the mass continuity equation, the first law of thermodynamics, and the ideal gas law. Since there is no analytical solution for these partial differential equations, numerical methods are used to achieve numerical integration in discrete grid space. However, the governing equations do not resolve all processes, independent of whether the equations are solved analytically or numerically. The physical processes on unresolved scales are incorporated into the equations for the resolved scales through source terms representing mass, momentum, and heat. Due to the typically unresolved nature of these processes, the physical processes on unresolved scales must be parameterized to describe their interaction with the resolved scales (Bauer et al., 2015).

According to this approach, two main challenges needed to be addressed to make accurate predictions: obtaining accurate initial conditions and understanding the laws governing atmospheric behavior (Pu and Kalnay, 2018). Over the next century, continuous and rapid developments focused on solving these two problems and transforming NWP from a proposition into an operationally practical tool. To obtain a more realistic initial state of the atmosphere, observing systems were developed. In parallel, data assimilation techniques were developed to fill the gap between incomplete observations and the required initial field, which allows for the integration of observational data

with short-range weather forecasts to produce more accurate initial conditions possible (Al-Yahyai et al., 2010).

Despite these continuing advancements in NWP, the predictability of the future atmospheric state is still limited by the incorrect representation of physical processes, parameterizations in the models, and the uncertainty from initial conditions. The ability to make predictions based on the currently available information and models is referred to as practical predictability. However, even with perfect models and initial conditions, there is still a theoretical limit to the predictability of the atmospheric state, which is known as intrinsic predictability. Due to the chaotic nature of the atmosphere (Lorenz, 1963), even infinitesimally small initial errors in a deterministic, nonlinear system like the atmosphere can lead to vastly different outcomes over time; this is also referred to as the "butterfly effect." Thus, a single deterministic forecast can only provide a prediction within a certain range and is unable to give quantitative information about the future atmospheric state. As a result, ensemble forecasting has been developed to provide quantitatively reliable information. This is also the direction in which NWP has shifted over the past 25 years, as stated by Buizza and Leutbecher (2015).

2.1.2 Ensemble forecasts

NWP inherently deals with uncertainties arising from two main sources: initial condition errors and model errors (Bauer et al., 2015). Despite the development of a relatively comprehensive observing system and data assimilation techniques, initial conditions for NWP models can only be estimated with finite accuracy. As a result, NWP models always start calculating forecasts from an atmospheric initial state that differs from the truth, and even a small error in the initial state will grow significantly with lead time. Due to the inherent non-linear nature of the governing equations, the growth of initial error is flow-dependent. Moreover, the models themselves are not perfect. The numerical representation of atmospheric dynamics and physics introduces model uncertainties related to factors such as the truncation of the equations of motion and the parameterization of sub-grid scale processes like cumulus convection. These two types of errors cannot be separated, as the initial conditions are estimated using a forecast model, meaning that initial condition errors are influenced by model deficiencies (Leutbecher and Palmer, 2008).

Ensemble forecasting offers a practical approach to addressing the complexity of stochastic dynamic equations by approximating them using the Monte-Carlo method (Leith, 1974). This method randomly sampled a finite number of points from the probability distribution representing the uncertainty in the atmosphere's initial state. These sampled points collectively form the ensemble of initial conditions, each point representing a possible initial state. Instead of explicitly predicting the movement of the whole probability distribution through phase space, ensemble forecasting approximates it by tracking the trajectories of the ensemble members through the phase space. (Wilks, 2011).

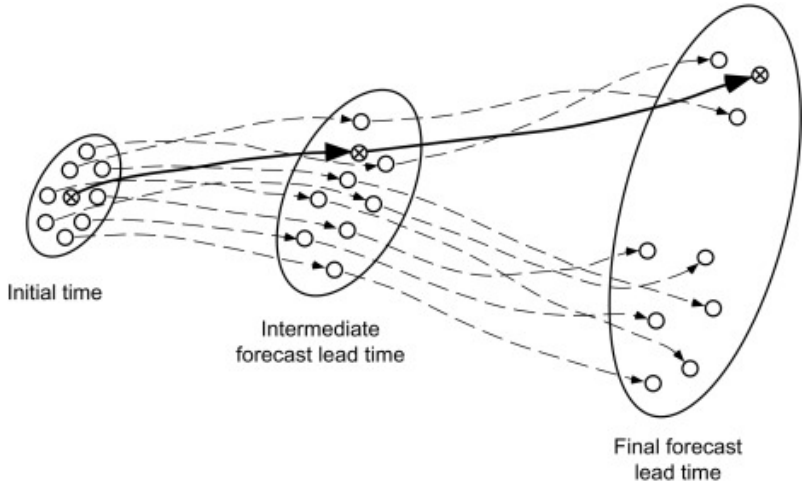


Figure 2.1: Schematic illustration of concept in ensemble forecasting, in an idealized two-dimensional phase space. This figure is adapted from Wilks (2011).

Figure 2.1 illustrates the concept of ensemble forecasting in a two-dimensional phase space. The crossed circle on the left side of the initial time ellipse represents the best initial value, which is the start of a deterministic forecast. The solid line represents the evolution of the deterministic forecast within the phase space, starting from the initial forecast through to the intermediate and final projections. The eight circles surrounding the central circle within the ellipse represent additional initial states, collectively approximating the variation present in the full distribution from which they were sampled. Initially, all ensemble members are very similar, but as the forecast progresses, they diverge due to the inherent chaotic nature of atmospheric dynamics. At the intermediate forecast lead time, all members produce similar forecasts, which means the forecast uncertainty is not much larger than at the initial time. However, between the intermediate and final forecast lead time, the trajectories diverge, thus the larger ellipse indicates the large uncertainty of the forecast. The large dispersion at the end is from initial condition error and model error, which is a single deterministic forecast unable to represent (the crossed circle at the final forecast lead time). The next two sections will introduce the methods used by operational forecast centers to choose the initial conditions to represent initial uncertainty and model uncertainty into those initial conditions.

Representation of initial uncertainty in ensemble forecasts

It must be noted that although the representation of initial uncertainty and model uncertainty are introduced separately, they are usually addressed together, not independently. In operational practice, producing each ensemble member requires rerunning the model every time, which implies a substantial computational cost. The limitation of computer power makes the selection of certain members to sufficiently represent the uncertainty become necessary. However, the actual probability density function of initial condition uncertainty is unknown, and it changes from day to day, which makes the selection of random samples from this distribution impossible (Wilks, 2011).

Various techniques are employed in ensemble forecasting to represent initial uncertainty in ensemble weather forecasting. These techniques can be categorized based on whether they aim to sample the

probability density function or selectively sample uncertainty dynamically in significant directions in state space. For instance, the Canadian Meteorological Service ensemble is based on initial conditions from an ensemble Kalman filter, where observations are incorporated into the model (Houtekamer et al., 2005). Though it is a computationally efficient data assimilation method, it has a limitation because it does not allow for the localization of ensemble-based covariances. Localization is a technique used to address the issue of spurious long-range correlations that arise due to sampling uncertainty in ensemble-based covariance estimates. By filtering out these correlations, localization helps to improve the accuracy and reliability of the assimilation process (Leutbecher and Palmer, 2008).

Another technique focuses on sampling the most dynamically relevant aspects of initial uncertainty. For example, the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble perturbs initial conditions based on leading singular vectors. They identify the initial uncertainties that will lead to the largest forecast uncertainties at the end of the specified period. By focusing on these dynamically most relevant directions, the singular vector approach ensures that the ensemble captures the most significant sources of uncertainty (Leutbecher and Palmer, 2008).

Representation of model uncertainty in ensemble forecasts

Apart from initial uncertainty, model uncertainty must be explicitly represented in ensemble forecast systems to prevent underdispersion. While initial condition uncertainty arises from imperfect observations of the atmosphere's starting state, model uncertainty stems from the inherent need to solve the governing equations numerically rather than analytically. Furthermore, the representation of unresolved scales of motion through parametrizations also contributes to model uncertainty (Leutbecher and Palmer, 2008).

In general, there are three methods for representing model error: the multi-model ensemble, the perturbed parameter ensemble, and stochastic-dynamic parametrization. The multi-model involves combining results from multiple quasi-independent climate models developed by different institutes around the world, and it has shown enhanced reliability over single-model ensemble forecasts (Leutbecher and Palmer, 2008). The perturbed parameter ensemble assumed that the correct tendency can be obtained by suitably perturbing the parameter values within a given parameterization scheme. However, for stochastic-dynamic parametrization, instead of using bulk formulas or an ensemble of parameterizations, it introduces stochastic elements to represent the inherent randomness and variability of subgrid processes (Palmer et al., 2005). For example, the IFS ensemble systems employed a Stochastically Perturbed Parametrisation Tendencies (SPPT) to represent model uncertainty with perturbations. This approach introduces random perturbations to the tendencies calculated by the physical parameterization schemes in the model, which can involve adjusting a distribution to control the characteristics and properties of the model error representation. Thus, compared to ensembles with initial perturbation only, the ensembles that account for model uncertainty are more reliable and can avoid being underdispersive (Palmer et al., 2009, page 2).

In all, the enhancement of the predictive skill of NWP has been achieved through scientific developments through advancements in better representation of unresolved processes in global models, the development of ensemble methods that provide forecast uncertainty estimates, and the development of techniques to determine the initial state. In addition, the increased computing power has enabled these advancements and contributed to the steady increase in forecast skills. The future NWP system will involve more computational tasks due to the need for higher resolution, but we can not expect future high-performance computing technology to keep developing, which will impose new constraints on addressing the science challenges (Bauer et al., 2015).

2.2 Data-driven Weather Prediction

Data-driven refers to an approach that primarily relies on gathering and analyzing data, in contrast to methods driven by process or theory (de Burgh-Day and Leeuwenburg, 2023). In the area of machine learning (ML), all ML techniques are data-driven (Schultz et al., 2021). Recently, Data-driven models have emerged as a promising alternative to traditional NWP models in weather forecasting. They extract information from large amounts of historical data using advanced machine learning techniques that do not rely on explicit physical equations to learn patterns and make forecasts (Reichstein et al., 2019).

These models are also often interchangeably referred to as "deep learning models" (Olivetti and Messori, 2023), "machine learning-based models" (de Burgh-Day and Leeuwenburg, 2023), or "AI models" (Ebert-Uphoff and Hilburn, 2023). To maintain consistency and avoid confusion throughout the master thesis, the term "data-driven weather prediction model" or "data-driven model" will be used to refer to the complete replacement of NWP models by a data-driven approach. The data-driven approach has shown the potential to improve forecast skill, especially for short-term predictions, and can complement or even replace computationally expensive NWP models in certain situations (Rasp et al., 2018). However, further challenges remain, including those related to the necessity of large and high-quality datasets, the interpretability of the models, and the possibility of extrapolating beyond the training data (Schultz et al., 2021).

2.2.1 Data-driven approaches for emulating components of numerical weather prediction models

Machine learning is an increasingly powerful tool that has proven to be computationally efficient. Much research has been conducted on applying machine learning to replace various components in the NWP workflow to increase efficiency.

The two fundamental elements in NWP models are the parameterization scheme and the dynamical core (see Fig. 2.2), and they are the two components that consume significant computational time. For example, within the ECMWF IFS model, the parameterization scheme accounts for about one-third of the total computational cost during model running (Chantry et al., 2021). Machine Learning (ML) is utilized as an alternative tool to emulate and speed up parameterization by

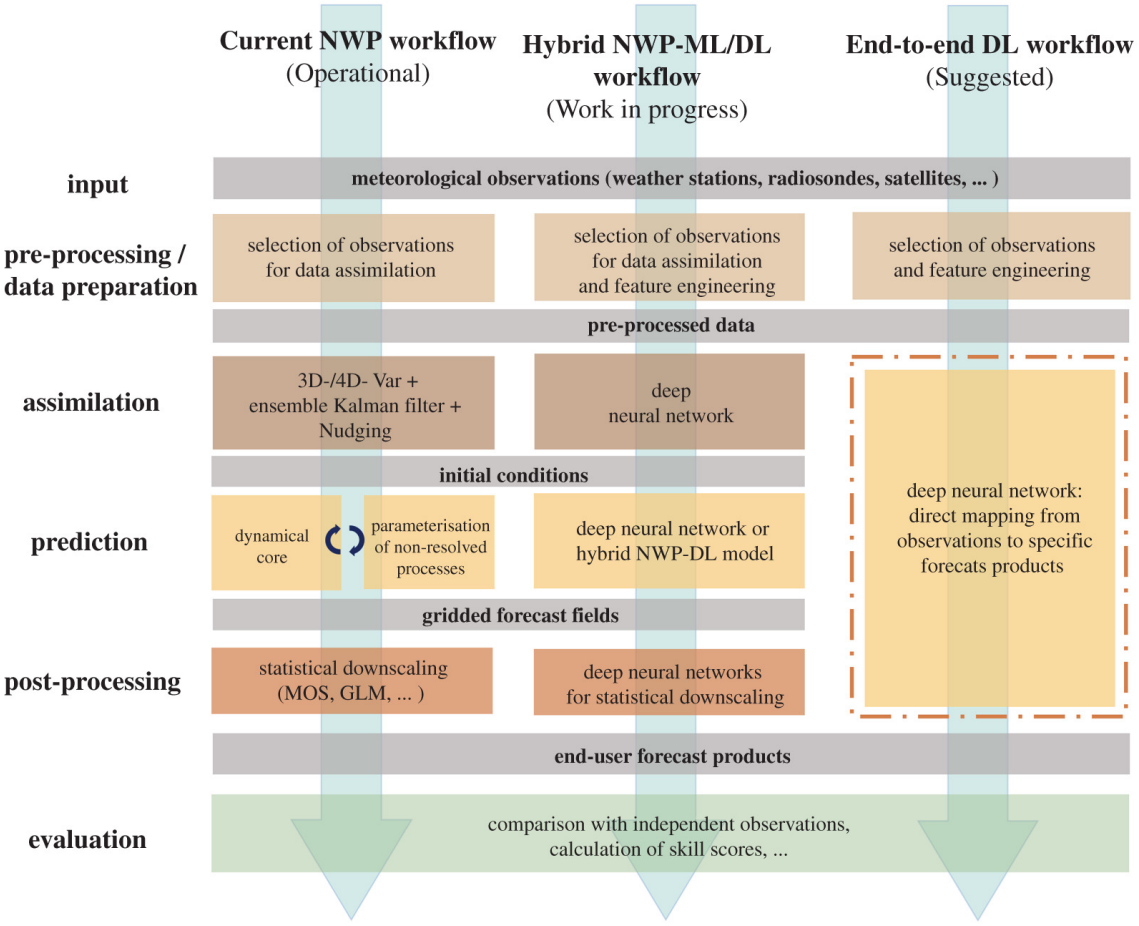


Figure 2.2: Idealized workflows of current numerical weather prediction(left), next-generation weather prediction with individual components substituted or augmented by ML and DL techniques(center), and purely data-driven Deep Learning (DL) weather forecasting systems(right). This figure is adapted from Schultz et al. (2021).

either emulating the whole parameterization scheme or sub-components of the scheme (Gettelman et al., 2021; Rasp et al., 2018). The Partial Differential Equations (PDEs) that represent the dynamical core in NWP models govern the fundamental physical processes of the atmosphere. The solving process is iterative and must be solved at every time step and grid point, making NWP models computationally intensive. There has been growing research exploring ML techniques to accelerate the solving process in the preconditioning and execution of solvers (Ackmann et al., 2020). Additionally, ML techniques have been employed in post-processing and downscaling to refine model raw output and increase accuracy and resolution (Harris et al., 2022). In addition to using machine learning techniques to enhance NWP models as a combination with physics-based models, a more radical and ambitious approach would be a fully data-driven replacement of the entire NWP model to provide forecasts from short-term to seasonal ranges (Figure 2.2, third column). This will also be the focus of the next section.

2.2.2 The development of data-driven weather prediction models

Early developments of data-driven weather prediction models

Dueben and Bauer (2018) took the first step and discussed the potential of Neural Networks for global weather prediction. They developed a toy model based on neural networks and compared it to the Lorenz 95 model, which has low complexity and includes only basic equations at coarse resolution. In their study, they only focused on predicting the geopotential height at 500 hPa and concluded that it is 'fundamentally' possible to generate global weather forecasts with coarse resolution for short-range prediction. The authors argued that while neural network models may excel in short-range regional forecasts, they would not be competitive on medium-long-range and climate timescales where maintaining physical consistency is key.

Rather than only extracting information from Global Circulation Models (GCMs) using machine learning techniques, Scher (2018) combined Convolutional Neural Networks (CNNs) with an autoencoder architecture to directly emulate the complete physics and dynamics of the GCMs themselves. Their conceptual study proved to be very promising, with the trained neural network able to skillfully predict the model state many time steps ahead while producing stable climate runs whose statistics closely matched those of the original GCM. Importantly, their neural network outperformed persistence and climatology baselines, which form some of the most important benchmarks for skill in weather forecasting. However, the GCM used was highly simplified, lacking both seasonal and diurnal cycles, oceans, and orography and using a coarse 625km resolution.

Building on this proof-of-concept, Scher and Messori (2019) adopted the same network architecture developed in (Scher, 2018) but applied it with more complex GCMs. Their study assessed for the first time how increasing the complexity of the underlying GCM affects the skill of the neural networks. While the neural networks continued demonstrating weather forecasting skills superior to baseline methods even for the more complex GCMs, the full climate statistics, particularly seasonal cycles, proved hard to reproduce. This highlighted the difficulty faced when using neural networks to emulate the behavior of comprehensively complex GCMs incorporating processes like seasonal variations.

After, subsequent studies focused on training Neural Network (NN) models using historical weather data to produce forecasts. Weyn et al. (2019) firstly trained a Convolutional Neural Network (CNN) on 24 years of ERA5 reanalysis data to predict 500 hPa geopotential heights and 300-700 hPa thicknesses over the Northern Hemisphere for medium-range forecasts. Their CNN application improved medium-range predictions over the persistence, climatology, and barotropic model for 500 hPa heights but was not better than operational weather forecasting models. Building on this work, Weyn et al. (2020) similarly developed a deep U-Net CNN for multi-meteorological variable forecasting. This model was significantly better compared to the previous one by remapping the data to a cubed-sphere grid, which minimized the distortions during convolution operations on the cube faces and supplied natural boundary conditions across face edges. They added additional capabilities to the CNN encoder-decoder architecture and utilized sequence prediction techniques.

With these improvements, the new data-driven model showed similar performance with the coarse-resolution ECMWF IFS model in global Root Mean square Error (RMSE) and Anomaly Correlation Coefficient (ACC). However, its performance still lagged behind high-resolution operational models and subseasonal-to-seasonal (S2S) forecasting systems. Despite these limitations, the data-driven model has demonstrated huge potential in competing with numerical weather prediction models.

Previous works showed the potential of using a CNN to produce skillful data-driven weather forecasts (Weyn et al., 2019, 2020; Rasp and Thuerey, 2021). Keisler (2022) explored using Graph Neural Networks (GNNs) instead of CNNs for data-driven weather forecasting. Their GNN model predicted 6 important atmospheric variables at 13 pressure levels in an autoregressive-iterative setup at 6-hour time steps. Their model outperforms the results of previous data-driven models from Rasp and Thuerey (2021) and Weyn et al. (2020). Its forecast skill is comparable with the current (in 2022 when the paper was published) operational global forecast system (GFS) model from the National Oceanic and Atmospheric Administration (NOAA) but still falls behind the higher-resolution IFS model from ECMWF. The authors argued three main reasons for the GNN's improved performance compared to models based on CNN: Firstly, the GNN is more flexible in handling the spherical geometry. Secondly, the forecast time step is relatively short, 6 hours. Thirdly, the Graph Neural Network (GNN) models a denser 3-D atmospheric state with more than one pressure level.

Advances in data-driven weather forecasting models since 2022

Significant progress has been made in data-driven weather prediction models since 2022 based on foundational key works before, with several models showing comparative and even outperforming the state-of-the-art IFS high-resolution models.

Pathak et al. (2022) employed a fourier-based neural network to develop FourCastNet. Trained with ERA5 reanalysis data, FourCastNet can produce 20 variables, including challenging ones like surface wind and precipitation, on five vertical levels with high horizontal resolution. This high resolution allows the model to resolve extreme events, such as tropical cyclones and atmospheric rivers, and allows for comparison with the high-resolution IFS of ECMWF. FourCastNet demonstrates competitive skill with the IFS high-resolution model up to a lead time of 7 days.

In the same year, Bi et al. (2023) used vision transformer architecture, and their model PanguWeather first reached better performance than the IFS high-resolution model for RMSE and ACC. Compared to FourCastNet, instead of predicting variables at each level separately like FourCastNet, it takes input weather variables with 13 vertical levels and feeds them into a single deep network, allowing vertical information flow (de Burgh-Day and Leeuwenburg, 2023). The second innovation was in the use of hierarchical temporal aggregation, where four versions of the model were used to predict at different lead times (1h, 3h, 6h, 24h) to avoid cumulative error from too many iterations. Lam et al. (2022) used GNNs to train their model GraphCast, which exceeds the skill of Pangu-Weather. GraphCast surpasses the ECMWF's high-resolution model on 90.0 % of the 2760 variable and lead time combinations.

Until now, there are still new models that keep coming out and target different needs of weather and climate modeling tasks. While previous models were mainly trained with ERA5 reanalysis, recent developments have broadened the scope of training data. Nguyen et al. (2023) introduced ClimaX, which was trained with heterogeneous climate datasets, enabling the model to generalize to diverse tasks based on different temporal and spatial horizons of interest. At the same time, there are models (Chen et al., 2023a,c) which are specifically developed for working with longer lead time predictions and have shown promising results. Besides, operational weather forecasting centers, such as ECMWF, are actively engaged in developing their own data-driven models (Lang et al., 2024).

While current state-of-the-art data-driven models have made remarkable achievements, most of them still rely on reanalysis data for training. Moreover, given the usual sample amount of training data, typically based on about 30 years of reanalysis data, substantial concerns have arisen regarding performance in extreme weather forecasting. Olivetti and Messori (2024) argued that current models are optimized for overall performance by averaging the forecast error, which may lead to poor performance in capturing extreme weather events. Another challenge for data-driven models is the development of probabilistic forecasts. Most models only target deterministic forecasts. A few data-driven models such as FourCastNet (Pathak et al., 2022) and PanguWeather (Bi et al., 2023) tried to generate ensembles by perturbing initial conditions, and Fuxi (Chen et al., 2023c) perturbed the model parameters. There is still an open question regarding the methods for quantifying uncertainties in data-driven models.

2.3 Heat Waves

Heat waves are often defined as prolonged periods with excessively high temperatures than normal. Prolonged extreme temperatures from heatwaves can devastate agricultural crops, increase energy demands, damage critical infrastructure, and trigger economic losses (Domeisen et al., 2022b). With increasing global warming, heatwaves are projected to increase in frequency, intensity, and duration in most regions (Perkins-Kirkpatrick and Gibson, 2017). Given the significant impact of heatwaves, accurate prediction of these extreme heat events is therefore crucial for preparedness and mitigating their widespread impacts on natural and human systems.

Heatwaves are typically defined based on a temperature threshold such as the 90th percentile or higher, often requiring a persistence of at least three consecutive days (Domeisen et al., 2022b). However, the specific definition can vary depending on the local climatology and associated impacts (Alexander et al., 2006; Perkins and Alexander, 2013; Russo et al., 2015). While definitions may differ across regions and applications, they generally aim to capture key heatwave characteristics including frequency, intensity, timing, and duration.

Since this thesis focuses on a case study, a detailed theoretical background on heat wave definitions and indices will not be covered. Instead, this section will introduce the processes that lead to heatwave development. Section 2.3.1 will introduce the key drivers and feedbacks crucial for

heatwave development, serving as the theoretical background for the discussion in Chapter 5. The theoretical background will conclude with Section 2.3.2, which will cover the unprecedented nature of the 2021 Northwest Pacific Heat Wave, the synoptic situation and near-surface air temperature evolution, as well as the drivers and mechanisms of this specific heat wave event.

2.3.1 Important drivers and feedbacks in the heat wave development

Understanding the processes that influence heatwave development and identifying the physical drivers of heatwaves allows models to represent these processes more accurately and provide improved forecasts. Heatwaves can result from a wide range of spatial and temporal scale processes with complex interactions (Barriopedro et al., 2023). Successful heatwave prediction requires considering these processes across a range of time scales. However, the relative contribution of each process and the necessary initialization can vary across different lead times (Domeisen et al., 2022b).

For short-term predictions (2-3 days), essential processes include the formation and maintenance mechanisms of quasi-stationary ridges or blocking patterns, anticyclonic flow anomalies, and diabatic heating from surface sensible heat fluxes. However, models often struggle to represent these flow patterns accurately (Grotjahn et al., 2015). On a time scale of up to 10 days, the representation of Rossby wave packets and the Madden-Julian Oscillation (MJO) can improve predictability (Tian et al., 2017; Fragkoulidis et al., 2018). For sub-seasonal scales, extreme heatwaves can show predictability in ensemble models, where ensemble members gradually cluster and shift toward warm anomalies (Domeisen et al., 2022a). The following discussion will present different processes ranging from large-scale to synoptic-scale and local-scale feedbacks to understand better the drivers and predictability of heatwaves across various time scales.

Atmospheric Blockings

In the extratropics, quasistationary anticyclonic flow anomalies, known as blockings in high latitudes and upper-level ridges (weak blockings) in low to mid-latitudes, are the primary drivers of heat waves. These anomalies disrupt the usual westerly flow, leading to a sharp transition from zonal to meridional pattern, which is the essential feature of blocking (Kautz et al., 2021).

There are two traditional approaches to defining blocking: one focuses on anomalies, while the other is based on absolute meteorological fields. Blocking can be identified using various dynamical parameters, such as negative potential vorticity anomalies in the upper troposphere (Schwierz et al., 2004), 500 hPa geopotential height (Tibaldi and Molteni, 1990), and potential temperature on the dynamical tropopause (Pelly and Hoskins, 2003). From the anomaly perspective, a blocking event can be identified by a 500 hPa geopotential height anomaly representing a reversal in the typical meridional gradient of geopotential height.

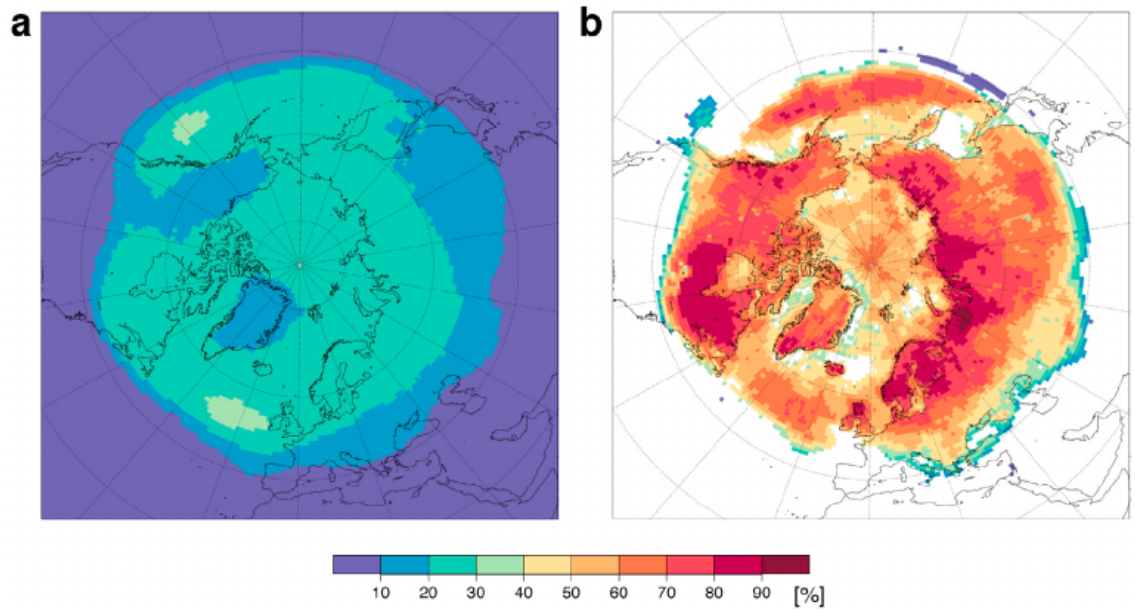


Figure 2.3: Blocking and high-temperature extremes. (a): Relative frequencies of northern hemispheric summer blocking events. (b): Percentage of having a high-temperature extreme and a co-located blocking. This figure is adapted from Pfahl and Wernli (2012)

Persistent blocking is strongly associated with extreme weather events, such as temperature and precipitation extremes. Pfahl and Wernli (2012) found a strong connection between atmospheric blocking and summer temperature extremes over large parts of the mid to high-latitude continents in the Northern Hemisphere. This connection becomes even more evident when considering weak blockings. Figure 2.3 shows that over large parts of high-latitude land regions, the percentage of warm temperature extremes related to weak blocking can reach 70-80%. This connection is particularly substantial over Siberia, Scandinavia, and the eastern North Pacific region.

Atmospheric blockings can create favorable conditions for surface temperature development through several processes. In the outer range of the blocking, the anticyclonic circulation can affect horizontal temperature advection. In the center of the blocking, persistent anticyclonic circulation drives strong subsidence, leading to adiabatic warming. Furthermore, the subsidence results in clear-sky conditions, enhancing shortwave radiation during the daytime (Trigo et al., 2004; Zschenderlein et al., 2020). The formation and maintenance of blocking are linked to various mechanisms across scales. Rossby wave-breaking events can strongly influence and lead to blocking events (Tamarin-Brodsky and Harnik, 2024). At the local scale, since the blocking anticyclone is characterized by a region of low potential vorticity air transported poleward within the upper troposphere, latent heating could enhance the transport of low potential vorticity air from the lower troposphere upward in warm conveyor belts, contributing to the formation of blocking (Madonna et al., 2014).

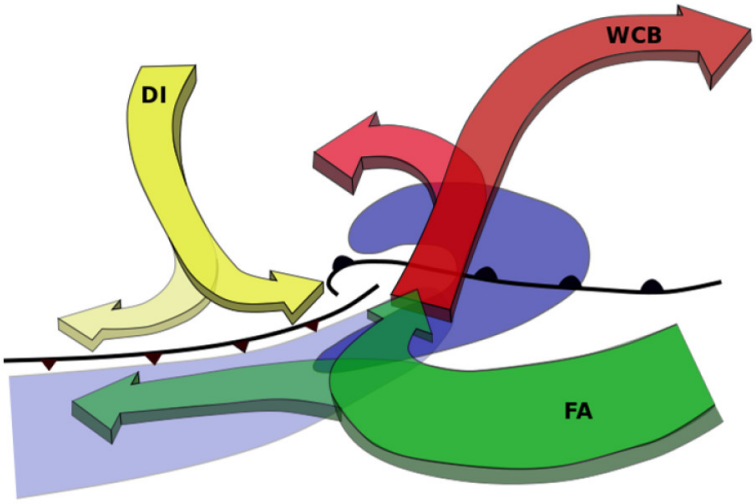


Figure 2.4: Schematic of cyclone-relative airflows over surface features. Includes cold and warm fronts (black), precipitation (dark blue), high TCWV (light blue), ascending warm conveyor belt (WCB; red), low-level feeder airstream (FA; green), and descending dry intrusion (DI; yellow). This figure is adapted from Dacre et al. (2019)

Warm Conveyor Belts

The Warm Conveyor Belt (WCB) was first identified through isentropic analysis of tropical cyclones. In this coordinate system moving with the cyclone, three distinct airstreams were observed (Green et al., 1966). It was further evidenced by using a Lagrangian approach, which identified its physical features, such as maximum ascent, increased potential temperature, and decreased specific humidity along the trajectory. This trajectory typically involves significant condensation and latent heating during the ascent phase (Wernli and Davies, 1997; Madonna et al., 2014).

WCBs can be linked to heat waves through several mechanisms. According to Pfahl et al. (2015), WCBs and the associated release of latent heat play a crucial role in the formation and maintenance of atmospheric blocking. A significant portion of air masses involved in Northern Hemisphere blocking undergoes heating of more than 2K within the three days before entering the blocking system, with a median heating of over 7K. AS The latent heat released during the ascent of air in WCBs can increase its potential temperature. This process also modifies potential vorticity (PV): below the level of maximum diabatic heating, the increase in potential temperature generates positive PV anomalies, while above this level, the decrease in heating leads to negative PV anomalies, known as PV destruction (Wernli and Davies, 1997). These negative PV anomalies brought by WCBs are crucial for the maintenance of atmospheric blocking and, consequently, for heatwave development.

Atmopsheric Rivers

Atmopsheric Rivers (ARs) are defined as narrow corridors characterized by enhanced water vapor transport, usually linked to low-level jets ahead of the cold front in the extratropical cyclone. The

ARs were usually identified by integrated water vapor (IWV) and integrated water vapor transport (IVT), but later the IVT method proved to be a more robust and appropriate parameter than IWV because it better described the horizontal moisture transport both at low latitudes and high latitudes (Ralph et al., 2020).

Though ARs are not typically important drivers during heat waves and are more important in leading to heavy precipitation events. However, anomalous atmospheric rivers can transport moisture and heat energy from low latitudes in the warm season. During the 2021 Pacific Northwest Heat Wave, an anomalous summertime AR played a crucial role in transporting warm and moist air from the Southeast Asian region all the way across the Pacific and made landfall over western North America(Mo et al., 2022). Though there is still no clear mechanism about how the ARs affected this heat wave, it likely provided moisture to the WCB airstream (Figure 2.4) and enhanced the latent heating release, which was crucial for the maintenance of atmospheric blocking as we discussed in the last section.

Land-atmosphere feedback and soil moisture deficit

In addition to the large-scale drivers mentioned earlier, regional to local feedback mechanisms can also influence the onset and duration of heat waves. In the Planetary Boundary Layer (PBL), the air directly interacts with the Earth's surface through processes of evapotranspiration (combined process of water evaporation from soil and transpiration from plants) and surface flux. The interaction between air and surface is governed by the land energy and water balance, and soil moisture plays an important role in these processes (Seneviratne et al., 2010).

Through evapotranspiration, moisture from the soil is transported back into the atmosphere. During periods of high soil moisture, the evapotranspiration rate is high, leading to higher latent heat flux and less sensible heat flux, resulting in lower surface temperatures. Conversely, during dry periods, the reduction in soil moisture content leads to a low evapotranspiration rate, meaning more energy is converted into sensible heat flux rather than latent heat flux, increasing surface temperature. The increased surface temperature will further dry the soil, continuing this cycle (Seneviratne et al., 2010). Fischer et al. (2007) investigated the role of land-atmosphere interaction in European summer heatwaves. By examining four major heatwaves using climate model simulations, they confirmed that soil moisture deficits and subsequent feedback mechanisms significantly amplify temperature extremes during heatwaves.

Miralles et al. (2014) presented a conceptual framework illustrating how soil moisture deficits and heat accumulation in the PBL contributed to extreme temperatures during mega-heatwaves. Fig.2.5 shows that desiccated soils contributed to higher surface sensible heat fluxes, which transfer more heat from the land surface to the overlying air, steadily heating the PBL. In addition to surface heating, warm air from higher altitudes is entrained into the PBL, further adding heat. The combined effect of surface sensible heat flux and warm air entrainment causes the PBL height to expand. As the height of the PBL increases, more warm air is entrained. At night, the heat generated during the daytime is preserved in the residual layer. Until the next day, the heat from the

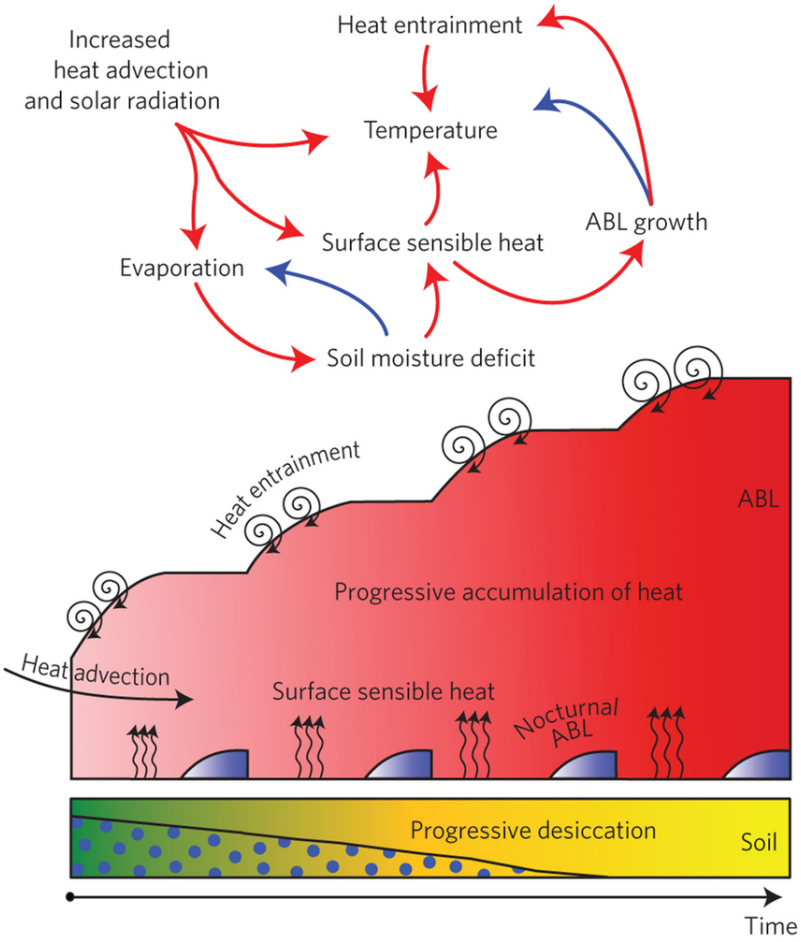


Figure 2.5: Conceptual model of land-atmosphere feedback. This figure is adapted from Miralles et al. (2014).

residual layer is re-entrained into the PBL. Over several days of development, this heat cycle leads to progressive heat accumulation in the PBL, enhancing soil desiccation and further escalating near-surface air temperatures.

Thus, an accurate representation of soil moisture and its interaction with the atmosphere is crucial for correctly simulating water and energy surface fluxes. For example, the Tiled ECMWF Scheme for Surface Exchanges over Land (TESSEL) is the operational land surface model employed in the IFS to simulate the evolution of soil conditions, vegetation states, and snow cover over continental regions at various spatial resolutions (Balsamo et al., 2009)

2.3.2 The 2021 Pacific Northwest Heat Wave

Overview: the unprecedented nature of the 2021 Pacific Northwest Heat Wave

From 25 June to 1 July 2021, the Pacific Northwest region of Canada and the United States experienced an unprecedented and extremely severe heatwave. Compared to the climatology from 1981 to 2021, near-surface air temperature anomalies soared to extreme highs of 16-20°C (Figure

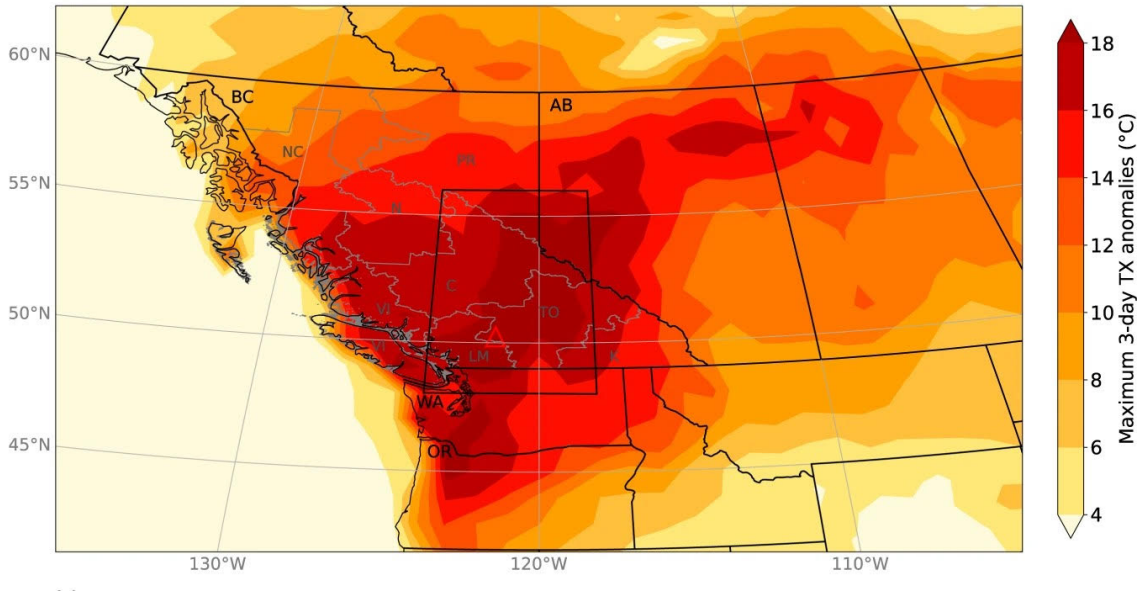


Figure 2.6: 3-day running mean of daily maximum near-surface air temperature anomalies with respect to 1981-2020 climatology, from June 23 to July 2, 2021, based on ERA5 reanalysis data. This figure is adapted from White et al. (2023).

2.6). On June 29, the village of Lytton in Canada (indicated by the red triangle in Figure 2.6) set a new national temperature record of 49.6°C, surpassing the previous record by an extraordinary 4.6°C. This temperature was reportedly the highest ever recorded north of 45° latitude worldwide (Environment and Climate Change Canada, 2022). Tragically, the next day, a catastrophic wildfire, exacerbated by drought conditions, devastated much of Lytton. This extreme heat event, far beyond historical experience, had devastating consequences, resulting in an estimated 740 excess deaths in the province British Columbia (Henderson et al., 2022).

This heatwave shattered numerous local historical records and ranked among the most extreme heatwaves worldwide (Thompson et al., 2022). According to White et al. (2023), the 2021 PNW heatwave was extraordinary even when compared to the infamous European heatwave in August 2003 and the Russian heatwave in July–August 2010. Although shorter in duration, the Pacific Northwest heatwave broke all-time temperature records by a significantly larger margin. Moreover, the maximum temperature anomalies, measured in standard deviations from normal, exceeded those of the 2003 European and 2010 Russian heatwaves (Thompson et al., 2022).

The exceptionally high near-surface temperatures during the heatwave were associated with remarkably high geopotential height and exceptionally dry soil conditions (Figure 2.7). From 28 June to 30 June, the average near-surface air temperature anomalies were exceptionally high, surpassing five times the daily standard deviation calculated from 1981-2010. Concurrently, the anomalies in geopotential height and soil moisture deficiency were also remarkably high, exceeding four and three times their respective standard deviations during the period from 25 June to 3 July (Bartusek et al., 2022).

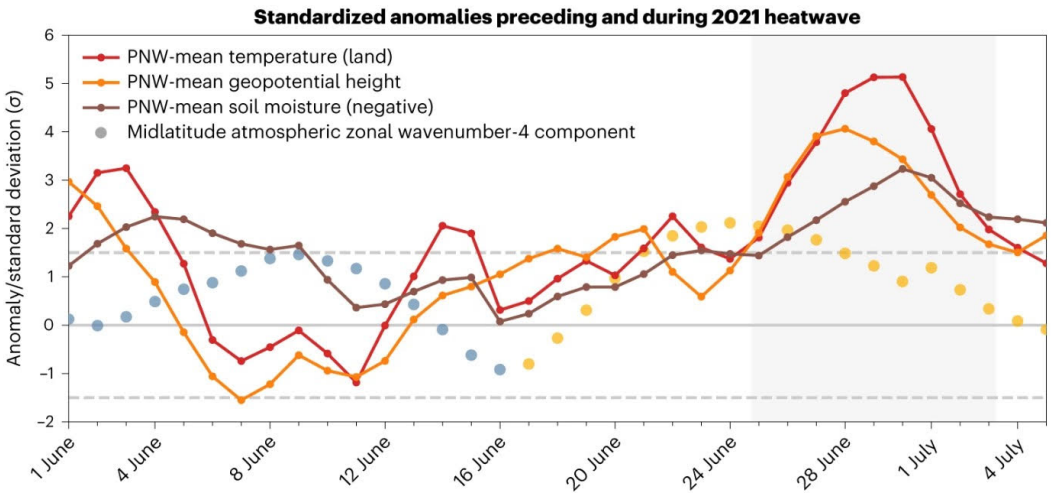


Figure 2.7: Evolution in ERA5 reanalysis throughout June averaged over the PNW (40–60°N, 110–130°W; land only) of near-surface air temperature (red line with dots), geopotential height on 500 hPa (orange line with dots) and soil moisture anomalies (brown line with dots), and the amplitude of a zonal wavenumber-4 disturbance in the midlatitude upper atmospheric circulation (colored blue when in negative phase and yellow in positive phase). Dashed grey lines represent ±1.5 standard deviations from their 1981–2010 mean. This figure is adapted from Bartusek et al. (2022).

Synoptic situation and near-surface air temperature evolution

The synoptic evolution leading to the high-temperature anomaly was associated with the development of an amplified upper-level ridge over the region. This ridge formed in connection with a rapidly deepening upstream cyclone (purple contour on fig. 2.8 (a)) that produced a WCB helping amplify the ridge (Oertel et al., 2023; Neal et al., 2022).

According to the analysis of Hotz et al. (2023), in the early stage of the heat wave (25 June to 26 June), near-surface air temperature anomalies in the PNW increased through both diabatic and advective processes (Fig.2.8 (i)). Air parcels ascended over the Pacific, moved poleward fed into the strengthening ridge, and then descended, contributing to positive temperature anomalies onshore (Fig.2.8 (b)). During the peak of heat wave (27 June - 30 June), the upper-level ridge was centered over the PNW, with average near-surface temperature anomalies exceeding 12 K ((Fig.2.8 (c), (e)). Temperature anomalies were primarily contributed diabatically at first, but the adiabatic contribution increased (Fig.2.8 (i)), associated with air parcels spiraling down anticyclonically into the ridge over the heat region (Fig.2.8 (d), (f)). After the peak (30 June - 1 July), total and diabatic temperature anomalies dropped (Fig.2.8 (i)), and this termination is associated with an onset of precipitation. Despite the upper-level ridge starting to shift east (Fig.2.8 (g)), advective temperature anomalies remained positive until 2 July, further indicating the termination was driven by convective cooling rather than cold air advection.

In summary, the near-surface air temperature development in 2021 PNW Heat Wave was associated with a quasi-stationary upper-level ridge downstream fueled by an upstream cyclone. Air parcel trajectory analysis revealed that the low-level air mass initially warmed due to latent heating upstream (from 24 June to 25 June). As this air mass subsided under the upper-level ridge, it

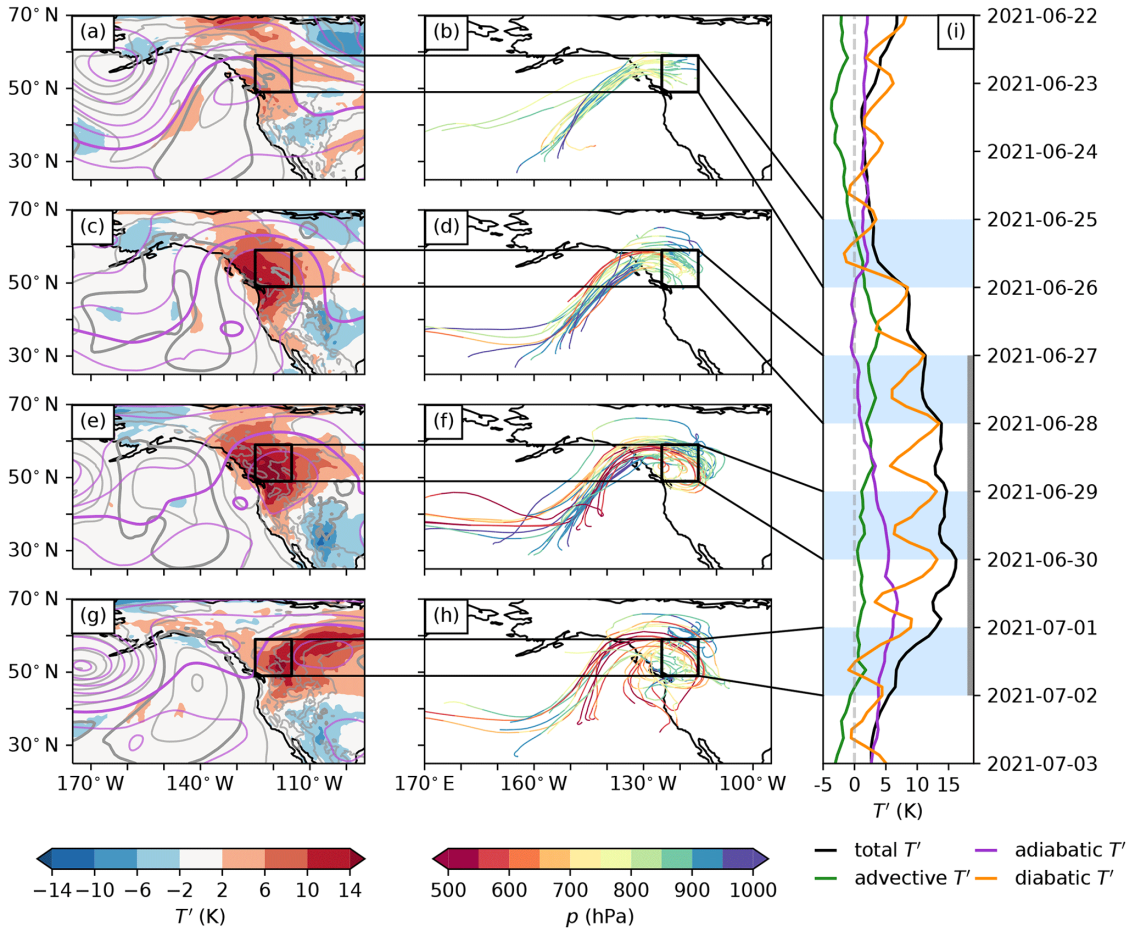


Figure 2.8: Synoptic evolution of the Pacific Northwest (PNW) heatwave, showing daily averages for (a, b) 25 June 2021, (c, d) 27 June 2021, (e, f) 29 June 2021, and (g, h) 1 July 2021. Panel (i) depicts near-surface temperature anomalies and contributions from 22 June to 3 July 2021. Left panels (a, c, e, g): color indicates temperature anomaly at 980 hPa; grey contours represent mean sea-level pressure (SLP) in 5 hPa intervals (1020 hPa bold); purple contours show geopotential height at 500 hPa in 100 m intervals (5800 m bold); black rectangle marks the PNW heatwave region. Middle panels (b, d, f, h): 30 trajectories with positive temperature anomalies, colored by pressure, arriving at land grid points on each date (10 trajectories each at 10, 30, and 50 hPa above ground level). Right panel (i): near-surface air temperature anomaly (black), advective component (green), adiabatic component (purple), and diabatic component (orange) from 24 June to 3 July 2021; grey line shows maximum 5-day daily temperatures. This figure is adapted from Hotz et al. (2023).

experienced further adiabatic warming. In the end, the heat wave eventually dissipated through convective cooling processes.

Drivers and mechanisms of the 2021 Pacific Northwest Heat Wave

Before discussing the drivers and mechanisms in detail, the predictive skill of this event and associated processes will be introduced first. According to Emerton et al. (2022), while seasonal forecasts from ECMWF indicated an increasing probability of above-normal temperatures for the Pacific Northwest region in June 2021, they did not consistently capture the signal for this record-breaking heat wave until about 2-3 weeks in advance. Lin et al. (2022) further evaluated

S2S forecast models and found that though forecasts beyond 2 weeks (initialized on 17 June) can indicate the above-normal temperature signals, they still struggled to predict the extreme intensity of the heat wave. It was not until the lead time was reduced to 5 to 11 days that the models began to accurately capture the intensity of this heat wave. They identified two key processes that contributed to this limited forecast skill before two weeks, an anomalous atmospheric river with its moisture transport (Mo et al., 2022), and the eastward progression of Rossby wave trains associated with the boreal summer intraseasonal oscillation in Southeast Asia. Aligning with previous findings, Oertel et al. (2023) focused on the medium-range forecasts and found that the magnitude of the heat wave could not be well predicted beyond a 7-day lead time. They discovered that a complex chain of synoptic precipitation events across the Pacific, involving enhanced WCBs, was crucial for amplifying the upper-level ridge downstream over Western North America. Only when these complex weather events were captured by the models could the extreme magnitude of the heat wave be accurately predicted.

The link between the 2021 PNW heat wave and a high-pressure system, referred to as a "heat dome," had been previously identified as the main dynamical component (Bartusek et al., 2022; Lin et al., 2022). This high-pressure system was characterized by subsidence, trapped air, and sensible heating as the dominant mechanisms driving the anomalous heat in the lower atmosphere. However, the conceptual model of a 'heat dome' overlooked the role of upstream diabatic heating processes (White et al., 2023). Based on a local wave activity diagnostic, Neal et al. (2022) revealed an essential role of diabatic heat released during cloud formation in an upstream cyclone in strengthening and amplifying this upper-level ridge associated with the heat wave. Oertel et al. (2023) further expanded on the diabatic heating perspective by quantifying the contribution of diabatic heat released by air masses originating from WCB ascent across the Pacific Ocean to the formation of the upper-level ridge.

While large-scale flow patterns played a role, Philip et al. (2022) pointed out that the observed near-surface air temperature anomaly during the 2021 PNW heat wave was more extreme than would be expected only based on these anomalous large-scale atmospheric flow patterns, suggesting the local topography and antecedent dry soil might have amplified this high-temperature anomaly. Schumacher et al. (2022) further investigated the complex interplay between dynamic and thermodynamic processes, finding that the latent heat release from the upstream cyclone not only helped to initiate the upper-level ridge and contributed to high-temperature anomaly in the upper-troposphere but also directly contributed to the extreme near-surface temperatures by mixing into the deep boundary layer facilitated by dry soil conditions. Further, this dry soil also acted to amplify surface heating by reducing evaporative cooling and enhancing sensible heating.

Several studies have highlighted the crucial role played by temperature anomalies in the upper troposphere, suggesting that the extreme magnitude of the 2021 PNW Heat Wave was controlled by this upper-tropospheric heat (Schumacher et al., 2022; Hotz et al., 2023). They argued that these positive temperature anomalies aloft helped suppress convective dampening, which would otherwise act to limit the near-surface extreme temperatures through convective mixing and precipitation processes. Based on moist convective instability theory, Zhang and Boos (2023) theoretically explained how upper-tropospheric temperature anomalies (represented by 500 hPa temperature)

could set an upper bound on surface temperatures during heat waves over mid-latitude continental regions. Further, Papritz and Röthlisberger (2023) quantified the sources of temperature anomalies near the surface and in the free troposphere and revealed that while local surface heating and subsidence were primarily responsible for the near-surface temperature anomaly, upstream diabatic heating substantially contributed to warming the air aloft in the upper troposphere which reinforce the idea the extreme buildup of heat near the surface during the heatwave was enabled by prior warming of the free troposphere aloft.

In summary, the 2021 PNW Heat Wave resulted from a combination and interplay between anomalous large-scale circulation patterns and thermodynamic processes. This unique combination of drivers and processes led to the record-breaking temperatures experienced during the heat wave.

3 Data and methods

3.1 Data

3.1.1 Observational reference: ERA-5 reanalysis

To better understand climate change and extreme weather, it is important to study the past. However, observational data is often incomplete and contains errors. Data assimilation methods combine observations with short-range weather forecasts to provide the best estimate of atmospheric conditions for weather prediction initialization. Reanalysis uses the same approach, but for past periods, gaps in observations are filled to create consistent long-term datasets. These datasets are valuable references for evaluating weather forecast model performance (ECMWF, 2023). ERA5 is the fifth-generation atmospheric reanalysis dataset produced by the Copernicus Climate Change Service (C3S) at the ECMWF. Covering the period from 1950 to the present, it provides a high-resolution (0.25° or approximately 31 km) representation of atmospheric conditions across 137 vertical levels, extending up to 80 km altitude. ERA5 is based on the ECMWF's Integrated Forecast System (IFS) high-resolution model (Cycle 41r2) and 4D-Var data assimilation (Hersbach, 2023).

In this thesis, ERA5 is not only used as ground truth to evaluate the performance of models but also used as the base state for the initial condition generation of the data-driven model (Section 3.2.3). The variables used at multiple levels and in detail are summarized in the following table (Table 3.2). Additionally, the climatology is computed from the ERA5 reanalysis data, which is interpolated to a grid spacing of 1° for the period from 1979 to 2019. For each grid point, a time period centered on 29 June is considered, spanning from June 15 to July 14.

3.1.2 Numerical Weather Prediction reference: ECMWF IFS forecasts

ECMWF IFS model, widely considered the best globally for medium-range weather forecasting, is chosen as the baseline numerical weather prediction model. The IFS undergoes regular updates to maintain its leading position, resulting in varying model configurations over time. In this thesis, the cycle 47r2 configuration is utilized. The IFS offers several atmospheric model configurations for different forecast ranges, with the medium-range forecast comprising the high-resolution single forecast (HRES) and the ensemble forecast (ENS). In this thesis, both the IFS HRES and ENS forecasts are employed.

IFS HRES

The HRES runs at a high resolution of 0.1° (approximately 9 km), with 137 vertical levels. It is run four times daily at 00, 06, 12, and 18 UTC. The forecast initialized from 00/12 UTC covers a period of 10 days, while those initialized from 06/18 UTC cover 3.75 days. The initial conditions are derived using an ensemble 4D-Var data assimilation system, which incorporates information from the forecast of the previous assimilation cycle and collected observations within a 3-hour assimilation window around the analysis time (Owens and Hewson, 2018). In this thesis, the IFS HRES forecasts initialized at 00/12 UTC are retrieved from WeatherBench2 (Rasp et al., 2023) with a re-gridded resolution of 0.25°.

IFS ENS

The ENS forecast consists of 51 members, including one unperturbed control forecast and 50 perturbed forecasts. The ensemble data assimilation (EDA) creates initial conditions by incorporating observation, model, and boundary condition errors for generating perturbed members. To further represent forecast uncertainty, singular perturbations are added to the initial conditions and designed to capture the most rapidly growing modes of atmospheric instability. The model uncertainty is represented by Stochastically Perturbed Parameterisation Tendencies (SPPT), which introduces spatially and temporally correlated perturbations to the model physics tendencies (Owens and Hewson, 2018). Until 2023, the horizontal resolution was 0.2 degrees. In Cy48r1, the resolution of the ENS is upgraded to 0.1 degrees, matching the resolution of HRES. The ensemble forecasts are initialized at 00/12 UTC and provide a longer range of forecasts up to 15 days. In this thesis, the IFS ENS are re-gridded to 0.25°.

3.1.3 Data-driven model: FourCastNet forecasts

In this Master's thesis, two versions of the data-driven model FourCastNet are used to compare with the forecast of the ECMWF Integrated Forecasting System (IFS) model. The updated version of the FourCastNet differs from the original model by incorporating a new neural operator and more dense vertical information input. The following section will describe the model architecture and discuss the differences and updates between these two versions.

FourCastNet Version 1

The Fourier ForeCasting Neural Network (FourCastNet) (Pathak et al., 2022) uses the Vision Transformer (ViT) as the architectural backbone, combined with Fourier Neural Operators (FNOs), resulting in the Adaptive Fourier Neural Operator (AFNO) model. The Vision Transformer (ViT) can process images by transforming them into a series of tokens, which are then fed into the transformer. This allows the model to capture long-range dependencies effectively, making it

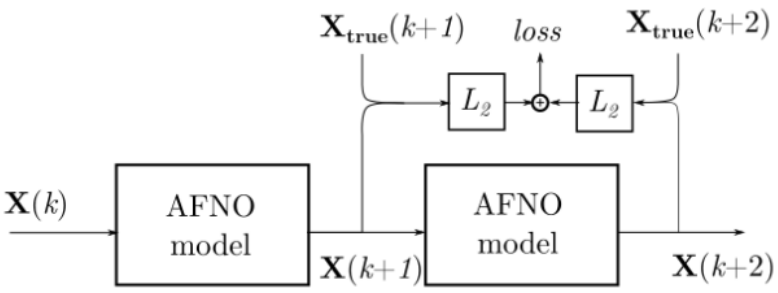


Figure 3.1: Two steps of training include pre-training and fine-tuning of FourCastNet; This figure is adapted from Pathak et al. (2022).

efficient for understanding spatial relationships in high-resolution atmospheric data. Furthermore, the Fourier Neural Operator enables the neural network to operate in the frequency domain, thereby allowing the network to learn complex functions and make accurate predictions (Falk et al., 2018). The integration of ViT and FNO in the AFNO model allows dependencies to be captured across spatial and channel dimensions.

FourCastNet (version 1) uses high-resolution input data comprising 6 surface variables and 5 atmospheric variables at four pressure levels (Table 3.1.4) and predicts the atmospheric state with a 6-hour temporal resolution. It is trained with ERA5 reanalysis data in two steps: pre-training and fine-tuning. The first step is to pre-train the model with meteorological variables in ERA5. Each variable is re-gridded and represented by a 2D field consisting of pixels (721×1440). Consequently, a single training data point at a specific time is represented by a tensor with dimensions of ($721 \times 1440 \times 20$). The training dataset spans years from 1979 to 2015, the validation dataset covers the years 2016 and 2017, and the testing dataset includes data after 2018.

Next, the training steps are described. Each training data point at a specific time is denoted as $X(k\Delta t)$, where k represents the time index, and Δt is fixed as 6 hours. In the pre-training step, the AFNO model is trained to learn the mapping from $X(t)$ to $X(t+\Delta t)$ in the next time step. In the subsequent fine-tuning step, the model uses the output $X(t+\Delta t)$ as input to generate the output $X(t+2\Delta t)$. The training loss is then calculated by comparing $X(t+\Delta t)$ and $X(t+2\Delta t)$ with their respective ground truth values (ERA5), $X_{true}(t+\Delta t)$ and $X_{true}(t+2\Delta t)$ (Figure 3.1). The model is optimized by minimizing the sum of these two computed training losses. After the pre-training and fine-tuning steps, the validation dataset is used to estimate the skill of the model through hyperparameter optimization.

FourCastNet Version 2

Based on the first version of the FourCastNet, FourCastNet2 utilizes Spherical Fourier Neural Operators (SFNO) to represent non-linear atmospheric dynamics better. Traditional FNOs use fast Fourier transforms, which can transform data to the frequency domain where global, long-range dependencies can be modeled more easily (Falk et al., 2018). However, a significant drawback

of using fast Fourier transforms is that they assume a Euclidean (flat) domain. When applied to spherical data, such as Earth's surface data, this assumption leads to issues like incorrect handling of the poles. This means that FFT does not naturally respect the spherical geometry, leading to artifacts when modeling data on a sphere. Instead, SFNOs employ a Harmonic Transform (SHT) that is suitable for spherical geometries, allowing for more accurate and realistic simulations of atmospheric dynamics. Moreover, SFNO has demonstrated long-term stability in maintaining plausible dynamics for year-long simulations, enhancing its predictive skill for medium-range to long-term weather forecasts (Bonev et al., 2023). FourCastNet2 is trained with a denser vertical level, encompassing 13 pressure levels, while maintaining the same spatial and temporal resolution as FourCastNet1 despite the architectural changes.

This thesis generates ensemble forecasts for two versions of FourCastNet, FourCastNet1 and FourCastNet2. Each ensemble consists of 50 members, with one control member. The ensemble forecasts for both versions are initialized at 00 UTC for each day from June 14, 2021, to July 4, 2021. For FourCastNet1, a total of 57 time steps with 6-hour intervals are generated, up to a maximum lead time of 336 hours. FourCastNet2 generates a total of 60 time steps with 6-hour intervals, up to a maximum lead time of 360 hours. Regarding the spatial area, the ensemble forecasts are generated globally and then extracted for the study area. Refer to Section 3.2.3 for the generation of initial conditions. Additionally, deterministic forecasts of FourCastNet2 are created with IFS HRES initial conditions, initialized at 00 UTC for each day from June 14, 2021, to July 4, 2021, a total of 60 time steps with 6-hour intervals are generated, up to a maximum lead time of 360 hours. The chosen variables in this thesis are summarized in Table 3.2, noted the specific humidity values of FourCastNet used in this thesis were derived from the dewpoint temperature, which was calculated from the variable of air temperature and relative humidity, in conjunction with the pressure.

3.1.4 Data-driven model: Pangu-weather forecasts

In this thesis, the Pangu-Weather model has been used for supplementary comparison. Pangu-Weather is a data-driven weather prediction model employing a transformer architecture - the 3D Earth-specific transformer (3DEST). Unlike traditional models operating in two dimensions, 3DEST processes data in three dimensions, considering latitude, longitude, and pressure levels, enabling information to flow horizontally and vertically. Pangu-Weather predicts 4 surface variables and 5 upper-air variables at 13 pressure levels with a 0.25° horizontal resolution (Table 3.1.4). It was trained with ERA5 reanalysis data from 1979 to 2017, using 2019 for validation and 2018, 2020, and 2021 for the test dataset. In contrast to FourCastNet, which follows a fixed interval approach, Pangu-Weather utilizes hierarchical temporal aggregation, training four different models with varying lead times (1h, 3h, 6h, 24h) to reduce cumulative forecast errors. For example, for a forecast with a 30-hour lead time, the 24-hour model will be executed once, followed by the 6-hour model executed once more (Bi et al., 2023). In this thesis, the Pangu-Weather model, initialized with the initial conditions from the IFS HRES forecast, is used and obtained from WeatherBench2

(Rasp et al., 2023) with a horizontal resolution of 0.25° and up to a maximum lead time of 240 hours.

	FourCastNet1	FourCastNet2	PanguWeather
Architecture	Adaptive Fourier Neural Operator	Spherical Fourier Neural Operators	3D Earth-specific Transformer
Resolution	$0.25° \times 0.25°$	$0.25° \times 0.25°$	$0.25° \times 0.25°$
Temporal steps	6h	6h	1,3,6,24h
Levels	surface and 4 pressure levels (50, 500, 850, 1000 hPa)	surface and 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa)	surface and 13 pressure levels (same as FourCastNet2)
Variables	Surface variables: T2M, U10, V10, MSL, TP, SP; Atmospheric variables: Z, T, RH, U, V	same as FourCastNet1 but with denser vertical levels	Surface variables: T2M, U10, V10, MSL; Atmospheric variables: Z, T, Q, U, V
Training datasets	ERA5 reanalysis (1979-2017)	ERA5 reanalysis (1979-2017)	ERA5 reanalysis (1979-2017)

Table 3.1: Summary features of data-driven weather prediction models. T2M: 2 m temperature, U10: 10 m zonal wind, V10: 10 m meridional wind, MSL: mean sea level pressure, TP: total precipitation, SP: surface pressure, Z: geopotential height, T: temperature, RH: relative humidity, Q: specific humidity, U: zonal wind, V: meridional wind.

Model/Dataset	Type	Δx	Δt	Chosen variables	Initial conditions
ERA5	Reanalysis	0.25°	6h (noted the temporal resolution for upper-level variables is 24 h, except for variables at 850 hPa and 500 hPa)	Single level - T2m, TCWV; multiple levels (13 levels) - T, RH, Z	-
IFS HRES	NWP deterministic forecasts	0.1° (remapped to 0.25°)	6h (noted the temporal resolution for upper-level variables is 24 h, except for variables at 850 hPa and 500 hPa)	Single level - T2m, TCWV; multiple levels (13 levels) - T, RH, Z	Operational IFS HRES ICs
IFS ENS	NWP ensemble forecasts (51 members)	0.2° (remapped to 0.25°)	6h	Single level - T2m, TCWV; multiple levels (4 levels) - T, RH, Z	Operational IFS ENS ICs
Pangu-Weather	Data-driven deterministic Forecast	0.25°	6 h	Single level - T2m, TCWV; multiple levels (13 levels) - T, Q, Z	IFS HRES ICs
ForecastNet1	Data-driven ensemble forecasts (51 members)	0.25°	6 h	Single level - T2m, TCWV; multiple levels (4 levels) - T, RH, Z	IFS ICs and gaussian noise ICs
ForecastNet2	Data-driven deterministic and ensemble forecasts (51 members)	0.25°	6h	Single level - T2m, TCWV; multiple levels (13 levels) - T, RH, Z	IFS ENS ICs, IFS HRES ICs and Gaussian noise ICs

Table 3.2: Summary of datasets used in this thesis. T2m: 2 m temperature, TCWV: total column water vapor, T: temperature, RH: relative humidity, Z: geopotential height, Q: specific humidity.

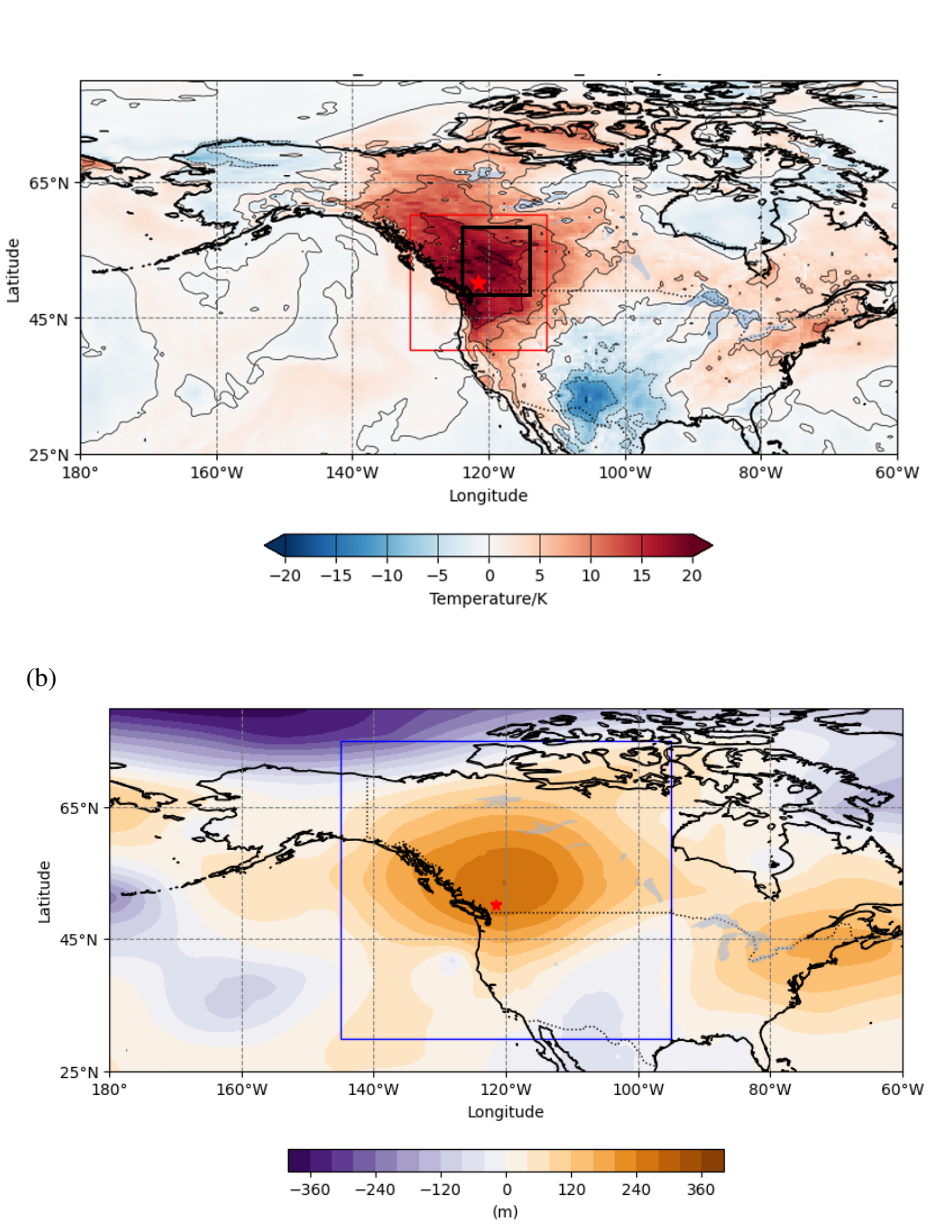


Figure 3.2: Study regions over the Pacific Northwest: (a) ERA5 two-meter temperature anomaly on 29 June 2021 at 00 UTC with respect to the ERA5 climatology for June and July from 1979 to 2019. The star represents the city of Lytton. The red solid box represents a 20° latitude by 20° longitude region centered on Lytton, while the black box encompasses the area between 49°N-59°N and 115°W-125°W. (b) ERA5 500 hPa geopotential height anomaly on 29 June 2021 at 00 UTC with respect to the ERA5 climatological mean for June and July from 1979 to 2019. The blue solid box represents the region spanning 145°W-95°W and 30°N-75°N.

3.2 Methods

3.2.1 Study domain

The focus of this thesis is on the Northwest Pacific region. The first study region is a 20° latitude by 20° longitude box centered on Lytton, following the same definition as in Oertel (2023). This region

is chosen to evaluate the model's forecast performance in predicting the magnitude of the 2-m temperature peak (29 June 2021 00UTC) during the 2021 Pacific Northwest Heat Wave. The model performance evaluation for 2-m temperature in Chapter 4 is based on this region. Additionally, the grouping of ensemble members based on their predictive skill in 2-m temperature, described in Section 4.1, is also based on the same study region. To specifically investigate the effect of land-atmosphere feedbacks on model performance, a second study region is defined that encompasses only land areas. This region, a black box with 10.5° latitude by 10.5° longitude, is employed in Chapter 5 to study the vertical profile of temperature and humidity.

Furthermore, to evaluate the forecast representation of the location and shape of the upper-level ridge, the 500 hPa geopotential height anomaly is used to characterize the upper-level ridge. The study region for this analysis is chosen as the box (145°W-95°W, 30°N-75°N) covering the high geopotential height anomaly over North America.

3.2.2 Evaluation metrics

In this section, several evaluation metrics used in the thesis are introduced. The equations are taken and adapted from Japan Meteorological Agency (2024).

Bias

The bias, also known as the mean error, represents the difference between forecast and verifying values and is defined as:

$$\mathrm{BIAS} \equiv \frac{\sum_{i=1}^{n} w_i D_i}{\sum_{i=1}^{n} w_i}, \tag{3.1}$$

$$D_i = F_i - A_i, \tag{3.2}$$

$$w_i = \cos \phi_i, \tag{3.3}$$

where F_i and A_i represent the forecast and verifying values for the i data point. D_i is the deviation between the forecast and the verifying value. n is the number of samples representing the total grid points of the field. If the forecast is perfect, bias equals zero, indicating no bias. When computing the average of a wide region, the differences in area due to the latitude need to be accounted for. In the equirectangular projection, the weighting coefficient w_i is replaced with the cosine of latitude to account for the convergence of longitude near the poles. In this thesis, the two-meter temperature fields chosen for evaluation by bias weighted at each grid point (i) by the cosine of latitude before calculating bias, with latitude expressed in degrees, for example:

$$T_{i_{\mathrm{weighted}}} = T_i \cos\left(\frac{\mathrm{lat} \cdot \pi}{180}\right) \tag{3.4}$$

Root Mean square Error

Root Mean Square Error (RMSE) measures the average magnitude of the errors. Unlike bias, which can show whether the forecast tends to be too high or too low, RMSE measures the overall accuracy by considering the square of the errors. This ensures that larger errors have a more significant impact on the RMSE value. It is defined by:

$$\mathrm{RMSE} \equiv \frac{\sqrt{\sum_{i=1}^{n} w_i D_i^2}}{\sqrt{\sum_{i=1}^{n} w_i}} \tag{3.5}$$

Where D_i is the deviation between the forecast and the verifying value as defined in eq. 3.2, RMSE is calculated as the square root of the average of squared differences between the predicted and verified values. w_i is the weighting coefficient as defined in eq. 3.3. In this thesis, the temperature and geopotential fields chosen for evaluation by RMSE are weighted at each grid point (i) by the cosine of latitude before calculating RMSE, with latitude expressed in degrees, for example:

$$T_{i_{\text{weighted}}} = T_i \sqrt{\cos\left(\frac{\text{lat} \cdot \pi}{180}\right)} \tag{3.6}$$

Anomaly Correlation Coefficient

The Anomaly Correlation Coefficient (ACC) is a widely used measure for spatial field verification. It quantifies the spatial correlation between forecast and observed anomalies, relative to climatology. ACC evaluates how well forecast anomalies reflect observed anomalies, indicating the accuracy of the forecast in predicting actual data (Andersson, 2015). It is defined as follows:

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} w_i (f_i - \bar{f})(a_i - \bar{a})}{\sqrt{\sum_{i=1}^{n} w_i (f_i - \bar{f})^2 \sum_{i=1}^{n} w_i (a_i - \bar{a})^2}}, \qquad (-1 \leqslant \mathrm{ACC} \leqslant 1) \tag{3.7}$$

where n is the number of samples, and f_i, a_i are respectively given by:

$$f_i = F_i - C_i, \quad \bar{f} = \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i}, \tag{3.8}$$

$$a_i = A_i - C_i, \quad \bar{a} = \frac{\sum_{i=1}^{n} w_i a_i}{\sum_{i=1}^{n} w_i}. \tag{3.9}$$

where F_i, A_i, C_i represent the forecast value, the verifying value, and the climatological value, respectively. Additionally, $\bar{f}$ and $\bar{a}$ are the mean value of f_i and a_i, and w_i is the weighting

coefficient as defined in eq. 3.3. The geopotential height on 500 hPa fields chosen for evaluation in this thesis by ACC has been weighted at each grid point (i) by the cosine of latitude. The Anomaly Correlation Coefficient (ACC) ranges from +1 to -1, where +1 indicates perfect correlation, 0 indicates no correlation (climatological average), and -1 indicates perfect anti-correlation. ACC values below 0.6 suggest that the positioning of synoptic scale features has little forecasting value (Owens and Hewson, 2018).

Forecast skill horizon

First, the concept of the 'forecast horizon' and 'forecast skill horizon' is differentiated. The forecast horizon is the time period between the initial forecast time and the future valid time, representing the length of the forecast period. The forecast skill horizon used by ECMWF is originally defined as the lead time when the ensemble ceases to be more skillful than a climatological distribution, based on the continuously ranked probability score (Buizza and Leutbecher, 2015). In general, the forecast skill horizon is the maximum lead time at which the forecast provides more accurate and reliable information than a reference baseline, such as the climatological distribution they defined.

In this thesis, rather than quantifying the practical skill limit against climatology, a 'skillful' near-surface air temperature forecast horizon is defined as the lead time when the forecast bias with respect to ERA5 reanalysis first becomes smaller than 5 K. For 500 hPa geopotential height, a skillful forecast horizon is defined as the lead time when the ACC value of the forecast compared with ERA5 becomes higher than 0.6 since lower values are insufficient for accurately positioning synoptic-scale features. The forecast skill horizons for these two variables must be interpreted separately due to their different definitions of 'skillful.

3.2.3 Generation of initial conditions

In this thesis, the approaches to producing ensembles for the data-driven model are all based on initial condition ensemble methods, which involve running the model multiple times with slightly different initial conditions. For FourCastNet Version 1 and Version 2, 50 ensemble members are generated using two initial conditions. One initial condition is derived from the IFS ENS analysis by extracting the initial state. The other initial condition is obtained by perturbing the ERA5 field with random Gaussian noise. The following two methods are then described using equations adapted from Bülte et al. (2024) and (Pathak et al., 2022).

Gaussian noise perturbation

The first approach is to add random perturbation to the initial state of the ERA5 field with a $0.25^\circ \times 0.25^\circ$ resolution to represent the initial condition error in the analysis. Pathak et al. (2022) first employed Gaussian noise to generate initial conditions:

$$X^{Gauss,m}_{lat,lon,t,0} = Y_{lat,lon,t} + \sigma\xi_{lat,lon,t} \qquad m = 1,\ldots,M, \tag{3.10}$$

where $Y_{lat,lon,t}$ represents the initial state of variables based on ERA5 reanalysis, lat and lon represent global grid points and t represents the initialization time. The Gaussian noise $\xi_{lat,lon,t} \sim N(0,1)$ is introduced, and σ is the tuning parameter to tune the noise level. Thus, Gaussian initial condition $X^{Gauss,m}_{lat,lon,t,0}$ is generated over the initialization time t, members m and variables. It should be noted that initial conditions are generated globally for all grid points in the models. In this thesis,$\sigma = 0.3$ is set as in Pathak et al. (2022), and $m = 50$ is set to generate 50 initial conditions for comparison with the same ensemble members of IFS.

IFS Initial condition

To compare with the IFS, the same initial conditions are used to compare the IFS models. The IFS ensemble forecasts $Z^{m}_{lat,lon,t,0}$ for each member (51 members in total) at time step 0 are extracted (starting point of the forecast) and used as initial conditions. Additionally, the IFS HRES forecasts at time step 0 are also chosen to initialize FouCastNet2.

$$X^{\mathrm{ifs},m}_{lat,lon,t,0} = Z^{m}_{lat,lon,t,0} \quad \text{for} \quad m = 1,\ldots,M. \tag{3.11}$$

As introduced in Section 3.1, all ensemble forecasts of FourCastNet (Version 1 and Version 2) are initialized at 00UTC. It should be noted that, compared to the initial conditions of IFS, ERA5 employs a longer assimilation window. Specifically, forecasts initialized from ERA5 at 00/12 UTC include observational data up to 9 hours ahead, whereas operational forecasts only extend 3 hours ahead (Lam et al., 2022). As a result, forecasts initialized from ERA5 at 00/12 UTC tend to perform better than those initialized at 06/18 UTC for short lead times. Therefore, care should be taken not to over-interpret the performance of forecasts initialized with ERA5-based Gaussian noise initial conditions at short lead times when evaluating them against ERA5 as the ground truth.

3.2.4 Classification of ensemble members

The purpose of classifying ensemble members into groups is to analyze differences between those that perform well (good members) and those that perform poorly (bad members) in predicting extreme temperatures during heatwaves. By separating ensemble members based on their predictive skill of near-surface air temperature, the processes contributing to successful or unsuccessful heatwave predictions in the data-driven models can be studied and compared.

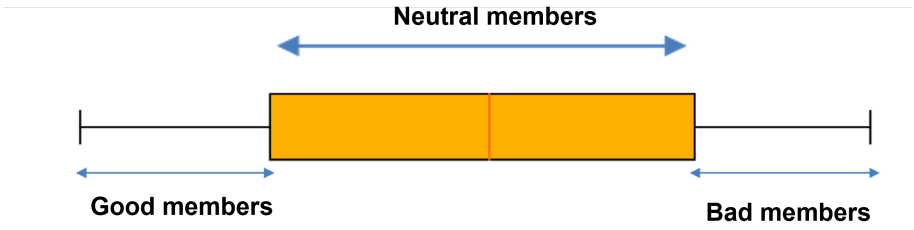


Figure 3.3: Schematic illustration of member groups. The box represents the 30-70 quantile range of RMSE, with the line inside the box indicating the median. This illustration is for conceptual purposes and is not derived from actual data.

Specifically, ensemble members are classified into three groups (good, bad, neutral) based on their forecast skill for 2-meter temperature during a heatwave period from June 27 to July 1, 2021, 00UTC. Classification uses the percentile rank of RMSE against ERA5 reanalysis, averaged over the study area (black box defined in Section 3.2.1). At each initialization time from 18 June to 28 June, the 15 members with RMSE smaller than the 30th percentile are "good members", the 15 members with RMSE larger than the 70th percentile are "bad members", and the remaining 20 are "neutral members" (see Figure 3.3).

4 Forecast evaluation for the 2021 Pacific Northwest Heat wave

On the peak day (29 June 00 UTC) of this heat wave, the heat wave magnitude represented by the two-meter temperature anomaly with respect to June to July climatological mean from 1979 to 2019 reached up to 15-20K over a large area of the Pacific Northwest (Figure 3.2a). The most extreme region can be identified around the city of Lytton, where the hottest temperature was recorded during this event. Although the amplitude of the upper-level ridge had peaked and matured on 26 June and 27 June and started to weaken by 28 June (Neal et al., 2022), the positive geopotential height anomaly at 500 hPa was still present over the North American continent on 29 June (Figure 3.2b).

According to Oertel et al. (2023), state-of-the-art numerical NWP models failed to capture the magnitude of the extremely high-temperature anomaly during the peak of this heat wave beyond a 7-day lead time, indicating the existence of a predictability barrier. This chapter aims to compare the performance of data-driven models against NWP models in predicting the magnitude of the heat wave peak temperature, following the same definition of heat wave peak magnitude as Oertel et al. (2023). Additionally, the predictive skill of data-driven models for the associated upper-level ridge pattern is investigated. Section 4.1 analyzes the forecast evolution initialized 14 days before the heat wave peak for 2-meter temperature and 500 hPa geopotential height. Section 4.2 compares their forecast skill horizons to quantify the predictive skill of models further.

4.1 Forecast evolution for the heat wave peak with lead time

4.1.1 Forecast evolution of 2-m temperature for the heat wave peak

In order to evaluate the prediction of heat wave magnitude at the peak, we first investigate the 2-m temperature forecast and compare how the forecasts evolve across lead time within different models. Figure 4.1 illustrates the domain-averaged 2m temperature forecasts valid at 29 June 00 UTC (4 p.m. in the Pacific Northwest), initialized between 15 June 00 UTC and 28 June 00 UTC. The ensemble and control forecasts are represented by boxes and diamonds, respectively. Additionally, the triangle represents the deterministic forecasts initialized with the operational IFS HRES initial conditions. These are shown for IFS (blue), FCNV2 with IFS initial conditions (green), FCNV2 with Gaussian noise initial conditions (yellow), and Pangu-Weather (red). For comparison, the

corresponding ensemble and control forecasts from FCNV1 with IFS and Gaussian noise initial conditions are also included (grey).

Before going deeper into the analysis of the forecast performance of 2-m temperature compared with ERA5 reanalysis, the general feature of the ensemble spread will first be introduced. Overall, as the initialization time approaches 29 June, the ensemble spread decreases for both data-driven models and NWP models, indicating a reduction in forecast uncertainty closer to the event. When comparing the control forecast with the ensemble spread, we can identify that for the IFS ENS, the control forecast consistently lies within the ensemble spread for almost all initialization time (except for initialization time on 15 June). However, for FCNV2 with IFS and Gaussian noise initial conditions, we can identify that their control forecasts occasionally lie outside the ensemble spread, particularly for forecasts initialized before 23 June. This may suggest that ensemble generation approaches based only on perturbing initial conditions may not fully capture the forecast uncertainty. While this approach accounts for initial condition uncertainty, it does not represent model uncertainty, as highlighted by Bülte et al. (2024). In contrast, the IFS ENS incorporates model uncertainty through stochastically perturbed parametrization tendencies (Section 2.1.2).

Moreover, compared with IFS ENS, the FCNV2 tends to have a smaller ensemble spread at longer lead time, especially for ensemble forecasts initialized with Gaussian noise initial condition. As the perturbed Gaussian noise is random and independent of the prevailing atmospheric condition, with the model integrated into a longer lead time, the noise will not be as strong as in the ifs initial condition. In contrast, IFS ENS initial condition employed the singular vector perturbation techniques can identify the most influential perturbation and better represent the forecast uncertainty even at the longer lead time (Section 2.1.2). It should be noted that the discussion of ensemble spread here is based only on the distribution of ensemble members, without direct computation and comparison to observational distributions. Previous discussions have pointed out the under-dispersion and underestimation of forecast uncertainty in data-driven model ensemble forecasts initialized by Gaussian initial conditions and IFS initial conditions (Bülte et al., 2024, Fig. 5b).

Next, the forecast bias with respect to ERA5 will be analyzed. The ERA5 reanalysis (red line) indicates that the 2-meter temperature on 29 June 00 UTC was approximately 303 K. Before 21 June, most model forecasts underestimated the intensity of the heat wave by more than 7 K. The Pangu-Weather forecast initialized on 19 June significantly underestimated the magnitude of this heat wave, showing a larger bias compared to FCNV2 and IFS. On 21 June, while the ensemble spread of the FCNV2 forecast is small, the IFS ensemble forecast shows a much larger ensemble spread and begins to anticipate the upcoming heat wave, but most ensemble members still underestimate the magnitude of the heat wave of more than 5 K.

Between 21 June and 22 June, the 2-m temperature forecast of FCNV2 and IFS experienced a so-called "forecast jump" (Richardson et al., 2024). They abruptly improved their forecast and started to capture the magnitude of this heat wave. The improvement between 21 June and 22 June of the ensemble mean for FCNV2 is even larger than that of IFS. Similarly, the operational Pangu-Weather also improved its forecast, and the underestimation on 22 June is less than 3K. Notably, between 20 June and 22 June, Pangu-Weather has already experienced a "forecast jump."

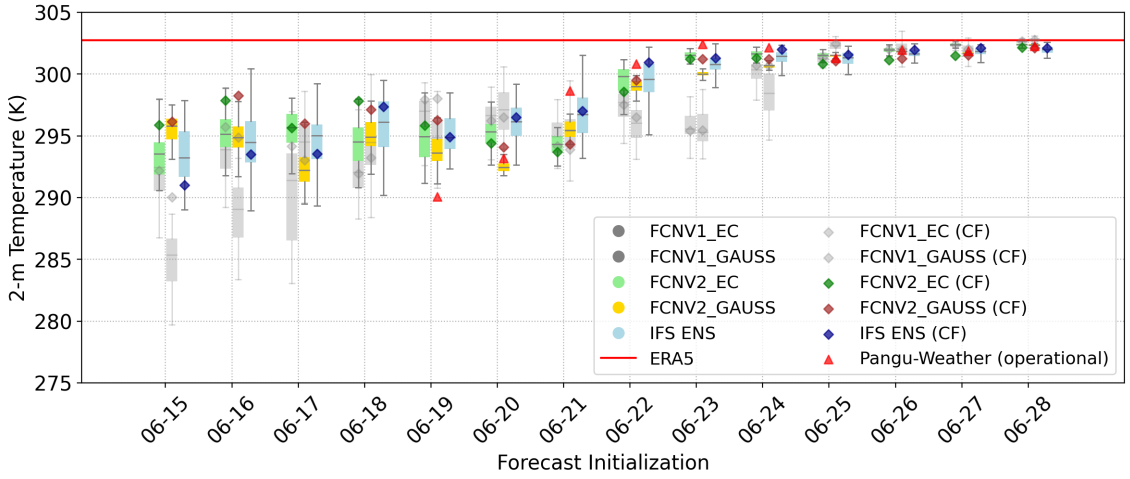


Figure 4.1: Distribution of ensemble forecasts of 2m Temperature on 29 June 2021 00 UTC. Averaged over a 20° latitude by 20° longitude box centered on Lytton, initialized between 15 and 28 June 2021, 00 UTC. The box marks the 25-75 quantile, whiskers mark the 1-99 quantile, and the line within the box represents the median. Colored diamonds show the control forecast (CF), colored triangles show the deterministic forecast. The red line represents the ERA5 reanalysis of 2m temperature at 00 UTC on 29 June. IFS_ENS: IFS ensemble forecasts; FCNV1_EC: ForeCastNet1 initialized with IFS initial conditions; FCNV1_GAUSS: ForeCastNet1 initialized with Gaussian noise initial conditions; PGW_HRES: Pangu-Weather initialized with IFS HRES initial conditions; FCNV2_EC: ForeCastNet2 initialized with IFS initial conditions; FCNV2_GAUSS: ForeCastNet2 initialized with Gaussian noise initial conditions.

After June 23, FCNV2, Pangu-Weather, and IFS accurately captured the magnitude of this heat wave.

Finally, it is worth noting that the performance of FCNV1 was significantly worse compared to the other models (grey box in Figure 4.1) throughout the initialization period, which consistently underestimated the magnitude of the heat wave. Unlike other models (FCNV2, IFS, and Pangu-Weather), FCNV1 did not exhibit a notable "forecast jump" between 21 June and 22 June. Instead, it only improved the forecast between 23 June and 24 June, which is two days later than the other two models. Even at a shorter lead time of 5 days (forecast initialized on 24 June), the FCNV1 initialized with Gaussian noise initial condition still had a forecast bias of around 5 K. Except for the change of architecture (Section 3.1.3), another distinctive difference between FCNV1 and FCNV2 is the different numbers of levels of input data (FCNV1 only with 4 pressure levels, as shown in Table 3.1.4). This might provide the FCNV2 with a more comprehensive representation of the atmospheric conditions, thus leading to the difference in the performance of 2-m temperature forecasts. In the next chapter, this will be further investigated.

4.1.2 Forecast evolution of 500 hPa geopotential height for the heat wave peak

As a correct prediction of heat wave magnitude is strongly linked to the right representation of the upper-level ridge, the forecast evolution of geopotential height on the 500 hPa level is further investigated. Figure 4.2 shows the forecast evolution of ACC for the 500 hPa geopotential height

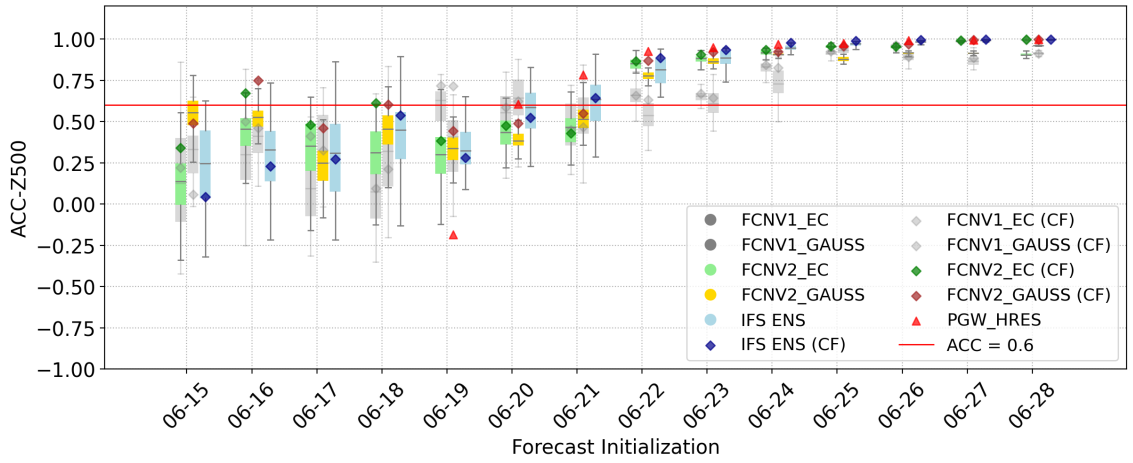


Figure 4.2: Distribution of ensemble forecasts of ACC of Z500 valid on 29 June 2021 00 UTC, averaged over the region ($145°W - 95°W, 30°N - 75°N$, described in Section 3.2.1), initialized between 15 and 28 June 2021 00 UTC. The box marks 25-75 quantiles, the whiskers are 1-99 quantiles, and the line is median. Colored diamonds represent the control forecast, colored triangles represent the deterministic forecast, and the red line represents an ACC standard value of 0.6.

anomaly valid on 29 June 00 UTC. The ACC is calculated between the model forecasts and ERA5 reanalysis data as the method described (Section 3.2.2), and the color scheme and plot elements follow the same conventions as described before (4.1.1).

First, the contrast in ensemble spread between FCNV2 initialized with Gaussian initial conditions (yellow box) and IFS ENS (blue box) is more noticeable in the forecast of geopotential height at a longer lead time, which further indicates the limitation of Gaussian initial condition. Next, we compare the ACC of forecasts with a threshold value of 0.6 (red line in Figure 4.2). For most forecasts initialized before 19 June, the ACC values remain below 0.6, indicating that both FCNV2 and IFS have limited skill in representing this high geopotential height anomaly before 19 June. The Pangu-Weather struggled to represent the upper-level ridge on June 19th, and the ACC value was negative, which corresponded with its inaccurate temperature forecast for that day (Figure 4.1).

For forecasts initialized on 20 June, the ACC of the ensemble median for IFS reaches 0.6, demonstrating improved skill in representing the upper-level ridge, although not fully correct. In contrast, the ACC of FCNV2 forecasts initialized with both IFS and Gaussian noise initial conditions remains below 0.6 until 21 June. Although Pangu-Weather had the poorest performance on 19 June, it showed a substantial improvement on 20 June, with the ACC value increasing from a negative value the previous day to 0.6. From forecasts initialized on 22 June, the ACC of all forecasts (except for FCNV1) surpasses 0.6, coinciding with the abrupt improvement in surface air temperature forecast (Section 4.1.1).

Compared to FCNV2 and IFS, FCNV1 (grey box) with both IFS and Gaussian noise initial conditions performs the worst, and it only crosses the ACC value of 0.6 on 24 June and still struggles in accurately representing the upper-level ridge even one day prior to the heat wave peak.

4.2 Analysis of forecast skill horizon

To quantify the skill of models in predicting the magnitude of the heat wave and the representation of upper-level ridge at the peak of the heat wave, in this section, the forecast skill horizon of two-meter temperature and 500 hPa geopotential height will be further analyzed based on forecast evolution in the last section (Section 4.1). As discussed in Section 3.2.2, the forecast skill horizon for two-meter temperature is defined as the lead time when the forecast bias of two-meter temperature becomes smaller than 5K. Based on the analysis presented in Section 4.1.1, this 5K bias baseline can provide an approximate estimation of when the forecast experienced a significant improvement, which likely occurred between June 21st and June 22nd. On June 21st, most of the forecast bias values were greater than 5K. To define a skillful 500 hPa geopotential height forecast, an ACC of 0.6 is used as the baseline since values below 0.6 are considered insufficient for accurately positioning synoptic-scale features. It needs to be noted that the forecast skill horizons for two-meter temperature and 500 hPa geopotential height should be interpreted separately, as the definition of "skillful" differs between these two variables.

We first start with the analysis of the forecast skill horizon of two-meter temperatures (Figure 4.3). Figure 4.3a and 4.3b show the forecast evolution of FCNV1 and FCNV2 ensemble forecasts, respectively, alongside the IFS ensemble forecast. Compared to the control forecasts (see Figure A.1), the ensemble forecasts exhibit greater consistency over initialization time. Further, forecast skill horizons defined before are extracted and compared across models in Figure 4.3c. Overall, the comparison shows that the forecast skill horizons for FCNV2, whether initialized with IFS initial conditions (green box) or Gaussian noise initial conditions (yellow box), are similar and comparable to the IFS ensemble. Both become skillful in predicting the heat wave magnitude at a lead time of around 7-8 days. Compared with FCNV2 and IFS, FCNV1 (grey box) exhibits shorter forecast skill horizons, lagging behind both FCNV2 and the IFS by about two days. FCNV1 initialized with IFS initial conditions performs slightly better than initialized with Gaussian noise initial conditions. Notably, the operational Pangu-Weather (red triangle) has a longer forecast skill horizon (8-9 days) than both IFS and FCNV2, which means its forecast bias is less than 5 K earlier than that of the other models (FCNV2 and IFS).

Next, the same analysis of the forecast skill horizon is applied to the 500 hPa geopotential height (Figure 4.4). Figure 4.4a and Figure 4.4b show the forecast evolution of ACC of 500 hPa geopotential height for FCNV1 and FCNV2, respectively, alongside the IFS ensemble forecast. Despite the detailed forecast evolution analysis discussed in the previous section, it needs to be noted that both FCNV2 and FCNV1 initialized with Gaussian noise initial conditions struggled to accurately represent the upper-level ridge, even at short lead times, as their ACC values remained below 1. (orange line in Figure 4.4a and 4.4b). This issue with 500 hPa geopotential height forecasts using Gaussian initial conditions was also identified by Bülte et al. (2024). Next, forecast skill horizons are extracted and compared across models in Figure 4.4c. IFS (blue box) exhibits a longer forecast horizon (8-9 days) for the 500 hPa geopotential height compared to FCNV1 (grey box) and FCNV2 (blue and green boxes). Pangu-Weather demonstrates its ability to represent this upper-level ridge at

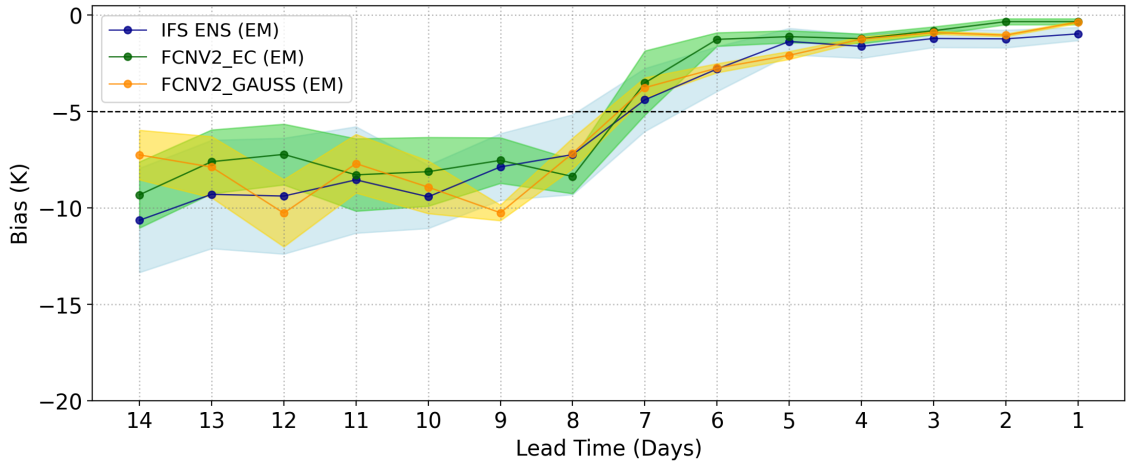
Model	T_{2m}	Z_{500}
IFS ENS	7.22 (7.45)	8.20 (8.36)
FCNV2_EC	7.31 (7.18)	7.61 (7.61)
FCNV2_GAUSS	7.36 (7.35)	7.69 (7.84)
FCNV1_EC	5.42 (5.46)	7.23 (7.28)
FCNV1_GAUSS	5.02 (5.43)	6.19 (7.19)
PGW_HRES	8.39	9.00

Table 4.1: Forecast skill horizons (in days) for two-meter temperature (T2m) and 500 hPa geopotential height (Z500). Values outside parentheses indicate the ensemble mean forecast skill horizon, while values inside parentheses represent the corresponding control forecast skill horizon. Note that the last row of Pangu-Weather represents the forecast skill horizon of a single deterministic forecast.

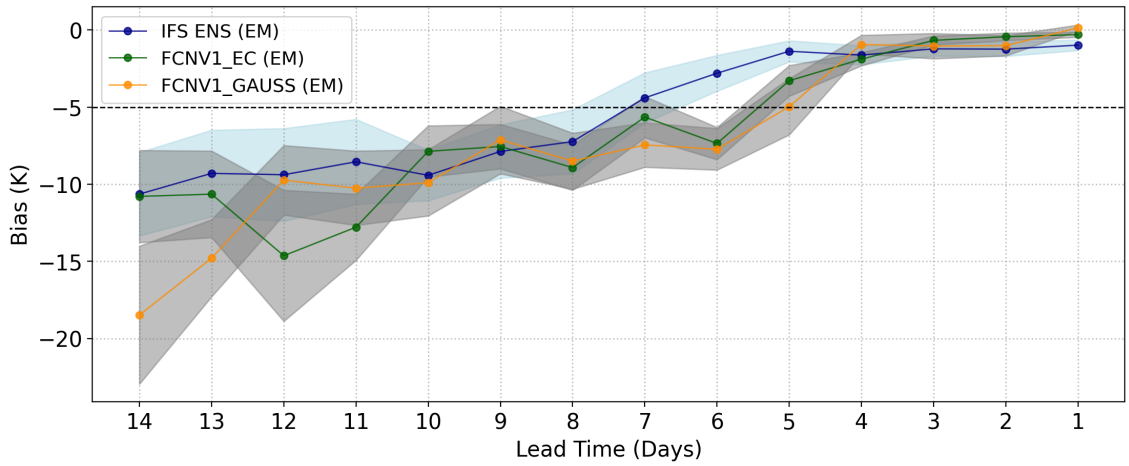
a lead time of 9 days. The significant improvement is more evident compared with control forecasts of IFS and FCNV2 (Figure A.2),

Finally, the forecast skill horizons for two-meter temperature and 500 hPa geopotential height across models are summarized in Table 4.1. FCNV2 initialized with either IFS or Gaussian initial conditions demonstrates similar skill in predicting the heat wave peak magnitude, achieving a temperature bias of less than 5K at a 7-day lead time, comparable to the IFS model. However, the IFS model captures the upper-level ridge 8 days before the peak, earlier than FCNV2. Pangu-Weather exhibits a longer forecast skill horizon, with a temperature bias of less than 5K around 8 days before the peak and skillfully representing the upper-level ridge 9 days prior. FCNV1 performed the worst and it only predicted the magnitude of heat wave peak forecasts at a lead time of 5 day and struggling to represent the upper-level ridge accurately. It is important to note that the forecast skill horizon is defined only with respect to the peak time of the heat wave, which might not provide a comprehensive picture of the performance of models during the entire heat wave period. Moreover, as mentioned by Pasche et al. (2024), although Pangu-Weather predicted the high temperature earlier, an analysis of the spatial distribution showed that it even predicted a higher temperature in the region where the ground truth data did not indicate one.

(a) Forecast evolution of 2-m temperature bias of FCNV2



(b) Forecast evolution of 2-m temperature bias of FCNV1



(c) Comparison of forecast skill horizon of 2-m temperature across models

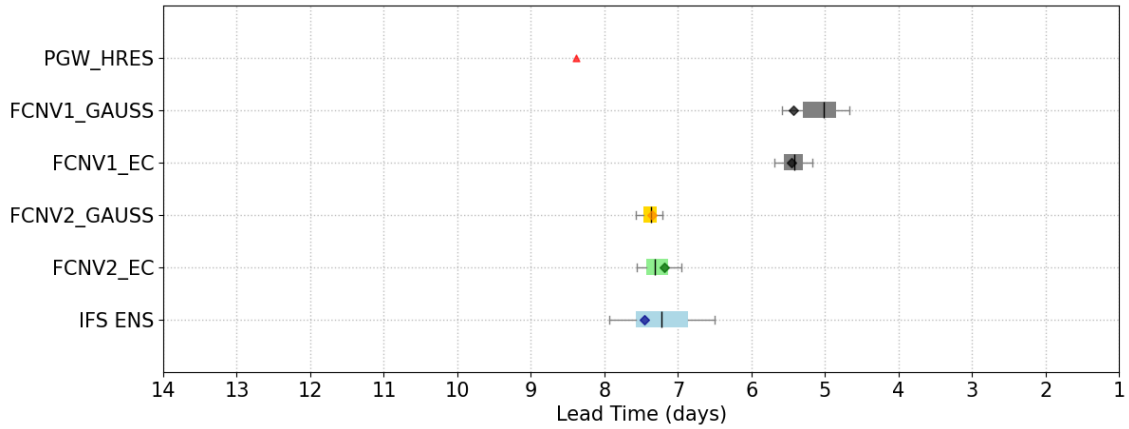
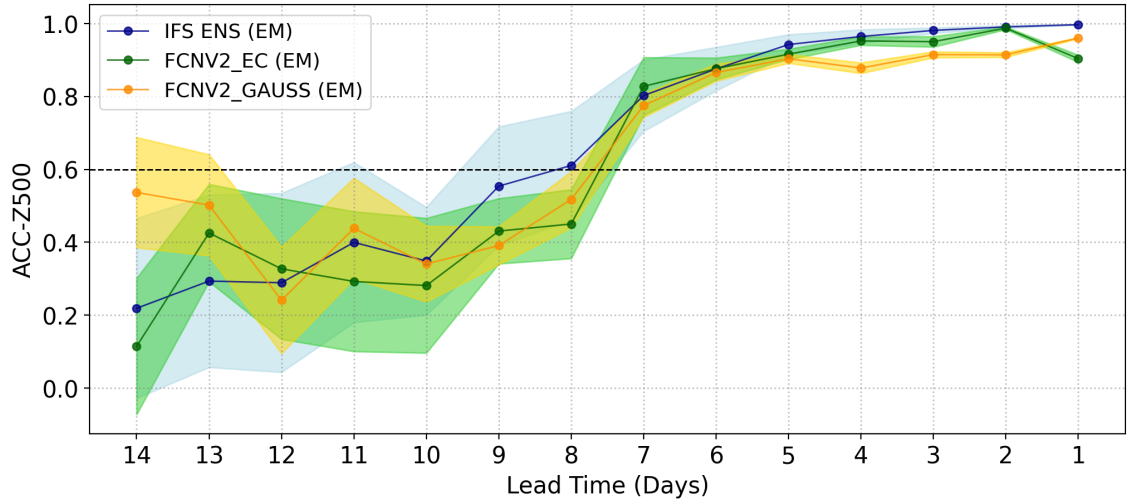
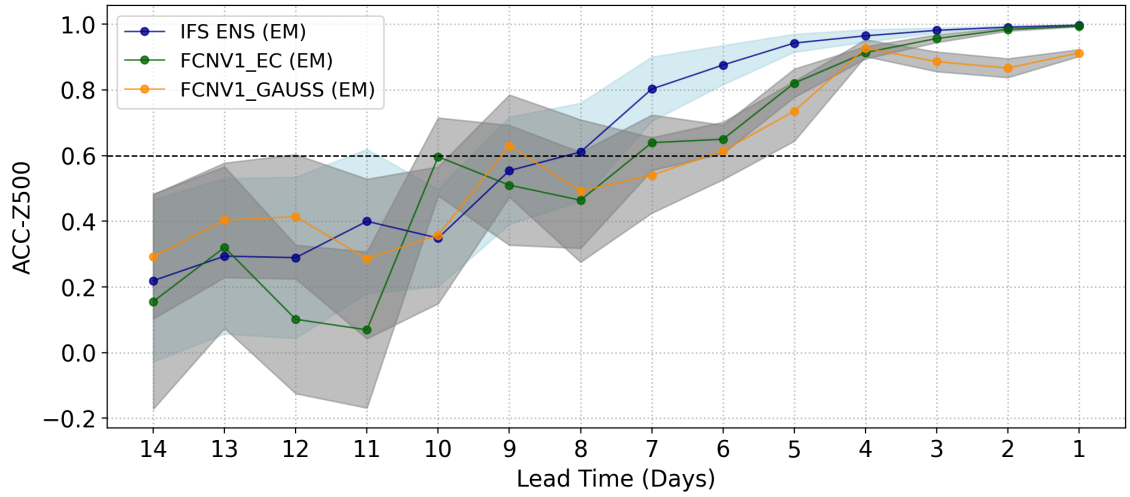


Figure 4.3: (a) Ensemble forecast evolution of two-meter temperature bias with respect to ERA5 reanalysis valid on 29 June 00UTC, initialized from 14 days to 1 day prior to 29 June 00UTC. The solid marked line represents the ensemble mean; the shading area represents ±1.5 standard deviations. The two-meter temperature bias averaged over 20° latitude by 20° longitude box. The dashed line represents a 5 Kelvin bias baseline. (b) Same as (a) but for FCNV1 ensemble forecast. (c) The forecast skill horizon of models for 2-m temperature. The line represents the forecast horizon of the ensemble mean, and the whisker represents the forecast horizon range, covering ±1.5 standard deviations from the mean forecast horizon. The diamond shape points represent control forecasts, and the triangle represents operation Pangu-Weather initialized with IFS high-resolution initial condition.

(a) Forecast evolution of 500 hPa geopotential height ACC for FCNV2



(b) Forecast evolution of 500 hPa geopotential height ACC for FCNV1



(c) Comparison of forecast skill horizon of 500 hPa geopotential height across models

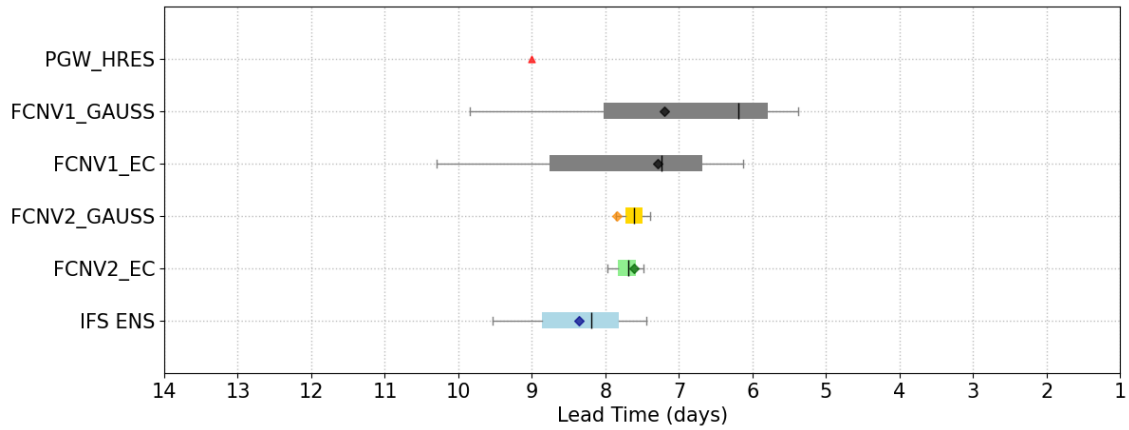


Figure 4.4: Same as Figure 4.3 but for ACC of 500 hPa geopotential height averaged over the region (145°W-95°W, 30°N-75°N).

5 Meteorological analysis of the 2021 Pacific Northwest Heat Wave

Results from the previous chapter revealed that data-driven models (FCNV2 and Pangu-Weather) predicted the peak magnitude during the 2021 Pacific Northwest Heat Wave with performance comparable to IFS. Although the predictive skill of representing the upper-level ridge was investigated, whether these data-driven models could effectively capture the link between the high-temperature anomaly and this upper-level ridge remains unclear. Secondly, while the large-scale circulation pattern acted as an important precursor during this event, the extreme near-surface temperatures would not have been as severe without additional thermodynamic processes involved in soil moisture deficits and upper-tropospheric heat as discussed in Section 2.3.2.

Thus, in this chapter, instead of focusing only on how the data-driven model depicts this extreme heat, the model evaluation is extended to the representation of drivers and processes that are already known to lead to the high near-surface temperatures during this event. By separating ensemble members of the data-driven model based on their predictive skill of near-surface air temperature (Method 3.2.4), the processes contributing to successful or unsuccessful heat wave predictions in the data-driven models can be studied and compared with the NWP model. Following an intensive analysis of the forecast on 29 June, the analysis in this chapter extends to the whole heat wave period (27 June to 1 July).

Section 5.1 investigates the representation of the large circulation pattern in data-driven ensemble forecasts. In Section 5.2, the time evolution of the vertical structure of temperature and moisture anomalies based on ERA5 reanalysis and associated thermodynamic processes are first discussed, laying the background for the next section. Lastly, Section 5.3 evaluates the representation of local thermodynamical processes in data-driven models, and the implication for near-surface air temperature diurnal evolution is discussed.

5.1 Representation of blocking patterns in the data-driven model

Figure 5.1 shows the composite-mean two-meter temperature anomaly and associated 500 hPa geopotential height anomaly (with respect to climatology) averaged from 27 June to 1 July for the "good members" in the FourCastNet2 and IFs ensemble forecasts initialized on 20 June, 22 June, and 24 June.

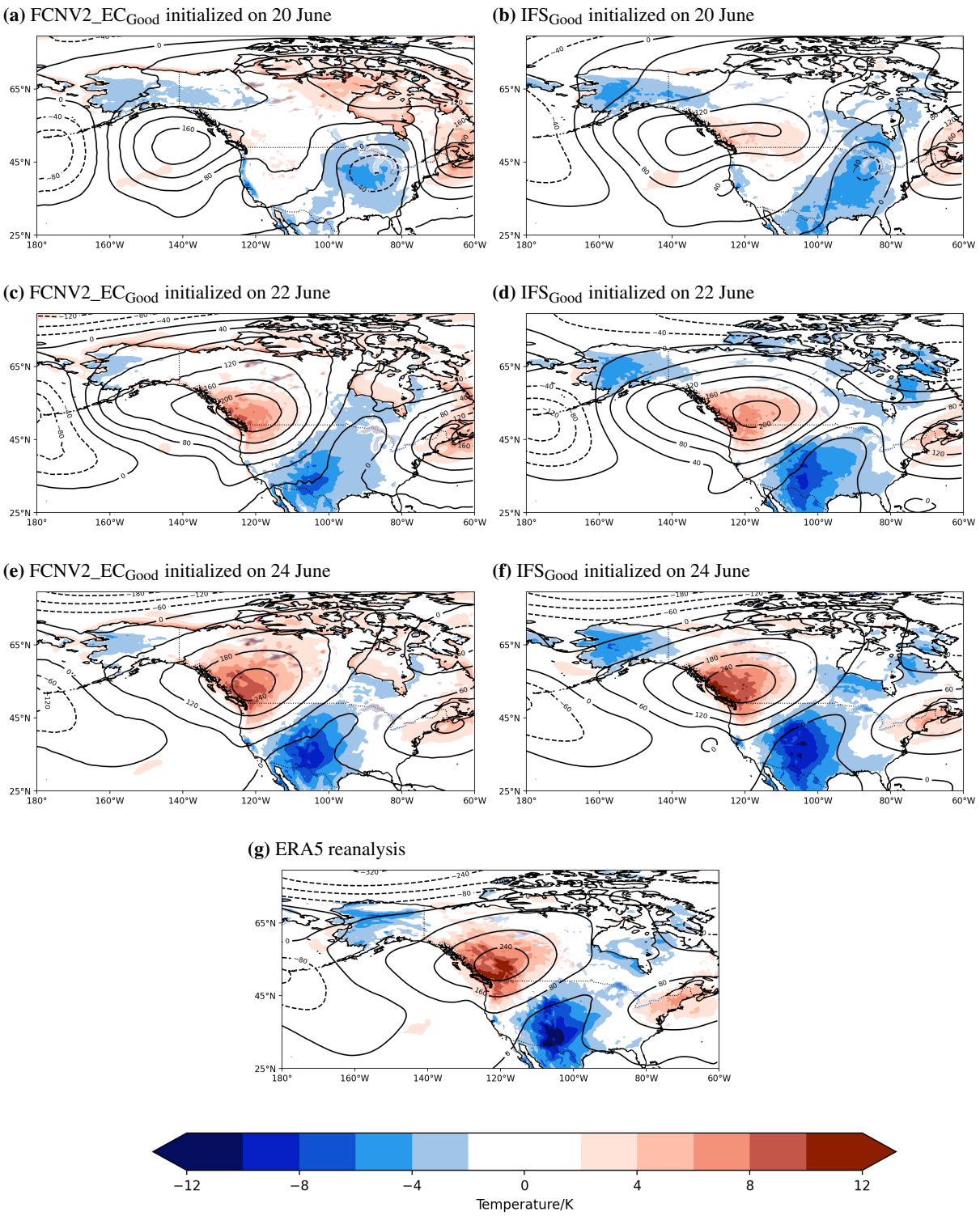


Figure 5.1: Composite mean of 2-m temperature anomaly (filled contours) and Z500 anomaly (solid black contours) for "good members" in FCNV2 initialized with IFS initial conditions (a, c, e) and IFS ENS (b, d, f) ensemble forecasts initialized on 20, 22, and 24 June, and ERA5 reanalysis (g), averaged from June 27 to July 1 (included), with respect to the ERA5 June-July climatology (1979-2019)

The "good members" are defined as those members with higher predictive skill in two-meter temperature in terms of RMSE. According to the ERA5 reanalysis (Figure 5.2g), an upper-level ridge characterized by an anomalous 500 hPa high geopotential height anomaly was located over Northwestern America during the heat wave (27 June to 01 July), exhibiting a high geopotential anomaly in the center up to 240 m. The two-meter temperature anomaly compared to climatology reached up to 12 K, and the location of the center of this blocking system coincides well with the high near-surface air temperature anomaly.

Next, we move the attention to the "good members" in the ensemble forecasts. In general, as the initialization time of the forecast is closer to the valid forecast time, both FCNV2 and IFS progressively improve in capturing the location and magnitude of the blocking system and associated high-temperature anomaly. On 20 June, while the IFS "good members" had already captured the reversal of the geopotential height anomaly gradient and the location of this system was already on the continent, the exceptional thickness was still underestimated (Figure 5.1b). In contrast, the FCNV2 "good members" struggled to predict the correct location of the center, with the predicted center of the blocking system still over the Northwest Pacific Ocean. (Figure 5.1a). By 22 June, both "good members" of IFS and FCNV2 predicted the location of the blocking system, but the magnitude of the high geopotential height anomaly remained underestimated. Concurrently, both models could predict the temperature anomaly, although the spatial extent and magnitude were still underestimated (Figure 5.1c and 5.1d). On 24 June, while the FCNV2 predicted the center location of the high geopotential height anomaly, the IFS model better captured the magnitude of the geopotential height anomaly at the center. As a result, compared to the ERA5 reanalysis, the IFS model more accurately depicted the spatial extent of the high-temperature anomaly. Although the IFS model still slightly underestimated the magnitude of the high-temperature anomaly, the FCNV2 underestimated the magnitude but depicted a larger spatial extent. (Figure 5.1e and 5.1f).

In contrast to the "good members" in the ensemble forecasts (Figure 5.1), the "bad members" (Figure 5.2) are defined as those members with lower predictive skill in two-meter temperature in terms of RMSE. A notable difference between the "good members" and "bad members" is identified in the forecasts initialized on 22 June (Figures 5.2c and 5.2d). The "good members" of both IFS and FCNV2 depicted a similar picture of the temperature anomaly and large-scale circulation pattern on this day. However, the difference between the "bad members" of FCNV2 and IFS is distinctive. The "bad members" in the IFS ensembles have already captured the shape and location of the blocking system and the associated temperature anomaly. In contrast, the "bad members" in the FCNV2 ensembles still struggled to capture the correct circulation pattern and the high-temperature anomaly. The observed difference between the "bad members" of FCNV2 and IFS suggests that FCNV2 may have more difficulty predicting anomalous conditions compared to IFS on 22 June. In other words, FCNV2 might be more significantly affected when extrapolating to such extreme conditions, as identified by Pasche et al. (2024). However, it is important to consider that the performance difference between "bad forecasts" of FCNV2 and IFS could also be influenced by the insufficient initial conditions used to generate the data-driven ensemble forecasts.

Despite the more pronounced divergence in performance between the ’good’ and ’bad’ members of the FCNV2 ensembles, the previous analysis still implies that FCNV2 may capture the relationship

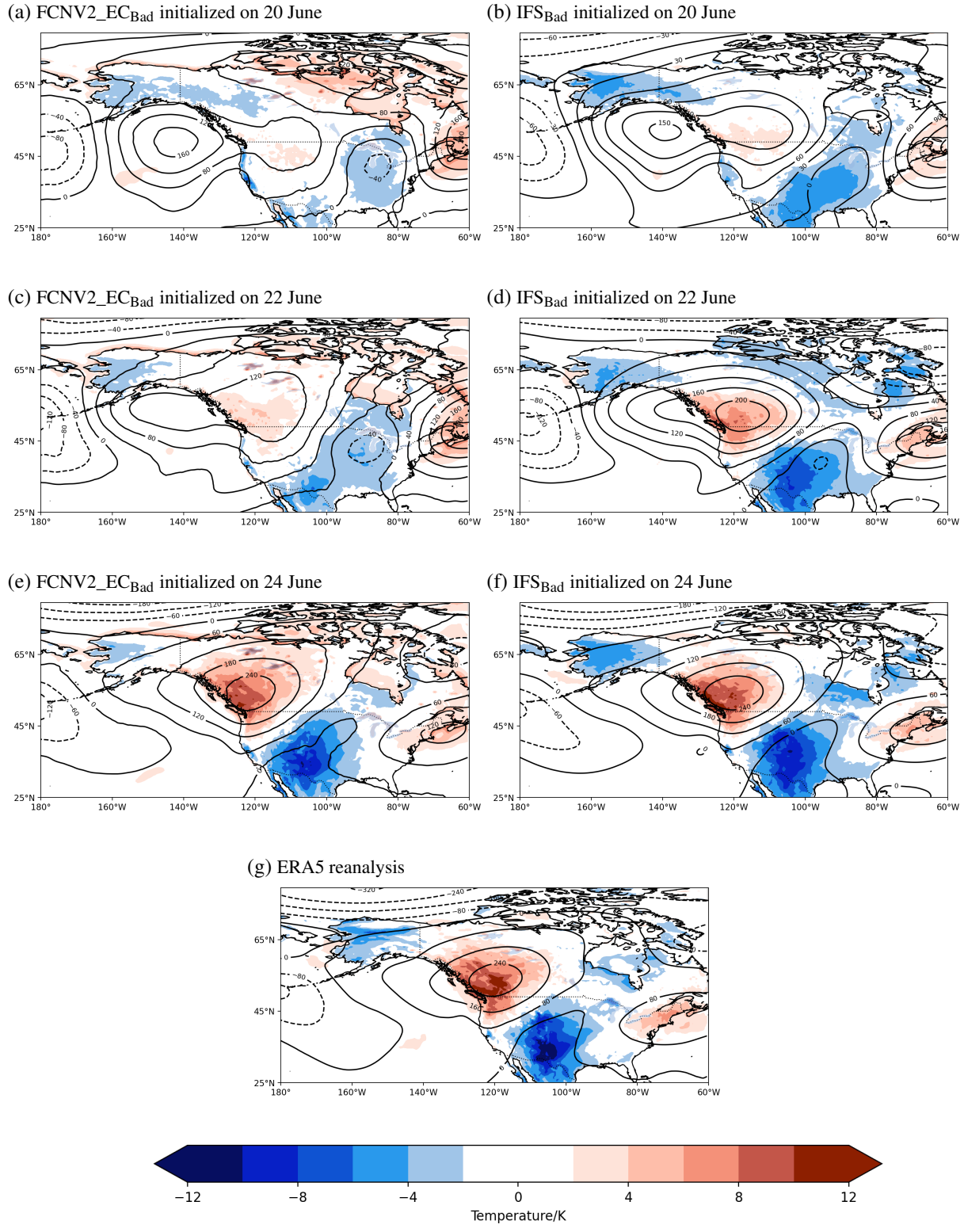


Figure 5.2: Composite mean of 2-m temperature anomaly (filled contours) and Z500 anomaly (solid black contours) for "bad members" in FCNV2 initialized with IFS initial conditions (a, c, e) and IFS ENS (b, d, f) ensemble forecasts initialized on 20, 22, and 24 June, and ERA5 reanalysis (g), averaged from June 27 to July 1 (included), with respect to the ERA5 June-July climatology (1979-2019)

between high-temperature anomalies and large-scale blocking patterns, as the members with good predictive skill in two-meter temperature forecasts also capture the blocking pattern better. Further, this relationship is evident from the spatial alignment of the temperature anomaly with the location of the high geopotential height anomaly in the FCNV2 ensembles.

5.2 Evolution of the vertical structure and associated processes

In this section, we first examine the temporal evolution of temperature and humidity vertical anomalies with respect to the climatology from 24 June to 5 July in the ERA5 reanalysis. This analysis aims to discuss several important processes involved in land-atmosphere feedback and upper-tropospheric heat during the development of the heat wave. It is important to note that our discussion of these processes is not only based on the analysis of the vertical profiles of temperature and humidity. Instead, it is also informed by and connected to many other studies, as the vertical profiles alone provide limited information.

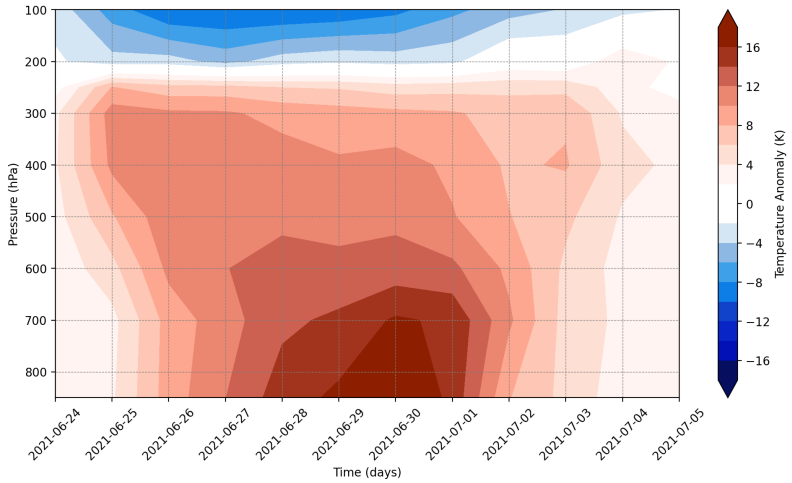
5.2.1 Evolution of the temperature anomaly vertical structure and associated processes

Figure 5.3a illustrates the temporal evolution of the vertical structure of temperature anomalies from 100 hPa to 850 hPa during the period from 24 June to 5 July. The figure reveals a distinct tilted pattern in the temperature anomalies throughout the vertical profile. Notably, significant temperature anomalies in the mid- to upper-tropospheric levels (300 hPa to 600 hPa) emerge from 24 June to 25 June, with anomalies exceeding 12 K. In contrast, during this time, the temperature anomalies in the lower troposphere (700 hPa to 850 hPa) only range from 2 K to 6 K. Hotz et al. (2023) used Lagrangian temperature anomaly decomposition to quantitatively analyze this upper tropospheric anomaly, separating it into positive advective diabatic and negative adiabatic components, suggesting that the warming of the air parcel in the upper troposphere is due to diabatic processes and the horizontal advection of warm air, which further confirms this warm aloft is a signature of WCBs (Section 2.3.1).

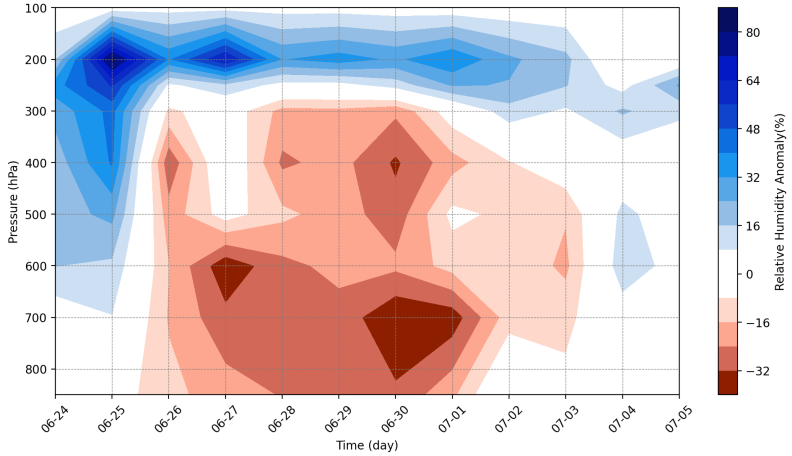
This heat signature of WCBs in the mid-to-upper-troposphere begins to weaken gradually on 27 June and continues until 2 July. Simultaneously, the temperature in the lower levels starts to rise, with the temperature anomaly beginning on 26 June and peaking between 29 June and 30 June, exceeding 16 K. This observed pattern further implies the top-down control of near-surface heat during this event, a significant feature as noted by Hotz et al. (2023) and Schumacher et al. (2022).

In addition to this feature, we can also identify the vertical extent of heat from 850 hPa to 600 hPa, indicating that the impact of the extreme heat event is not limited to the near-surface layers but extends to higher altitudes in the lower atmosphere. This vertical extent of heat may be related to atmospheric boundary layer processes and mixing within the lower atmosphere. As Schumacher

(a) Evolution of vertical temperature anomaly vertical structure



(b) Evolution of relative humidity anomaly vertical structure



(c) Evolution of specific humidity anomaly vertical structure

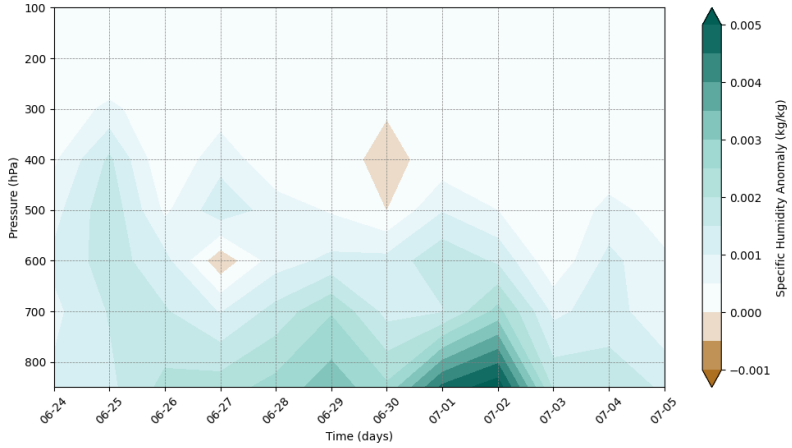


Figure 5.3: (a) Time heightight plot of temperature anomaly with respect to the June and July climatology from 1979 to 2019, averaged over the land domain ($49°N - 59°N$ and $115°W - 125°W$) between 24 June and 1 July, 2021. (b) Time-height plot of relative humidity anomaly. (c) Time-height plot of specific humidity anomaly.

et al. (2022) described, desiccated soils contributed to higher surface sensible heat fluxes, which transfer more heat from the land surface to the overlying air, steadily heating the planetary boundary layer. In addition to surface heating, warm air from the upper troposphere was gradually entrained and mixed into the growing PBL. The combined effect of surface sensible heat flux and warm air entrainment caused the PBL height to expand. As the PBL height increased, it was further connected with the upper-tropospheric heat. In the PBL, the heat kept accumulating as the process as Miralles et al. (2014) described: at night, the heat generated during the daytime was preserved in the residual layer. The next day, the heat from the residual layer was re-entrained into the PBL. Over several days of development, this heat cycle led to progressive heat accumulation in the PBL, further enhancing soil desiccation and escalating near-surface air temperatures.

But it needs to be noted this upper-tropospheric heat was not primarily responsible for the near-surface temperature anomaly, the contribution of the warm air aloft was more important in serving as a prerequisite for the heat accumulation in the PBL by suppressing moist convection that would have had a cooling effect (Schumacher et al., 2022; Papritz and Röthlisberger, 2023). In summary, the dry soil conditions, strong surface sensible heating, and the resulting deep PBL facilitated the mixing process and allowed for the accumulation of heat within the boundary layer over several days. These factors played a crucial role in the sustained development of the lower-level temperature anomaly, ultimately leading to extreme temperatures exceeding 16 K during the peak of the heat wave between 29 June and 30 June (Figure 5.3a).

5.2.2 Evolution of the moisture anomaly vertical structure and associated processes

The discussion in the last section suggests that the dry soil conditions played a crucial role in facilitating the growth of the PBL height. Through enhanced surface sensible heating, the dry soils further exacerbated the near-surface temperatures. Next, we examine the temporal evolution of specific humidity vertical profiles in ERA5 reanalysis and build a more comprehensive picture of the thermodynamical processes that contributed to the development of this extreme heat wave event.

As shown in Figure 5.3c, a column of positive specific humidity anomaly (0.001 to 0.002 kg/kg) extending from 850 hPa to the upper troposphere can be identified from 24 June to 26 June. Given the antecedent dry soil conditions and limited precipitation in the region prior to the onset of the heat wave (Section 2.3.2), which likely suppressed local evapotranspiration, the moisture contributing to the observed (ERA5 reanalysis) high specific humidity column from 24 June to 26 June is more likely advected from remote source. From 26 June, the moisture column height gradually decreased, and the high moisture column can be identified in the low-level atmosphere and persist until the end of the heat wave. The following discussion will focus on the source of this moisture and the associated processes.

During the 2021 Pacific Northwest heat wave, an anomalous warm-season atmospheric river (AR) played a significant role in transporting tropical moisture from Southeast Asia across the North

Pacific Ocean to North America (Mo et al., 2022). This AR made a prolonged landfall during the period of June 24-26, bringing a substantial amount of moisture into western Canada. Concurrently, the WCB ascent occurred over the East Pacific from 25 June to 28 June, with the outflow of WCB directly feeding into the upper-level ridge. The timing of the AR's landfall and the WCB's ascent suggests that the moisture transported by the AR likely provided moisture for the ascending WCB (Oertel et al., 2023). Thus, this specific humidity column from 24 June to 26 June might be attributed to the combined effects of ARS and WCBs during this period.

While the interaction between the AR and WCB may explain the moisture transport and the formation of the high specific humidity column between June 24 and 26 June, the trajectory analysis by Baier et al. (2023) provides additional insights and explains the second high specific humidity column (28 June to 30 June) in the lower level (Figure 5.3c). Baier et al. (2023) further traced back the air causing this heat wave 3 weeks before reaching Northwestern America, showing 9 to 6 days before this event (20 June to 23 June), the air ascended by the convective lift along the Meiyu-Baiu front over Eastern Asia and thus lead to strong diabatic heating, which is concentrated into two WCBs Oertel et al. (2023). During the last 6 days before the event (23 June to 29 June), the air mass that landed over Northwestern America descended and entered the PBL. Despite this descending motion, which typically leads to a decrease in moisture content, the specific humidity within the PBL remained high. This suggests that the increased lower atmospheric specific humidity observed between 28 to 30 June may have primarily been a result of the advection of the moist air mass from the tropics, which subsequently descended into the PBL. Given the dry soil conditions during this period, only a small amount of the observed moisture can be associated with local evaporative processes, as the dry soils would have limited the potential for evaporative cooling. (Figure 5.3c). However, it needs to be noted that there are no clear quantifications of the moisture sources in other studies. The description provided here is based on the interpretation of the observed specific humidity profiles and the air mass trajectory analysis conducted by Baier et al. (2023).

Though the positive specific humidity anomaly is observed during the heat wave period (Figure 5.3c), the negative relative humidity anomaly profile can be identified for the atmospheric column from the lower atmosphere to the upper troposphere (Figure 5.3b). The negative relative humidity anomaly suggests that despite the increased specific humidity, the air was still unsaturated relative to its temperature. This explains why the relatively high specific humidity did not lead to increased cloud formation or precipitation. Instead, the moisture from the West Pacific that mixed into the warm and dry air mass might act as a short-lived greenhouse gas, trapping solar radiation and further warming the lower atmosphere as a potential mechanism stated by (Mo et al., 2022). The negative relative humidity anomaly is also consistent with the clear sky conditions observed during this event. These clear sky conditions allowed for increased incoming shortwave radiation, which further contributed to the warming of the lower atmosphere during the heat wave.

Finally, the termination of this heat wave and the associated mechanism are briefly discussed. After 1 July, the temperature anomaly in the lower level significantly dropped, and by 02 July, the temperature of the whole atmospheric column had weakened, thus marking the end of this heat wave (Figure 5.3a). The termination of the heat wave occurred when the atmosphere was no longer stable as warming aloft entrained to lower-atmopshere, causing the convective damping of the

near-surface temperature (Zhang and Boos, 2023; Hotz et al., 2023). The high specific humidity anomaly and low relative humidity can also be observed after 1 July (Figure 5.3b). The increased moisture content in the atmosphere, as indicated by the specific humidity anomaly, likely provided the necessary conditions for the heavy precipitation that accompanied the termination of this heat wave Hotz et al. (2023).

In summary, by combining the analysis of the temporal evolution of the vertical structure of temperature and humidity during the 2021 Pacific Northwest heat wave with findings from other studies, several important processes leading to the development of this extreme event have been identified. First, the upper-tropospheric heat driven by WCBs during this event was an important prerequisite for heat accumulation in the lower atmosphere. The warm air aloft increased the stability of the atmospheric column and suppressed the cooling effect of moist convection, which would have otherwise limited near-surface temperature extremes. Second, dry soil conditions played a crucial role in modulating the surface energy balance. The dry soils facilitated the development of a deep planetary boundary layer (PBL), allowing the lower atmosphere to connect with the upper-tropospheric heat source. More importantly, the deep PBL enhanced heat accumulation near the surface, leading to high-temperature anomalies. This, in turn, exacerbated soil drying, creating a positive feedback loop that further amplified the heat wave.

The combination of these processes, including upper-tropospheric heat suppressing convective cooling and dry soils facilitating PBL growth and heat accumulation, created a favorable environment for the development and intensification of extreme surface temperatures during the 2021 PNW heat wave.

5.3 Representation of processes in the data-driven models

After discussing several important thermodynamical processes leading the high-temperature development during the 2021 Pacific Northwest heat wave using ERA5 reanalysis data, the following section will examine the time evolution of the vertical temperature and humidity profile forecast by data-driven models and how they represent the aforementioned processes in the 2021 Pacific Northwest Heat Wave.

5.3.1 Forecast evolution of vertical temperature structure

Good members and bad members in data-driven ensemble forecasts

We first examine the temporal evolution of the vertical temperature anomaly profile in good members and bad members of FCNV2 and FCNV1 ensembles. Figure 5.4 shows the evolution of the vertical temperature anomaly forecast initialized from 20 June to 24 June 00 UTC. The left column represents good members with better predictive skills for near-surface air temperature,

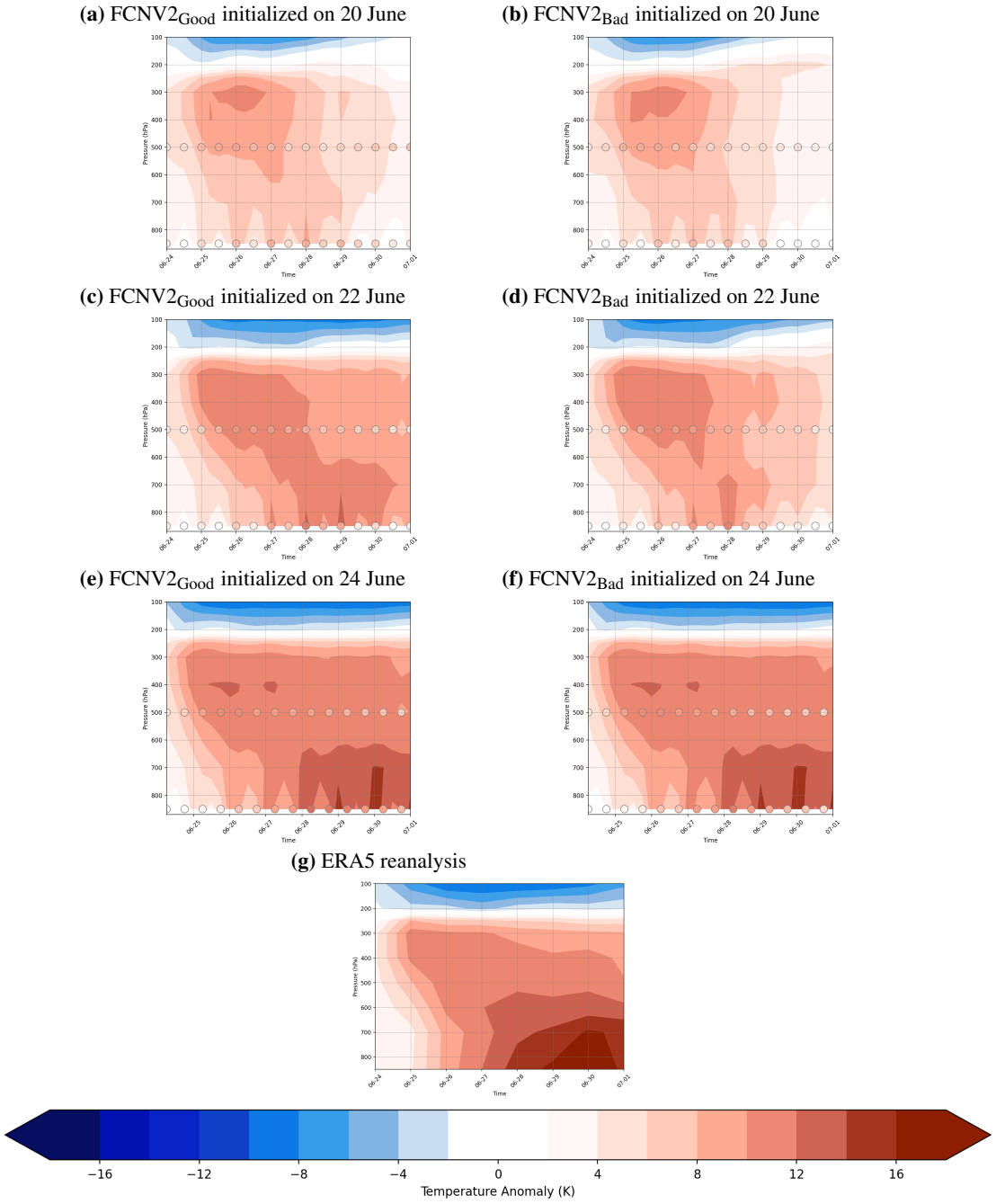


Figure 5.4: Time-height plots of temperature forecast anomalies (24 June - 1 July, 2021) of FourCastNet2 initialized with IFS initial conditions (FCNV2), averaged over the land domain, with respect to the June-July climatology (1979-2019). Scatter points at 500 hPa and 850 hPa represent FourCastNet1 (FCNV1) forecasts. The first column (a, c, e) shows the composite mean of "good members," while the second column (b, d, f) shows the composite mean of "bad members." Each row represents a different initialization time (20, 22, and 24 June 00 UTC). Last row: ERA5 reanalysis (g).

while the right column represents bad members with worse predictive skills. Scatter points at 500 hPa and 850 hPa indicate the FCNV1 ensemble forecasts, available only at these two levels.

Overall, both good members and bad members across lead times can predict that the heat first emerges from the upper-troposphere on 24 June to 25 June, though the forecast of the magnitude of this upper-tropospheric heat varies. Second, for both good members and bad members in the forecasts at earlier initialization time (Figure 5.4a, 5.4b) (20 June), they all struggle to predict the development of the temperature anomalies in the lower atmosphere during heat wave period.

Next, the difference between good members and bad members in FCNV2 across initialization time is discussed. From the forecast initialized on 18 June (Figure A.3a, A.3b), both good and bad ensemble members only capture the warm signal of upper-tropospheric heat. From the forecast initialized on 20 June (Figure 5.4a and 5.4b), both the good and bad ensemble members captured the signal of this warm aloft (300 hPa - 400 hPa), however, they still did not predict any further heat development in the lower-level atmosphere. A notable difference between the good and bad members is their representation of the heat mixing between 500 and 600 hPa levels.

The forecasts initialized on 22 June (Figure 5.4c and 5.4d) show the improvement forecast of upper-tropospheric heat, consistent with the time when it better predicted the upper-level ridge and near-surface temperature forecasts as we discussed in Section 4.2. Notably, compared to the bad members, the good ensemble members better depicted the persistence of upper-tropospheric heat and the mixing of heat between the upper-troposphere and lower atmosphere. From the forecast initialized on 24 June (Fig. 5.4e, f), the difference between good and bad members becomes smaller, they all capture the upper-tropospheric heat and the development of high temperature in the lower-level atmosphere during the heat wave period (27 June to 1 July). However, compared to ERA5 reanalysis, there is still a slight underestimation of the vertical extent of heat at lower-atmopshere.

Finally, the difference between FCNV1 and FCNV2 is discussed. For forecasts initialized at earlier initialization times (18 June and 20 June) (scatter points in Figure A.3a, b and Figure 5.4a, b), the difference between FCNV2 and FCNV1 is not significant as both of them have underestimated the magnitude of upper-tropospheric heat. The difference becomes larger for the forecast initialized on 22 June (Figure 5.4c, d). Compared to FCNV2, FCNV1 underestimates the upper-tropospheric heat more significantly and also fails to predict the heat development in the lower atmosphere. The poor performance of FCNV1 in representing upper-tropospheric heat and subsequent heat development is consistent with its poor performance in predicting near-surface temperature as discussed in Section 4.1. FCNV1 underestimated the upper-tropospheric heat at 500 hPa and likely failed to capture the vertical heat mixing between 500 hPa and 850 hPa.

In summary, as the initialization time progresses, FCNV2 exhibits an improved representation of the upper-tropospheric heat and its subsequent development in the lower levels. A comparison between the good and bad members reveals that the good members, characterized by better predictive skill for near-surface temperature, depict the heat mixing between the upper troposphere and lower levels more accurately. This heat mixing is linked to the development of a PBL, which is promoted by dry soil conditions and the entrainment of warm air from aloft into the PBL.

Inter-comparison between deterministic forecasts

Before going into the moisture vertical profile, we first discuss the vertical temperature anomaly profiles in Pangu-Weather and IFS HRES, and compared them with the FCNV2.

Fig 5.5 shows the evolution of vertical temperature forecast anomaly profile in IFS HRES (first row), FCNV2 (second row) and Pangu-Weather(third row), initialized from 18 June to 24 June. Comparing the forecast evolution across the three models, all three models start capturing the emergence of upper-tropospheric heat and the subsequent development of heat in the lower levels as the initialization time progresses from 18 June to 24 June.

Next, we turn our attention to the difference between each model. For IFS HRES, it have shown strong upper-trophspheric heat signal (10 - 12 K) from the forecast initialized on 18 June (Figure 5.5 (a)). However, for FCNV2 and PanguWeather, the magnitude of upper-tropospheric heat are still significantly underestimated (5.5(e), (i)). Moving to the forecasts initialized on 20 June (Fig 5.5 (b), (f), (j)), IFS HRES shows the onset of heat development at lower levels starting from 26 June, whereas FCNV2 fails to capture any signals of this lower-level heat development, and PanguWeather starts to depict the lower-level heat development but does not predict its prolonged duration. For the forecast initialized on 22 June (Fig 5.5 (c), (g), (k)), both IFS HRES and PanguWeather capture the upper-tropospheric heat and the subsequent heating in the lower-level atmosphere, albeit still underestimating the intensity of the strong heat at the lower level. In contrast, FCNV2 fails to capture the strong heating in both the upper troposphere and the heat development in the lower atmosphere. It is only in the forecasts initialized on 24 June (Fig 5.5 (d), (h), (l)) that all three models, including FCNV2, accurately depict this evolution of the vertical temperature anomaly profile. However, compared to IFS and Pangu-Weather, we can identify FCNV2 underestimated the high temperature anomaly in the lower-level atmosphere.

In summary, IFS HRES outperforms PanguWeather and FCNV2 in predicting the development of upper-level heat and subsequent heating in the lower atmosphere. IFS HRES captures the evolution of the vertical temperature profile earlier and more accurately than the other two models. While PanguWeather captures the lower-level heat development earlier than FCNV2, both models underestimate the intensity and duration of the heating compared to IFS HRES in the forecasts initialized before 22 June. FCNV2 lags behind IFS HRES and PanguWeather, only accurately representing the evolution of the vertical profile at the initialization time of 24 June.

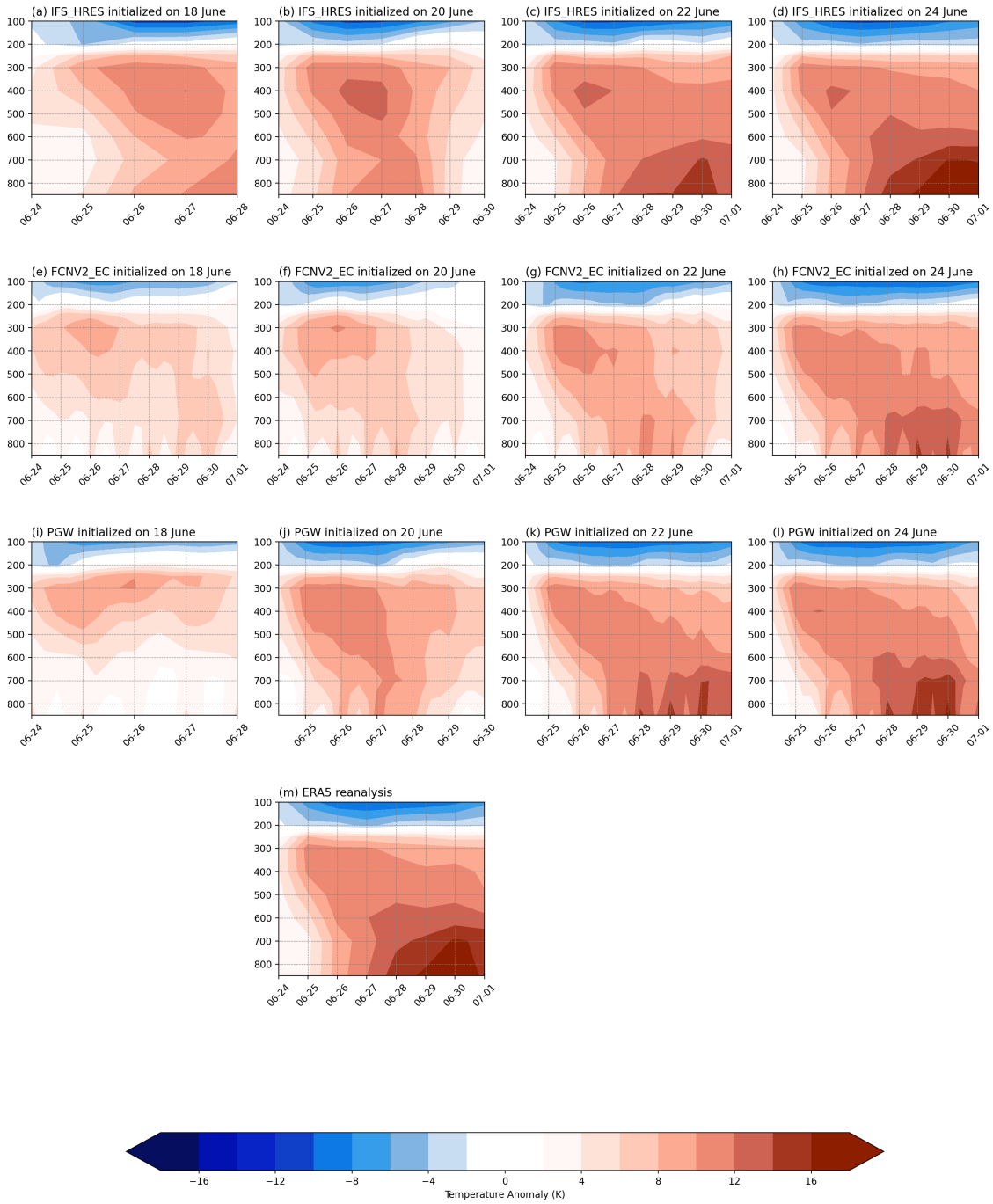


Figure 5.5: Time-height plots of temperature forecast anomalies, averaged over the domain from 24 June to 1 July, 2021, with respect to the June-July climatology (1979-2019) for deterministic forecasts. The first row (a-d) shows IFS HRES, the second row (e-h) shows FCNV2 initialized with IFS HRES initial conditions, and the third row (i-l) shows PanguWeather forecasts initialized with IFS HRES initial conditions. Each column represents a different initialization time from 18 June 00 UTC to 24 June 00 UTC. Note that IFS HRES has a temporal resolution of 24 h, while PanguWeather and FCNV2 have a resolution of 6 h.

5.3.2 Forecast evolution of vertical moisture structure

The analysis in Section 5.2 demonstrated that dry soil conditions played a critical role in the development of lower atmospheric temperatures during this heat wave event. Dry soil conditions

led to the creation of a deep PBL, allowing heat accumulation and mixing with upper-tropospheric heat within the PBL. In addition, dry soil facilitated the surface energy balance, promoting sensible heat flux over latent heat flux, thus contributing to heating the lower atmosphere (Schumacher et al., 2022). To indirectly investigate the representation of dry soil conditions and associated land-atmosphere feedback in data-driven models, this section examines their representation of the evolution of the vertical moisture profile.

Good members and bad members in FCNV2 ensemble forecast

We first examine the temporal evolution of the specific humidity anomaly vertical profile in good members and bad members in FCNV2 and FCNV1 ensembles. Fig. 5.6 shows the evolution of the specific humidity forecast initialized from 18 June to 24 June 00 UTC, the scatter points over 500hPa and 850 hPa represent the forecast of FCNV1 as it only available in these two levels.

To begin with the overall pattern of FCNV2 forecast(Figure 5.6), firstly, both good members and bad members across lead times can capture this high column of specific humidity extending from 850 hPa to the upper troposphere, which emerged from June 24 to June 26, as discussed in Section 5.2. Second, as the initialized time progresses, we can identify that forecasts start to capture the growing low-level moisture appearing on 29 June, through the magnitude varies. Notably, for the forecast initialized on 22 June (Figure 5.6c, d), both good and bad members overestimate the low-level moisture (0.005 kg/kg), while in the ERA5 reanalysis, it is only up to 0.0030 - 0.0035 kg/kg (Figure 5.6g). In contrast, the low-level moisture forecast initialized on 24 June (Fig 5.6e, f) predicts lower moisture compared to the forecast initialized on 22 June, but still remains slightly higher than the ERA5 reanalysis.

Next, the difference between good and bad members is discussed. From forecasts initialized on 18 June (Figure A.4a, A.4b), the good members captured the increase in low-level moisture better than the bad members during the heat wave peak (28 June to 30 June). On 20 June (Figure 5.6a, b), both good and bad members predicted an increase in low-level moisture reasonably well, but it needs to be noticed at this time (20 June) that the FCNV2 forecast of low-level temperature was still underestimated (Figure 5.4a, b). The forecast initialized on 22 June (Fig 5.6c, d) shows an overestimation of low-level moisture by both good and bad members, though the good members had a lower overestimation between 28 June and 29 June compared to the bad members. For the forecast initialized on 24 June (Fig 5.6e, f), the difference between the good member and bad members was not as evident.

Finally, we compare the evolution of vertical specific humidity anomaly between FCNV1 and FCNV2. Overall, the depiction of low-level moisture between FCNV1 and FCNV2 is similar. As the initialization time progresses, the forecast of low-level moisture also grows, and we can also identify the evident overestimation in FCNV1 that appeared from the forecast initialized on 22 June (Scatter points in Fig 5.6c, d).

In summary, FCNV2 represents the high moisture column that appeared on 24 June. However, both good members and bad members in FCNV2 overestimated the low-level moisture from the forecast initialized on 22 June. The overestimation of the good member is slightly lower than that of the bad member, but the difference is not evident. From the forecast initialized on 24 June, the overestimation seems to be "corrected."

In the last section, the difference between good members and bad members in depicting the vertical temperature anomaly profile is their representation of the mixing of heat between the upper troposphere and lower-level atmosphere. Connecting their depiction of the vertical profiles of moisture and temperature, one possible explanation for the overestimation of low-level moisture by FCNV2 during the heat wave peak from forecasts initialized on 22 June is that, while FCNV2 demonstrated improved predictability of the upper-level ridge and high-temperature anomaly on 22 June (see Fig 4.1 and 4.2), it does not initialize or account for soil moisture conditions, which suggests it may not adequately represent the temperature - soil moisture feedback (see Section2.3.1).

Consequently, FCNV2 could overestimate the evapotranspiration rate during the heat wave as it starts to predict the high near-surface temperature anomaly. This overestimation would lead to excessive moisture being transferred from the land surface to the lower atmosphere, resulting in the observed overestimation of low-level atmospheric moisture forecasts during the heat wave's peak. The lack of proper representation of the soil moisture deficit and its modulating effect on evapotranspiration in FCNV2 could be a key factor contributing to this overestimation. For the "corrected" moisture prediction from the forecast initialized two days later (24 June), one assumption is that as the initialization time approaches the actual event, the initial conditions might give the FCNV2 more information about the connection between temperature and moisture, thus FCNV2 could capture the right development of lower-level atmospheric moisture.

In the next section, we compare FCNV2 with IFS and PanguWeather models to further investigate this hypothesis. These comparisons can help us understand whether the overestimation of low-level moisture is specific to FCNV2.

Inter-comparison between deterministic forecasts

Fig 5.7 shows the evolution of vertical-specific humidity forecast anomaly profile in IFS HRES (first row), Pangu-Weather (second row), and FCNV2 (third row), initialized from 18 June to 24 June. Comparing the forecast evolution across the three models, all three models start to capture the high column of specific humidity extending from 850 hPa to the upper troposphere as the initialization time progresses from 18 June to 24 June.

Next, we shift our focus to the forecast of low-level moisture and compare the performance of different models. From forecasts initialized on 20 June (second column in fig 5.7), we can identify all three models (PanguWeather, IFS HRES, and FCNV2) begin to capture the development of low-level moisture on 29 June. Notably, Pangu-Weather (Fig 5.7 (j)) significantly overestimates

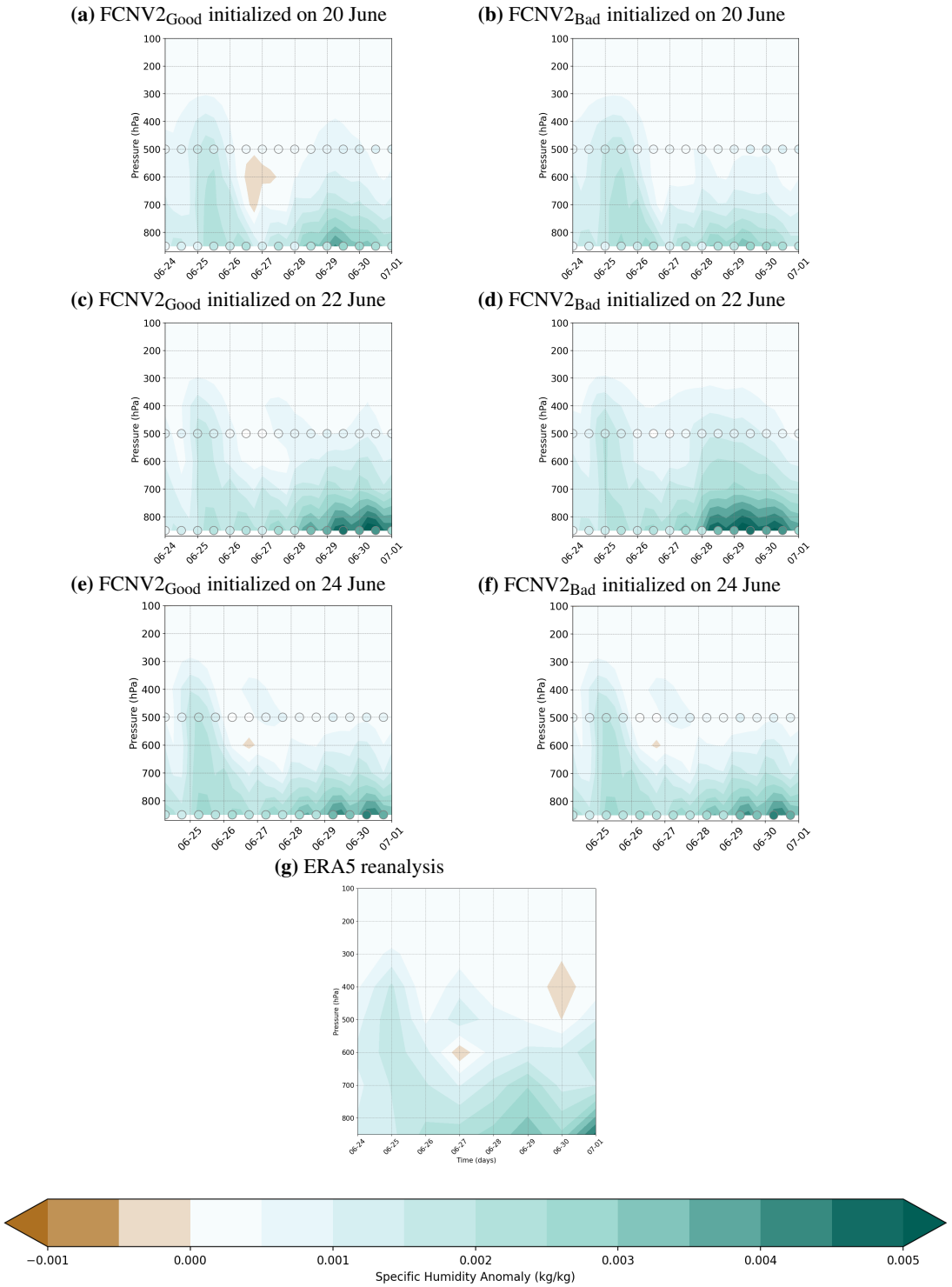


Figure 5.6: Time-height plots of specific humidity forecast anomalies (24 June - 1 July, 2021) of FourCastNet2 initialized with IFS initial conditions (FCNV2), averaged over the land domain, with respect to the June-July climatology (1979-2019). Scatter points at 500 hPa and 850 hPa represent FourCastNet1 (FCNV1) forecasts. The first column (a, c, e) shows the composite mean of "good members," while the second column (b, d, f) shows the composite mean of "bad members." Each row represents a different initialization time (20, 22, and 24 June 00 UTC). Last row: ERA5 reanalysis (g).

the low-level moisture, with values reaching up to 0.005 kg/kg. In contrast, both IFS HRES and FCNV2 do not exhibit any overestimation in their forecasts initialized on 20 June (Fig 5.7 (b), (f)).

Interestingly, when we examine the forecasts initialized on 22 June (third column in Figure 5.7), we notice Pangu-Weather appears to have "corrected" its previous overestimation of moisture. However, FCNV2 has started to overestimate the low-level moisture, which is consistent with our investigation in the ensemble forecast. Remarkably, IFS HRES provides a more accurate representation of low-level moisture across initialization times. From the forecast initialized on 24 June (fourth column in Figure 5.7), the moisture overestimation of FCNV2 improved but was still higher than the ERA5 reanalysis.

In summary, the comparison between models highlights that IFS HRES demonstrates a more consistent performance in representing the evolution of vertical specific humidity across initialization times, especially for the representation of low-level moisture. In contrast, Pangu-Weather and FCNV2 exhibit overestimation of the low-level atmospheric moisture at certain initialization times, though this overestimation is not consistent across all initialization times.

The inconsistent overestimation of low-level moisture in data-driven models could impact the modeled surface energy balance, as higher moisture content leads to increased evaporative cooling, potentially reducing the modeled surface temperatures. However, the evaluation of heat wave magnitude in the previous chapter 4 considered a larger region, which might have made this deficiency less apparent. To better understand the impact of the low-level moisture overestimation on the modeled near-surface air temperature in data-driven models, the next section will focus on a smaller region over land and investigate the diurnal evolution of near-surface air temperature.

Implication for near-surface air temperature diurnal evolution

Fig 5.8 shows the forecast of near-surface air temperature evolution during the heat wave period in FCNV2 and IFS ENS initialized from 20 June to 24 June. Before analyzing the forecast, We first examine the near-surface temperature evolution based on ERA5 reanalysis (red line in Figure 5.8). The near-surface air temperature gradually increased at the early stage of the heat wave from 27 June, and we can identify the diurnal cycle between 28 and 30 June (the peak of a heat wave) became quite stable.

The stable diurnal cycle during the peak of the heat wave shown in the ERA5 reanalysis reflected several important features during 2021 Pacific NorthWest Heat Wave. Firstly, the dry soil conditions limited the amount of moisture available for evaporative cooling, which typically helps moderate temperature fluctuations. With less moisture available, the surface heat flux was primarily driven by the incoming solar radiation during the day and the upward sensible heating from the warm, dry ground at night (Neal et al., 2022). Furthermore, we analyzed the negative relative humidity column in Section 5.2, which suggests the upper-tropospheric heat and the presence of a persistent high-pressure system during the heat wave likely contributed to the stable atmospheric conditions, with clear skies and minimal cloud cover. This allowed for uninterrupted solar heating during

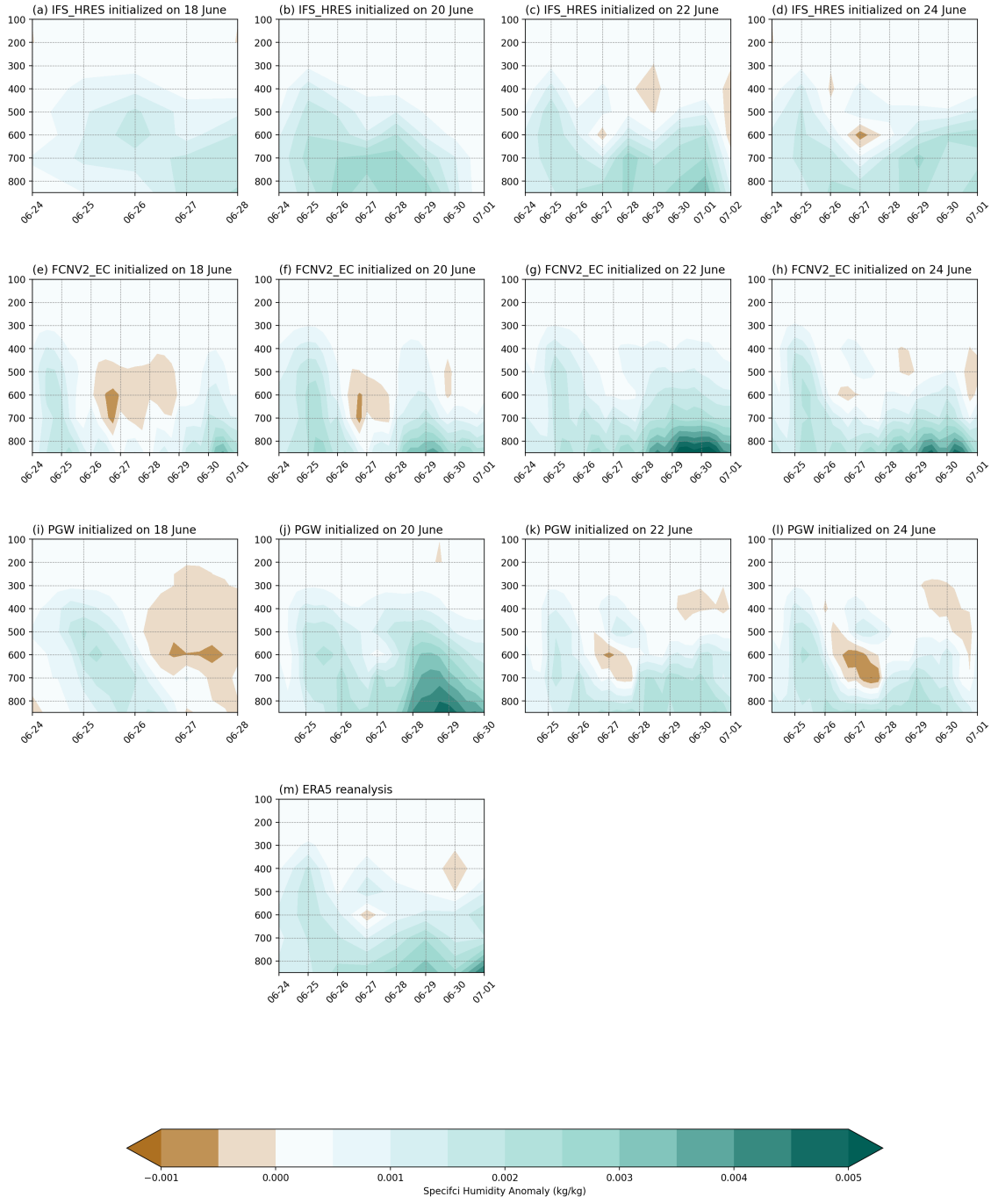


Figure 5.7: Time-height plots of specific humidity forecast anomalies, averaged over the domain from 24 June to 1 July, 2021, with respect to the June-July climatology (1979-2019) for deterministic forecasts. The first row (a-d) shows IFS HRES, the second row (e-h) shows FCNV2 initialized with IFS HRES initial conditions, and the third row (i-l) shows PanguWeather forecasts initialized with IFS HRES initial conditions. Each column represents a different initialization time from 18 June 00 UTC to 24 June 00 UTC. Note that IFS HRES has a temporal resolution of 24 h, while PanguWeather and FCNV2 have a resolution of 6 h.

the day and efficient radiative cooling at night, resulting in a consistent diurnal temperature cycle during the heat wave peak.

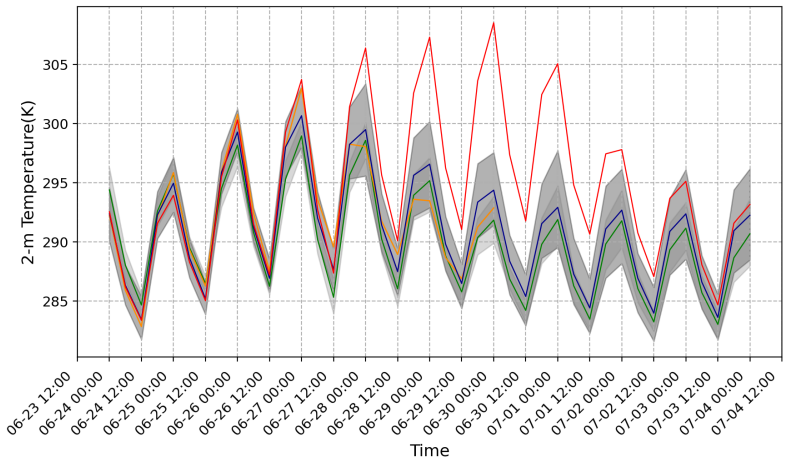
Next, we move attention to the differences in the forecast of the time evolution of two-meter temperature among the IFS ENS, FCNV2, and Pangu-Weather. From the forecasts initialized on 20 June (Figure 5.8a), we can identify that the maximum temperatures after 28 June, shown in the reanalysis, are significantly underestimated by all three models. The underestimation is most pronounced at 00 UTC, with a difference of more than 12 K. At this initialization time (20 June), the underestimation might largely be attributed to the misrepresentation of the large-scale circulation pattern, where the forecast skill of the circulation pattern is still limited (Figure 4.2). However, we can still observe that the underestimation between 28 June 00 UTC and 30 June 00 UTC is more pronounced in FCNV2 (green line) and Pangu-Weather (orange line) compared to the IFS (blue line). This underestimation in FCNV2 and PanguWeather is more evident at 00 UTC (16:00 local time) but not as pronounced at 12 UTC (04:00 local time).

Based on the forecasts initialized on 22 June (Figure 5.8b), all three models (IFS ENS, FCNV2, and PanguWeather) show significant improvement in their temperature forecasts compared to the forecasts initialized two days earlier. The improvement in temperature evolution forecasts could be largely due to the improved representation of large-scale circulation patterns in all three models at the time of initialization. Despite the overall improvement, IFS and FCNV2 still underestimate temperatures from June 28 to July 1, while the forecast of PanguWeather is closer to the time evolution depicted in ERA5 reanalysis. Moving on to the forecast initialized on June 24 (Fig. 5.8c), we see that all three models can reasonably predict the time evolution at the heat wave period.

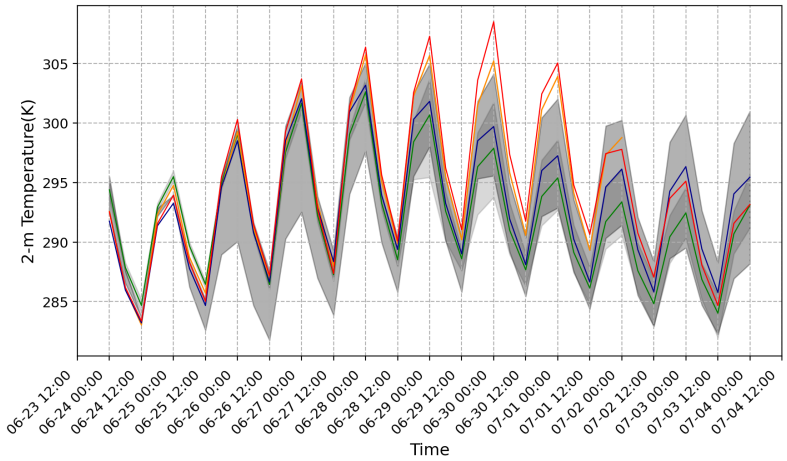
As shown before, compared to IFS, Pangu-Weather and FCNV2 have a larger underestimation of temperature at 00 UTC than at 12 UTC, resulting in a smaller diurnal temperature range. We further analyze the diurnal evolution aggregated between 28 June 00 UTC and 30 June 00 UTC, as shown in Figure 5.9. The composite diurnal evolution shows that from the forecast initialized on 20 June (Figure 5.9a), while all models underestimate the daytime peak temperature, PanguWeather exhibits a smaller diurnal evolution than other two models. From the forecast initialized on 22 June (Figure 5.9b), though all three models experience a smaller diurnal cycle than the ERA5 reanalysis, Pangu-Weather depicts the diurnal evolution closer to ERA5 reanalysis. The forecast of FCNV2 has a smaller diurnal cycle than IFS and Pangu-Weather. From the forecast initialized on 24 June, we can identify that three models all represent the diurnal evolution well match to ERA5 reanalysis (Figure 5.9c).

The moisture overestimation observed in PanguWeather (see Figure 5.7(j)) may contribute to the differences in the diurnal temperature evolution between PanguWeather and IFS in the forecast initialized on June 20 ((Figure 5.9a). During the daytime, from 28 June to 30 June, the overestimated moisture in the lower atmosphere by PanguWeather suggests the overestimation of evaporative cooling, as it would overestimate the amount of energy consumed by evapotranspiration rather than being used for sensible heating. This extra latent cooling may suppress daily maximum temperatures more than what was observed (ERA5). Furthermore, the excessive moisture predicted by PanguWeather may promote the formation of low clouds, which are effective at reflecting

(a) Forecasts of time evolution initialized on 20 June 00UTC



(b) Forecasts of time evolution initialized on 22 June 00UTC



(c) Forecasts of time evolution initialized on 24 June 00UTC

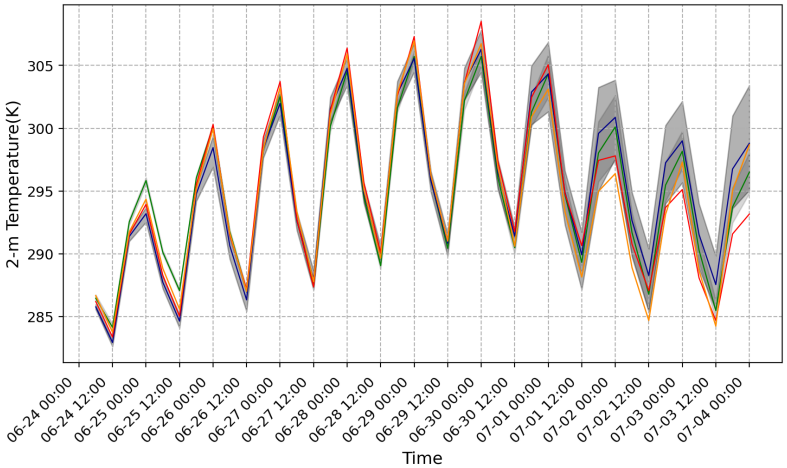
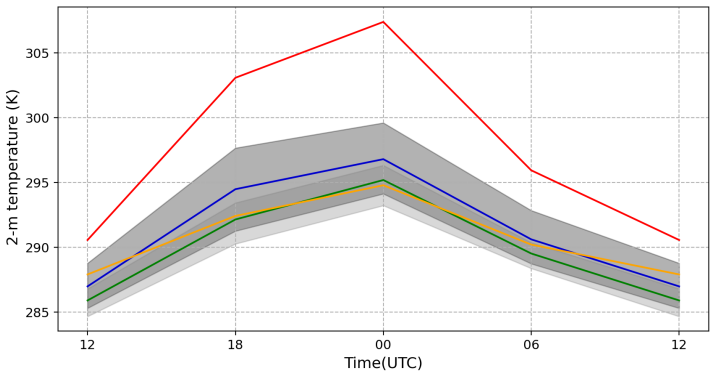


Figure 5.8: Time evolution of 2-meter temperature during the heat wave period (24 June to 4 July). Lines represent ensemble means: blue (IFS ENS), green (FourCastNet2 initialized with IFS), orange (Pangu-Weather initialized with IFS HRES). Red line: ERA5 reanalysis. Shading: 10-90% range of ensemble forecasts, light grey (FourCastNet2), dark grey (IFS ENS).

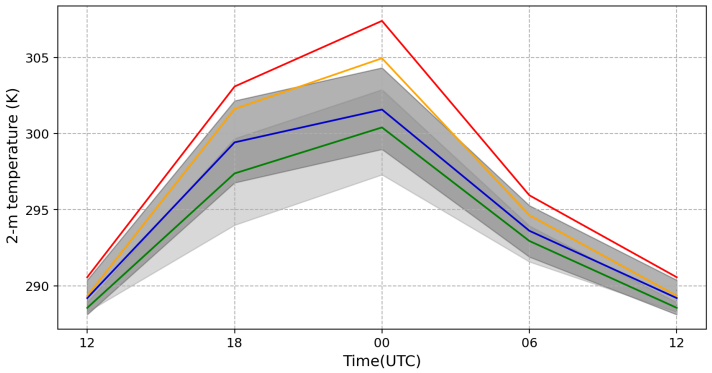
incoming solar radiation. Compared to the clear sky situation during the 2021 Pacific Northwest heat wave, this cloud radiative effect would reduce surface heating and hinder the rise in daily maximum temperatures. Conversely, the overestimation of atmospheric moisture can reduce radiative cooling of the surface at nighttime. Therefore, the overestimation of low-level atmospheric moisture in PanguWeather from the forecast initialized on 20 June may indicated by the dampened diurnal temperature evolution compared to IFS. Following the 'correction' of the low-level moisture forecast in Pangu-Weather at the initialization on June 22, the previously observed diurnal evolution discrepancy between IFS and Pangu-Weather is no longer evident (Figure 5.9b).

Similarly, the dampened diurnal temperature evolution of FCNV2 can also be identified by comparing it with IFS (Figure 5.9b), though the dampened diurnal evolution of FCNV2 is not evident as in Pangu-Weather, this may mainly be because the aggregated time we choose (only the peak of heat wave), as shown in Figure 5.8b, FCNV2 experienced a smaller diurnal evolution compared to IFS is most evident between 28 June and till 7 July.

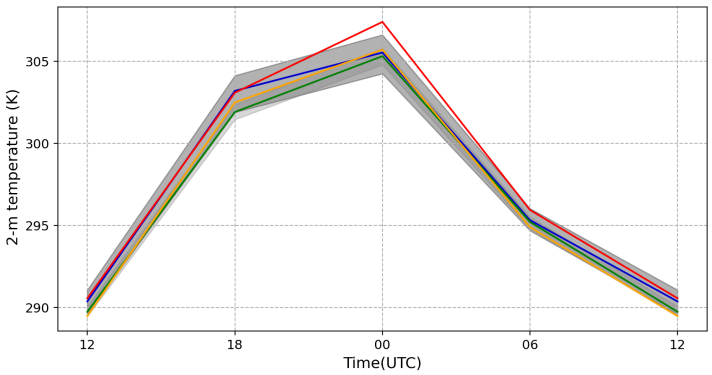
(a) Composite diurnal evolution initialized on 20 June 00 UTC



(b) Composite diurnal evolution initialized on 22 June 00 UTC



(c) Composite diurnal evolution initialized on 24 June 00 UTC



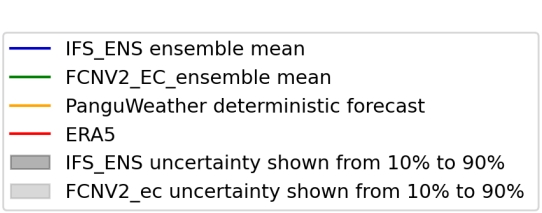


Figure 5.9: Composite diurnal evolution initialized from 20 June 00UTC to 24 June 00UTC, aggregated between 28 June 00 UTC and 30 June 00 UTC (included). Lines represent ensemble means: blue (IFS ENS), green (FourCastNet2 initialized with IFS IC), orange (Pangu-Weather initialized with IFS HRES IC). Red line: ERA5 reanalysis. Shading: 10-90% range of ensemble forecasts, light grey (FourCastNet2), dark grey (IFS ENS)

6 Conclusion and discussion

The development and maintenance of heat waves involve various drivers and their complex interplay across different spatial and temporal scales. Moreover, in the context of climate change, the non-linear effect of global warming increases the likelihood of unprecedented and more extreme heat waves (Domeisen et al., 2022b). While state-of-the-art numerical weather prediction models have shown significant improvement in heat wave prediction and even demonstrated potential for extended-range forecasts on the time scale of 3-4 weeks, models still exhibit errors across the entire range of heat wave drivers. Substantial improvements in heatwave prediction in the short term remain unlikely (Barriopedro et al., 2023). In recent years, data-driven models have experienced rapid advancements and demonstrated comparable performance to physics-based NWP models for medium-range weather forecasting. Despite showing promising average performance, their ability to predict extreme events is still uncertain, as these rare conditions often lie outside of the training data distribution, limiting their ability to extrapolate reliably (Olivetti and Messori, 2024).

The main research aim of this thesis is to investigate the uncertainty and potential of data-driven models in predicting extreme heat events and to compare their performance with a state-of-the-art NWP model through a case study that is totally out-of-sample. More importantly, this thesis aims to evaluate the performance of data-driven models in capturing and representing the key drivers and important processes that contribute to the development of heat waves. The record-breaking 2021 Pacific Northwest Heat Wave is selected as the case study because it was so extreme that it far exceeded the range of historical temperature observations, making it even more challenging to define the return period with confidence (Philip et al., 2022). To quantify uncertainty in heat wave prediction using data-driven models, ensemble forecasts are generated from two versions of the FourCastNet using IFS initial conditions and Gaussian noise initial conditions to compare with IFS ENS. Additionally, the Pangu-Weather deterministic forecasts are included for inter-comparison between models. After a detailed evaluation of the predictive skill for the magnitude of the heat wave peak, the representation of the large-scale circulation pattern and local thermodynamic processes in data-driven models is investigated. In the following discussion, the research questions raised in Chapter 1 are addressed:

1. **At what lead time do data-driven models start providing skillful predictions of the intensity for the peak of the heat wave and the associated anomalous atmospheric circulation pattern, and how do they compare with numerical weather prediction models?**

   In predicting the peak magnitude during the 2021 PNW Heat Wave, all models (FourCastNet1, FourCastNet2, Pangu-Weather, and IFS) were not accurately capturing the extreme magnitude

beyond the lead time of 7 days, and they all experienced significant improvement before they could capture the extreme magnitude.

When comparing the lead times at which their forecast bias of two-meter temperature started to become less than 5 K (considered skillful for predicting the heat wave peak magnitude), ensemble forecasts of FourCastNet2, initialized with either IFS initial conditions or Gaussian noise initial conditions, demonstrated similar forecast skill. They could predict the peak magnitude 7 days ahead of the peak, comparable to the performance of the IFS ensembles. For Pangu-Weather, its bias falls below 5 K around 8 days before the peak of the heat wave. However, at a longer lead time of 10 days, Pangu-Weather experienced the largest errors compared to IFS and FourCastNet2. In contrast, FourCastNet1 performed the worst, only capturing the peak magnitude 5 days ahead of the peak. At the lead time of 6 days, all models except FourCastNet1 could accurately predict the peak magnitude of this heat wave.

In predicting the associated anomalous atmospheric circulation pattern, represented by the 500 hPa geopotential height anomaly, IFS ensemble forecasts showed a skillful representation of this pattern at a lead time of 8 days. FourCastNet2 performed slightly worse, demonstrating skillful predictions around 7 days before the peak. Pangu-Weather demonstrated a skillful representation of this anomalous circulation pattern as early as 9 days but showed almost no skill at the lead time of 10 days. FourCastNet1 struggled the most, failing to represent this circulation pattern accurately, even at short lead times (5 days).

2. **To what extent can data-driven model capture the relationship between extreme temperature anomalies and the associated anomalous large-scale atmospheric circulation patterns?**

   By grouping the ensemble members of FourCastNet2 based on their predictive skill for near-surface air temperatures and categorizing them as "good members" and "bad members," it is found that those members that more accurately predict near-surface air temperatures during the heat wave period also better represent the position and magnitude of the anomalous large-scale circulation patterns. This implies that FourCastNet2 may have learned the link between the high surface temperature anomaly and this large-scale circulation pattern. Additionally, the location of the high near-surface temperature anomalies predicted by FourCastNet2 aligned well with the high-pressure center. However, while the "good members" of FourCastNet2 performed comparably to the "good members" of the IFS ensemble, the "bad members" of FourCastNet2 struggled more with capturing high-temperature anomalies and associated circulation patterns compared to the "bad members" of IFS.

3. **How do data-driven models represent the local thermodynamical processes during the heat wave?**

   Several thermodynamical processes involving upper-tropospheric heat and land-atmosphere feedback represented by data-driven models are investigated indirectly by examining the

forecasts of the vertical profiles of temperature and specific humidity anomalies. First, the data-driven models (Pangu-Weather and FourCastNet2) have the ability to predict the upper-tropospheric heat and the subsequent heat development in the lower atmosphere, but the timing of their predictions differs. For FourCastNet2, while it captures the upper-tropospheric heat signal at earlier initialization times, it only predicts the magnitude of upper-tropospheric heat and lower-atmosphere heat development at later initialization times (24 June). Pangu-Weather predicts both upper-tropospheric heat and lower-level heat development earlier than FourCastNet2 (22 June). In contrast, IFS consistently depicts the vertical temperature anomaly profile, capturing upper-tropospheric heat and subsequent lower-atmosphere heat development earlier (20 June) than both FourCastNet2 and Pangu-Weather. For FourCastNet1, although it captures the upper-tropospheric heat signal, the magnitude is significantly underestimated, and it fails to capture the subsequent heat development in the lower-level atmosphere.

Notably, the investigation of the forecast evolution of the vertical moisture anomaly profile reveals that FourCastNet2 and Pangu-Weather tend to overestimate low-level atmospheric moisture at certain initialization times (20 June for Pangu-Weather, 22 June for FourCastNet2). In contrast, IFS does not exhibit this overestimation of moisture across lead times. This discrepancy is further indicated by the difference in their forecasts of the diurnal evolution of near-surface air temperature, where Pangu-Weather and FourCastNet2 predict a smaller diurnal temperature range than IFS when the overestimation of moisture occurs. The overestimation of low-level atmospheric moisture by Pangu-Weather and FourCastNet2 suggests that these models may not adequately represent the land-atmosphere feedback and the surface energy budget. During daytime, the overestimated moisture suggests that models tend to predict more evaporative cooling, resulting in less energy available for sensible heating. Additionally, the overestimated moisture favors cloud formation, reducing incoming solar radiation and decreasing the energy available for surface heating. At night, the overestimated low-level moisture contributes to higher minimum temperatures due to an enhanced greenhouse effect trapping outgoing longwave radiation.

By evaluating the predictive skill of data-driven models and investigating their representation of large-scale circulation patterns and local thermodynamical processes, this thesis provides unique insights into the potential strengths and weaknesses of data-driven models in an out-of-sample case. During the 2021 PNW Heat Wave, FourCastNet2 and Pangu-Weather showed comparable skill to IFS ENS in predicting the peak magnitude of the heat wave and the associated large-scale circulation pattern. However, the investigation into the representation of local thermodynamic processes in data-driven models suggests that the IFS provides more robust and consistent predictions in terms of the vertical profile of temperature and moisture anomalies. In contrast, data-driven models tend to struggle with vertical profiling of moisture and overestimate low-level atmospheric moisture at certain initialization times. This limitation suggests that these models might not effectively capture the feedback between the surface and the atmosphere, as they do not include meteorological variables related to surface conditions. On the other hand, the IFS explicitly includes a scheme (TESSEL) (see Section 2.3.1) to represent the coupling between the surface and the atmosphere,

leading to a more accurate representations of low-level moisture. However, the findings only indirectly link the overestimation of moisture in data-driven models to near-surface air temperature by suggesting its impact on the diurnal evolution of temperature, rather than establishing a direct causal relationship. Thus, future studies involving comprehensive sensitivity testing of data-driven models are crucial to understand the dependency between predicted variables and the underlying physical processes represented by data-driven models.

Pasche et al. (2024) also evaluated data-driven models and compared them to IFS HRES during the 2021 PNW Heat Wave, focusing only on deterministic forecasts rather than probabilistic forecasts. They found that although Pangu-Weather and GraphCast provided comparable forecasts during the heat wave when compared to IFS HRES, data-driven models struggled more with extrapolating such extreme conditions. They pointed out that the models faced the most difficulty on the peak heatwave days, not primarily due to longer lead times from their initialization but because of the inherent predictability barrier of the extreme situation itself, regardless of the initialization time. This finding aligns with the finding of this thesis, which observed that all models experienced significant improvement in their forecast evolution around one week before the heat wave peak. This further illustrates that the extreme nature of the heat wave itself posed the biggest challenge to predictability for both data-driven and NWP models.

In addition to comparing data-driven models and NWP models, this thesis compared two versions of the FourCastNet model. In the evaluation of FourCastNet1, it was discovered that its forecasts contained significant outliers (not shown in the thesis). These outliers were later identified as artifacts resulting from issues with the model architecture. By improving the architecture in FourCastNet2, the model demonstrated increased stability over longer lead times. Another key difference between FourCastNet1 and FourCastNet2 was the input data, particularly regarding vertical level information. Although both models underwent the same training process, FourCastNet2 had access to more detailed vertical level data as inputs. This additional input information likely contributed to the improved performance of FourCastNet2, especially in cases where the vertical temperature profile was critical for surface heat development.

In this thesis, the evaluation has only focused on forecasts initialized at 00 UTC, using ERA5 as the ground truth for validating the forecasts. However, it is important to note that ERA5 initialized at 00 UTC has a larger assimilation window (9 hours) compared to the IFS forecasts (3 hours). To enable a more comprehensive and fair comparison between NWP models and data-driven models, future studies should extend the analysis to include forecasts initialized at different times. Additionally, instead of using ERA5 as the ground truth, the forecasts of NWP models should consider comparing against their own analysis, which would give a more fair comparison for IFS, especially at a short lead time.

This thesis employed two initial conditions to generate ensemble forecasts in a data-driven weather prediction model. Both approaches rely on running the model multiple times with slightly varying initial conditions. Consequently, the resulting ensembles can only capture uncertainties arising from the initial conditions but fail to account for uncertainties inherent in the model itself (Bülte et al., 2024). In contrast, NWP models include stochastic schemes to represent uncertainties due to model

integration, leading to more reliable probabilistic forecasts, which are especially crucial for extreme events (Leutbecher et al., 2017). Thus, the development of ensemble forecasting techniques for data-driven models is still in its early stages. Future research is still needed to explore ensemble forecasting in data-driven models and evaluate its effectiveness in predicting extreme events.

7 Abbreviations

NWP Numerical Weather Prediction

IFS Integrated Forecasting System

HRES high-resolution forecast

ENS ensemble forecast

ECMWF European Centre for Medium-Range Weather Forecasts

PNW Pacific Northwest

WCB Warm Conveyor Belt

WCBs Warm Conveyor Belts

ML Machine Learning

DL Deep Learning

GCM Global Circulation Model

GCMs Global Circulation Models

CNN Convolutional Neural Network

CNNs Convolutional Neural Networks

NN Neural Network

NWP Numerical Weather Prediction

S2S subseasonal-to-seasonal

GNN Graph Neural Network

GNNs Graph Neural Networks

ACC Anomaly Correlation Coefficient

ARs Atmopsheric Rivers

PBL Planetary Boundary Layer

TESSEL Tiled ECMWF Scheme for Surface Exchanges over Land

SPPT Stochastically Perturbed Parametrisation Tendencies

PDEs Partial Differential Equations

RMSE Root Mean square Error

Appendix

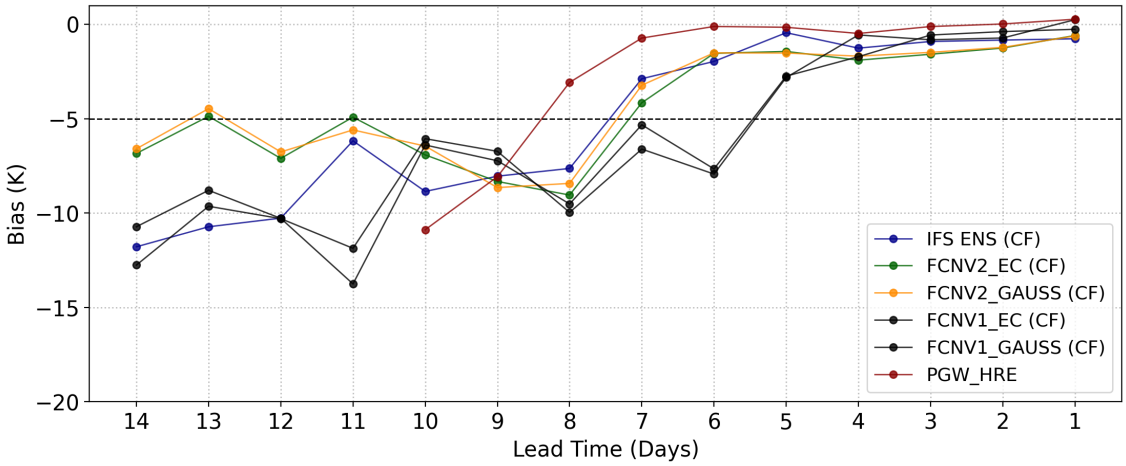


Figure A.1: Control forecast and deterministic forecast evolution of two-meter temperature bias with respect to ERA5 reanalysis valid on 29 June 00UTC, initialized from 14 days to 1 day prior to 29 June 00UTC. The solid marked line represents the control forecast and deterministic forecast (red line: operational Pangu-Weather); The two-meter temperature bias averaged over 20° latitude by 20° longitude box. The dashed line represents a 5 K bias baseline.

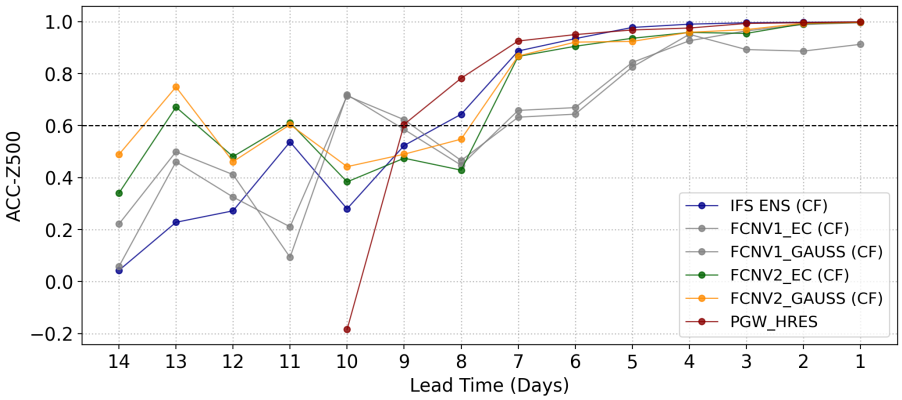


Figure A.2: Same as Figure A.1 but for ACC of 500 hPa geopotential height averaged over the region (145°W-95°W, 30°N-75°N).

(a) $FCNV2_{Good}$ initialized on 18 June (b) $FCNV2_{Bad}$ initialized on 18 June

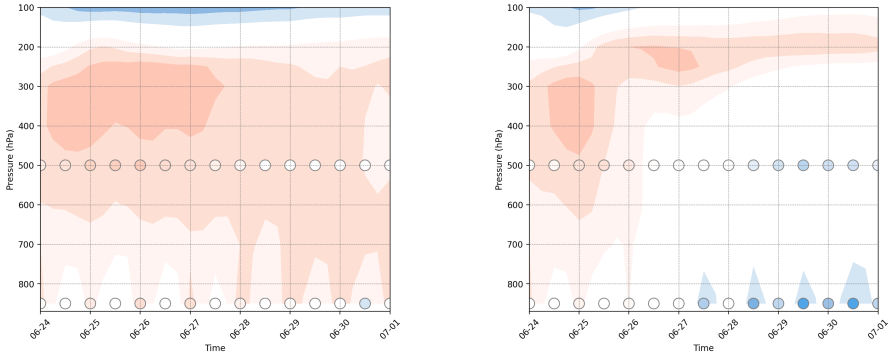


Figure A.3: Time-height plots of temperature forecast anomalies (24 June - 1 July), comparison of $FCNV2_{Good}$ and $FCNV2_{Bad}$ initialized on 18 June, scatter points represent FCNV1 forecast.

(a) $FCNV2_{Good}$ initialized on 18 June (b) $FCNV2_{Bad}$ initialized on 18 June

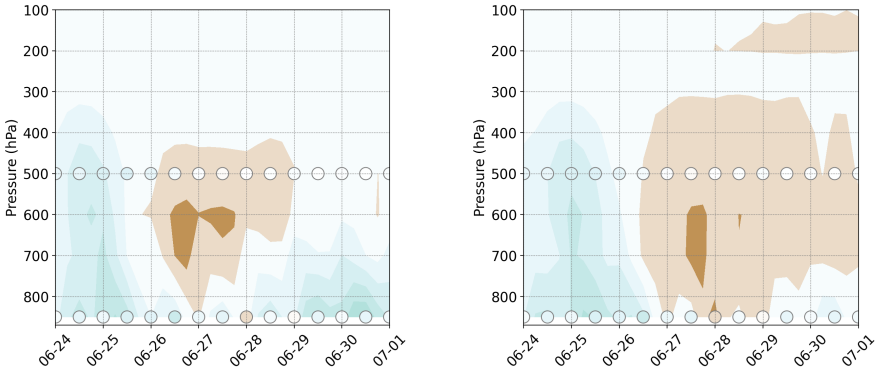


Figure A.4: Time-height plots of specific humidity forecast anomalies (24 June - 1 July), comparison of $FCNV2_{Good}$ and $FCNV2_{Bad}$ initialized on 18 June, scatter points represent FCNV1 forecast.

Bibliography

Ackmann, J., P. D. Düben, T. N. Palmer, and P. K. Smolarkiewicz, 2020: Machine-learned preconditioners for linear solvers in geophysical fluid flows. arXiv, https://doi.org/10.48550/ARXIV.2010.02866.

Al-Yahyai, S., Y. Charabi, and A. Gastli, 2010: Review of the use of numerical weather prediction (nwp) models for wind energy assessment. *Renewable and Sustainable Energy Reviews*, **14 (9)**, 3192–3198, https://doi.org/10.1016/j.rser.2010.07.001.

Alexander, L. V., and Coauthors, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, **111 (D5)**, https://doi.org/10.1029/2005jd006290.

Andersson, E., 2015: *User Guide to ECMWF Forecast Products*. URL /mnt/data/ecmwf_user_guide_update_v1.2_20151123.pdf, livelink 4320059, Page 83 of 129.

Baier, K., M. Rubel, and A. Stohl, 2023: The 3-week-long transport history and deep tropical origin of the 2021 extreme heat wave in the pacific northwest. *Geophysical Research Letters*, **50 (24)**, https://doi.org/10.1029/2023gl105865.

Balsamo, G., A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, and A. K. Betts, 2009: A revised hydrology for the ecmwf model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, **10 (3)**, 623–643, https://doi.org/10.1175/2008jhm1068.1.

Barriopedro, D., R. García-Herrera, C. Ordóñez, D. G. Miralles, and S. Salcedo-Sanz, 2023: Heat waves: Physical understanding and scientific challenges. *Reviews of Geophysics*, **61 (2)**, https://doi.org/10.1029/2022rg000780.

Bartusek, S., K. Kornhuber, and M. Ting, 2022: 2021 north american heatwave amplified by climate change-driven nonlinear interactions. *Nature Climate Change*, **12 (12)**, 1143–1150, https://doi.org/10.1038/s41558-022-01520-4.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55, https://doi.org/10.1038/nature14956.

Ben Bouallègue, Z., 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. *Bulletin of the American Meteorological Society*, **105 (6)**, E864–E883, https://doi.org/10.1175/

bams-d-23-0162.1.

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 (7970)**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bjerknes, V., 1904: Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik. *Meteorologische Zeitschrift*, **21**, 1–7, (The problem of weather prediction, considered from the viewpoints of mechanics and physics). Translated and edited by E. Volken and S. Brönnimann. – Meteorologische Zeitschrift 18 (2009), 663–667.

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical fourier neural operators: Learning stable dynamics on the sphere. arXiv, https://doi.org/10.48550/ARXIV.2306.03838.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, **141 (693)**, 3366–3382, https://doi.org/10.1002/qj.2619.

Bülte, C., N. Horat, J. Quinting, and S. Lerch, 2024: Uncertainty quantification for data-driven weather models. arXiv, https://doi.org/10.48550/ARXIV.2403.13458.

Chantry, M., S. Hatfield, P. Dueben, I. Polichtchouk, and T. Palmer, 2021: Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, **13 (7)**, https://doi.org/10.1029/2021ms002477.

Charlton-Perez, A. J., and Coauthors, 2024: Do ai models produce better weather forecasts than physics-based models? a quantitative evaluation case study of storm ciarán. *npj Climate and Atmospheric Science*, **7 (1)**, https://doi.org/10.1038/s41612-024-00638-w.

Chen, K., and Coauthors, 2023a: Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv, preprint, https://doi.org/https://arxiv.org/abs/2304.02948.

Chen, L., F. Du, Y. Hu, F. Wang, and Z. Wang, 2023b: Swinrdm: Integrate swinrnn with diffusion model towards high-resolution and high-quality weather forecasting. arXiv, preprint, https://doi.org/https://arxiv.org/abs/2306.03110.

Chen, L., X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, 2023c: Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. arXiv, preprint, https://doi.org/https://arxiv.org/abs/2306.12873.

Dacre, H. F., O. Martínez-Alvarado, and C. O. Mbengue, 2019: Linking atmospheric rivers and warm conveyor belt airflows. *Journal of Hydrometeorology*, **20 (6)**, 1183–1196, https://doi.org/10.1175/jhm-d-18-0175.1.

de Burgh-Day, C. O., and T. Leeuwenburg, 2023: Machine learning for numerical weather and climate modelling: a review. *EGUsphere*, **2023**, 1–48, https://doi.org/10.5194/egusphere-2023-350, URL https://egusphere.copernicus.org/preprints/2023/egusphere-2023-350/.

Domeisen, D. I. V., and Coauthors, 2022a: Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. *Bulletin of the American Meteorological Society*, **103 (6)**, E1473–E1501, https://doi.org/10.1175/bams-d-20-0221.1.

Domeisen, D. I. V., and Coauthors, 2022b: Prediction and projection of heatwaves. *Nature Reviews Earth amp; Environment*, **4 (1)**, 36–50, https://doi.org/10.1038/s43017-022-00371-z.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, **11 (10)**, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for ai weather prediction. *Nature*, **619 (7970)**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

Ebert-Uphoff, I., and K. Hilburn, 2024: The outlook for AI weather prediction. *Nature*, **619**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

ECMWF, 2023: Fact sheet: Reanalysis. ECMWF, accessed: 07.06.2024, https://www.ecmwf.int/en/about/media-centre/focus/2023/fact-sheet-reanalysis.

Emerton, R., C. Brimicombe, L. Magnusson, C. Roberts, C. Di Napoli, H. L. Cloke, and F. Pappenberger, 2022: Predicting the unprecedented: forecasting the june 2021 pacific northwest heatwave. *Weather*, **77 (8)**, 272–279, https://doi.org/10.1002/wea.4257.

Environment and Climate Change Canada, 2022: Canada's top 10 weather stories of 2021. Environment and Climate Change Canada, https://www.canada.ca/en/environment-climate-change/services/top-ten-weather-stories/2021.html.

Falk, T., and Coauthors, 2018: U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, **16 (1)**, 67–70, https://doi.org/10.1038/s41592-018-0261-2.

Fischer, E. M., S. I. Seneviratne, D. Lüthi, and C. Schär, 2007: Contribution of land-atmosphere coupling to recent european summer heat waves. *Geophysical Research Letters*, **34 (6)**, https://doi.org/10.1029/2006gl029068.

Fragkoulidis, G., V. Wirth, P. Bossmann, and A. H. Fink, 2018: Linking northern hemisphere temperature extremes to rossby wave packets. *Quarterly Journal of the Royal Meteorological Society*, **144 (711)**, 553–566, https://doi.org/10.1002/qj.3228.

Gettelman, A., D. J. Gagne, C. Chen, M. W. Christensen, Z. J. Lebo, H. Morrison, and G. Gantos, 2021: Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, https://doi.org/10.1029/2020ms002268.

Green, J., F. Ludlam, and J. McIlveen, 1966: Isentropic relative-flow analysis and the parcel theory. *Quarterly Journal of the Royal Meteorological Society*, **92 (392)**, 210–219.

Grotjahn, R., and Coauthors, 2015: North american extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends.

Climate Dynamics, **46 (3–4)**, 1151–1184, https://doi.org/10.1007/s00382-015-2638-6.

Harris, L., A. T. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, 2022: A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, **14 (10)**, https://doi.org/10.1029/2022ms003120.

Henderson, S. B., K. E. McLean, M. J. Lee, and T. Kosatsky, 2022: Analysis of community deaths during the catastrophic 2021 heat dome: Early evidence to inform the public health response during subsequent events in greater vancouver, canada. *Environmental Epidemiology*, **6 (1)**, e189, https://doi.org/10.1097/ee9.0000000000000189.

Hersbach, H., 2023: ERA5 reanalysis now available from 1940. *ECMWF Newsletter*, **175**.

Hotz, B., L. Papritz, and M. Röthlisberger, 2023: Understanding the vertical temperature structure of recent record-shattering heatwaves. *EGUsphere*, https://doi.org/10.5194/egusphere-2023-1703.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly Weather Review*, **133 (3)**, 604–620, https://doi.org/10.1175/mwr-2864.1.

IPCC, 2021: *Summary for Policymakers*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1-32 pp.

Japan Meteorological Agency, 2024: Outline of the operational numerical weather prediction at the japan meteorological agency. Japan Meteorological Agency, accessed: 2024-05-25, https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2024-nwp/pdf/outline2024_Appendix_A.pdf.

Kautz, L.-A., O. Martius, S. Pfahl, J. G. Pinto, A. M. Ramos, P. M. Sousa, and T. Woollings, 2021: Atmospheric blocking and weather extremes over the euro-atlantic sector – a review. *Weather and Climate Dynamics*, https://doi.org/10.5194/wcd-2021-56, [preprint].

Keisler, R., 2022: Forecasting global weather with graph neural networks. arXiv, https://doi.org/10.48550/ARXIV.2202.07575.

Lam, R., and Coauthors, 2022: Graphcast: Learning skillful medium-range global weather forecasting. arXiv, https://doi.org/10.48550/ARXIV.2212.12794.

Lang, S., and Coauthors, 2024: Aifs - ecmwf's data-driven forecasting system. arXiv, preprint, https://doi.org/10.48550/ARXIV.2406.01465.

Leith, C. E., 1974: Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, **102 (6)**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lessig, C., I. Luise, B. Gong, M. Langguth, S. Stadler, and M. Schultz, 2023: Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning. arXiv, preprint, https://doi.org/https://arxiv.org/abs/2308.13280.

Leutbecher, M., and T. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227 (7)**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Leutbecher, M., and Coauthors, 2017: Stochastic representations of model uncertainties at ecmwf: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, **143 (707)**, 2315–2339, https://doi.org/10.1002/qj.3094.

Lin, H., R. Mo, and F. Vitart, 2022: The 2021 western north american heatwave and its subseasonal predictions. *Geophysical Research Letters*, **49 (6)**, https://doi.org/10.1029/2021gl097036.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, **20 (2)**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Madonna, E., H. Wernli, H. Joos, and O. Martius, 2014: Warm conveyor belts in the era-interim dataset (1979–2010). part i: Climatology and potential vorticity evolution. *Journal of Climate*, **27 (1)**, 3–26, https://doi.org/10.1175/jcli-d-12-00720.1.

McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher, 2024: Identifying and categorizing bias in ai/ml for earth sciences. *Bulletin of the American Meteorological Society*, **105 (3)**, E567–E583, https://doi.org/10.1175/bams-d-23-0196.1.

Miralles, D. G., A. J. Teuling, C. C. van Heerwaarden, and J. Vilà-Guerau de Arellano, 2014: Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, **7 (5)**, 345–349, https://doi.org/10.1038/ngeo2141.

Mo, R., H. Lin, and F. Vitart, 2022: An anomalous warm-season trans-pacific atmospheric river linked to the 2021 western north america heatwave. *Communications Earth amp; Environment*, **3 (1)**, https://doi.org/10.1038/s43247-022-00459-w.

Neal, E., C. S. Y. Huang, and N. Nakamura, 2022: The 2021 pacific northwest heat wave and associated blocking: Meteorology and the role of an upstream cyclone as a diabatic source of wave activity. *Geophysical Research Letters*, **49 (8)**, https://doi.org/10.1029/2021gl097699.

Nguyen, T., J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, 2023: Climax: A foundation model for weather and climate. arXiv, preprint, https://doi.org/https://arxiv.org/abs/2301.10343.

Oertel, A., and Coauthors, 2023: Everything hits at once: How remote rainfall matters for the prediction of the 2021 north american heat wave. *Geophysical Research Letters*, **50 (3)**, https://doi.org/10.1029/2022gl100958.

Olivetti, L., and G. Messori, 2023: Advances and prospects of deep learning for medium-range extreme weather forecasting. *EGUsphere*, https://doi.org/10.5194/egusphere-2023-2490, [preprint].

Olivetti, L., and G. Messori, 2024: Do data-driven models beat numerical models in forecasting weather extremes? a comparison of ifs hres, pangu-weather and graphcast. *EGUsphere*, https://doi.org/10.5194/egusphere-2024-1042, [preprint].

Owens, R. G., and T. D. Hewson, 2018: *ECMWF Forecast User Guide*. ECMWF, Reading, https://doi.org/10.21957/m1cs7hc, section 2; Section 5; Section 6.

Palmer, T., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. Tech. rep., ECMWF, 2 pp. https://doi.org/10.21957/PS8GBWBDV.

Palmer, T., G. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, **33 (1)**, 163–193, https://doi.org/10.1146/annurev.earth.33.092203.122552.

Papritz, L., and M. Röthlisberger, 2023: A novel temperature anomaly source diagnostic: Method and application to the 2021 heatwave in the pacific northwest. *Geophysical Research Letters*, **50 (23)**, https://doi.org/10.1029/2023gl105641.

Pasche, O. C., J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, 2024: Validating deep-learning weather forecast models on recent high-impact extreme events. https://doi.org/10.48550/ARXIV.2404.17652, preprint, 2404.17652.

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

Pelly, J. L., and B. J. Hoskins, 2003: A new perspective on blocking. *Journal of the Atmospheric Sciences*, **60 (5)**, 743 – 755, https://doi.org/https://doi.org/10.1175/1520-0469(2003)060%3C0743:ANPOB%3E2.0.CO;2.

Perkins, S. E., and L. V. Alexander, 2013: On the measurement of heat waves. *Journal of Climate*, **26 (13)**, 4500–4517, https://doi.org/10.1175/jcli-d-12-00383.1.

Perkins-Kirkpatrick, S. E., and P. B. Gibson, 2017: Changes in regional heatwave characteristics as a function of increasing global temperature. *Scientific Reports*, **7 (1)**, https://doi.org/10.1038/s41598-017-12520-2.

Pfahl, S., C. Schwierz, M. Croci-Maspoli, C. M. Grams, and H. Wernli, 2015: Importance of latent heat release in ascending air streams for atmospheric blocking. *Nature Geoscience*, **8 (8)**, 610–614, https://doi.org/10.1038/ngeo2487.

Pfahl, S., and H. Wernli, 2012: Quantifying the relevance of atmospheric blocking for co-located temperature extremes in the northern hemisphere on (sub-)daily time scales. *Geophysical Research Letters*, **39 (12)**, https://doi.org/10.1029/2012gl052261.

Philip, S. Y., and Coauthors, 2022: Rapid attribution analysis of the extraordinary heat wave on the pacific coast of the us and canada in june 2021. *Earth System Dynamics*, **13 (4)**, 1689–1713, https://doi.org/10.5194/esd-13-1689-2022.

Price, I., and Coauthors, 2023: Gencast: Diffusion-based ensemble forecasting for medium-range weather. arXiv, preprint, https://doi.org/10.48550/ARXIV.2312.15796.

Pu, Z., and E. Kalnay, 2018: Numerical weather prediction basics: Models, numerical methods, and data assimilation. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. Cloke, and J. Schaake, Eds., Springer, Berlin, Heidelberg, 2–3.

Ralph, F. M., M. D. Dettinger, J. J. Rutz, and D. E. Waliser, Eds., 2020: *Atmospheric Rivers*. Springer International Publishing, 252 pp.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, **115 (39)**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Rasp, S., and N. Thuerey, 2021: Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, **13 (2)**, https://doi.org/10.1029/2020ms002405.

Rasp, S., and Coauthors, 2023: Weatherbench 2: A benchmark for the next generation of data-driven global weather models. arXiv, preprint, https://doi.org/10.48550/ARXIV.2308.15560.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven earth system science. *Nature*, **566 (7743)**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Richardson, D. S., H. L. Cloke, J. A. Methven, and F. Pappenberger, 2024: Jumpiness in ensemble forecasts of atlantic tropical cyclone tracks. *Weather and Forecasting*, **39 (1)**, 203–215, https://doi.org/10.1175/waf-d-23-0113.1.

Russo, S., J. Sillmann, and E. M. Fischer, 2015: Top ten european heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, **10 (12)**, 124 003, https://doi.org/10.1088/1748-9326/10/12/124003.

Scher, S., 2018: Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, **45 (22)**, https://doi.org/10.1029/2018gl080704.

Scher, S., and G. Messori, 2019: Weather and climate forecasting with neural networks: using general circulation models (gcms) with different complexity as a study ground. *Geoscientific Model Development*, **12 (7)**, 2797–2809, https://doi.org/10.5194/gmd-12-2797-2019.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **379 (2194)**, 20200 097, https://doi.org/10.1098/rsta.2020.0097.

Schumacher, D. L., M. Hauser, and S. I. Seneviratne, 2022: Drivers and mechanisms of the 2021 pacific northwest heatwave. *Earth's Future*, **10 (12)**, https://doi.org/10.1029/2022ef002967.

Schwierz, C., M. Croci-Maspoli, and H. C. Davies, 2004: Perspicacious indicators of atmospheric blocking. *Geophysical Research Letters*, **31 (6)**, https://doi.org/10.1029/2003gl019341.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, **99 (3–4)**, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004.

Tamarin-Brodsky, T., and N. Harnik, 2024: The relation between rossby wave-breaking events and low-level weather systems. *Weather and Climate Dynamics*, **5 (1)**, 87–108, https://doi.org/10.5194/wcd-5-87-2024.

Thompson, V., and Coauthors, 2022: The 2021 western north america heat wave among the most extreme events ever recorded globally. *Science Advances*, **8 (18)**, https://doi.org/10.1126/sciadv.abm6860.

Tian, D., E. F. Wood, and X. Yuan, 2017: Cfsv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous united states. *Hydrology and Earth System Sciences*, **21 (3)**, 1477–1490, https://doi.org/10.5194/hess-21-1477-2017.

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. *Tellus A: Dynamic Meteorology and Oceanography*, **42 (3)**, 343, https://doi.org/10.3402/tellusa.v42i3.11882.

Trigo, R. M., I. F. Trigo, C. C. DaCamara, and T. J. Osborn, 2004: Climate impact of the european winter blocking episodes from the ncep/ncar reanalyses. *Climate Dynamics*, **23 (1)**, 17–28, https://doi.org/10.1007/s00382-004-0410-4.

Wernli, H., and H. C. Davies, 1997: A lagrangian-based analysis of extratropical cyclones. i: The method and some applications. *Quarterly Journal of the Royal Meteorological Society*, **123 (538)**, 467–489, https://doi.org/10.1002/qj.49712353811.

Weyn, J. A., D. R. Durran, and R. Caruana, 2019: Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, **11 (8)**, 2680–2693, https://doi.org/10.1029/2019ms001705.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, **12 (9)**, https://doi.org/10.1029/2020ms002109.

White, R. H., and Coauthors, 2023: The unprecedented pacific northwest heatwave of june 2021. *Nature Communications*, **14 (1)**, https://doi.org/10.1038/s41467-023-36289-3.

Wilks, D. S., 2011: Chapter 7 - statistical forecasting. *International Geophysics*, D. S. Wilks, Ed., International Geophysics, Vol. 100, Academic Press, 215–300, https://doi.org/10.1016/B978-0-12-385022-5.00007-5.

Zhang, Y., and W. R. Boos, 2023: An upper bound for extreme temperatures over midlatitude land. *Proceedings of the National Academy of Sciences*, **120 (12)**, https://doi.org/10.1073/pnas.2215278120.

Zschenderlein, P., S. Pfahl, H. Wernli, and A. H. Fink, 2020: A lagrangian analysis of upper-tropospheric anticyclones associated with heat waves in europe. *Weather and Climate Dynamics*, **1 (1)**, 191–206, https://doi.org/10.5194/wcd-1-191-2020.

Acknowledgments

First, I would like to thank Prof. Dr. Peter Knippertz and Dr. Julian Quinting for allowing me to explore this exciting topic. I am especially thankful for their invaluable feedback and guidance throughout the year, allowing me to learn from new perspectives continually. Peter, thank you for giving me the opportunity to join in the exciting workshop and for always helping me navigate from confusion to clarity in every discussion.

I am incredibly grateful to Dr. Jannik Wilhelm and M.Sc. Nina Horat. I could not think of better mentors than them. Throughout the entire journey, they provided me with careful guidance and encouragement to explore. Jannik, thank you for all the help and inspiration you have given me. Whether it is technical support, academic discussions, or writing suggestions, you are always there for me. Nina, I have always learned so much from you. You can always see the key points of things and directly give me the most effective help. My gratitude is beyond my words.

I would also like to thank Dr. Alexander Lemburg and Dr. Annika Oertel for bringing fresh perspectives during the initial conception of the thesis. Their academic enthusiasm always touches me.

Lastly, I would like to thank my friends and family for their endless support, love, and joy.

Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 15.06.2024

Yangfan Zhou