Karlsruhe Institute of Technology

Online bias correction in data-driven weather forecast models

Master's Thesis of

Nicole Knopf

At the KIT Department of Physics
Institute of Meteorology and Climate Research (IMKTRO)

First examiner: Dr. Julian Quinting
Second examiner: Prof. Dr. Peter Knippertz

01. June 2024 – 23. June 2025

Karlsruher Institut für Technologie
Fakultät für Physik
76131 Karlsruhe

Online bias correction in data-driven weather forecast models (Master's Thesis)

I declare that I have developed and written the enclosed thesis completely by myself. I have not used any other than the aids that I have mentioned. I have marked all parts of the thesis that I have included from referenced literature, either in their original wording or paraphrasing their contents. I have followed the by-laws to implement scientific integrity at KIT.

Karlsruhe, 13. June 2025

. . . . . .

(Nicole Knopf)

Abstract

In recent years, Machine Learning (ML)-based Weather Prediction (MLWP) has emerged as a promising alternative to traditional Numerical Weather Prediction (NWP) systems. These data-driven models, such as the transformer-based Pangu-Weather, offer substantial computational advantages and have demonstrated competitive skill in medium-range forecasts. However, despite their impressive performance, MLWP models are prone to systematic forecast biases that increase with lead time and vary seasonally and regionally. Such biases, particularly in temperature forecasts at the 850-hPa level, present a major challenge for the operational applicability of MLWP, especially for longer lead times and in subseasonal contexts.

To address this issue, this thesis explores online bias correction strategies that iteratively adjust forecasts at each prediction step, thereby mitigating the propagation of errors throughout the forecast chain. The correction techniques are implemented within the standardized WeatherBench2 (WB2) framework and applied to Pangu-Weather forecasts using ERA5 reanalysis data as a reference. Two statistical methods are investigated: Multiple linear regression (MLR) and Extreme Gradient Boosting (XGBoost). Both methods are trained to correct the systematic error of temperature forecasts at each grid point on a coarsened spatial resolution and subsequently re-interpolated to high resolution.

Results show that offline bias correction with XGBoost leads to a substantial reduction in systematic errors, both globally and regionally. In particular, the method effectively corrects pronounced cold and warm biases in critical regions such as the Southeast Pacific and the Caribbean. XGBoost consistently outperforms MLR due to its ability to capture nonlinear relationships and regional variability. In contrast, the online correction shows more limited improvements under the current implementation. Nonetheless, it offers conceptual advantages in autoregressive forecast settings by addressing error accumulation during forecast generation. The findings highlight the practical value of offline correction while also emphasizing the potential of online methods. Future work should further explore and refine online correction strategies to better exploit their integration into the iterative nature of MLWP systems.

Zusammenfassung

In den letzten Jahren hat sich die datengetriebene Wettervorhersage mittels maschinellen Lernens (MLWP) als vielversprechende Alternative zu traditionellen numerischen Wettervorhersagemodellen (NWP) etabliert. Modelle wie das auf Transformern basierende Pangu-Weather bieten erhebliche Vorteile in Bezug auf den Rechenaufwand und zeigen eine mit klassischen Verfahren vergleichbare Vorhersagegüte im mittelfristigen Bereich. Trotz dieser Fortschritte weisen MLWP-Modelle systematische Vorhersageverzerrungen auf, die mit zunehmender Vorhersagezeit ansteigen und regional sowie saisonal variieren. Solche systematischen Fehler, insbesondere bei Temperaturvorhersagen auf dem 850-hPa-Niveau, stellen eine wesentliche Herausforderung für den operationellen Einsatz dar, insbesondere bei längeren Vorhersagezeiträumen im subseasonalen Bereich.

Zur Lösung dieses Problems untersucht die vorliegende Arbeit Strategien zur Online-Bias-Korrektur, bei denen die Vorhersage nach jedem Zeitschritt angepasst wird, um die Fehlerfortpflanzung im Verlauf des Prognoseprozesses zu verringern. Die Korrekturverfahren werden im standardisierten WB2-Framework implementiert und auf Vorhersagen des Pangu-Weather-Modells angewendet, wobei ERA5-Reanalysedaten als Referenz dienen. Zwei statistische Methoden werden evaluiert: die multiple lineare Regression (MLR) und das Extreme Gradient Boosting (XGBoost). Beide Ansätze werden auf einem räumlich reduzierten Gitter trainiert und anschließend auf die ursprüngliche Auflösung rückinterpoliert.

Die Ergebnisse zeigen, dass eine Offline-Korrektur mit XGBoost systematische Fehler sowohl global als auch regional deutlich reduziert. Insbesondere lassen sich ausgeprägte Kalt- und Warmverzerrungen in Regionen wie dem südöstlichen Pazifik und der Karibik effektiv korrigieren. XGBoost übertrifft MLR dabei konsistent, da es nichtlineare Zusammenhänge und regionale Unterschiede besser abbilden kann. Die Online-Korrektur zeigt hingegen im aktuellen Setup nur begrenzte Verbesserungen. Dennoch bietet sie konzeptionelle Vorteile für autoregressive Vorhersagemodelle, da sie Fehler bereits während der Prognosebildung adressieren kann. Die Ergebnisse unterstreichen den praktischen Nutzen von Offline-Korrekturen und weisen zugleich auf das Potenzial von Online-Ansätzen hin, das in zukünftigen Arbeiten gezielt weiterentwickelt werden sollte – insbesondere im Hinblick auf eine engere Verzahnung mit der iterativen Struktur von MLWP-Modellen.

Contents

List of Figures

List of Tables

Abbreviations

3DEST	3D Earth-specific transformer
AI	Artificial Intelligence
C3S	Copernicus Climate Change Service
CNN	convolutional neural network
CRPS	Continous Ranked Probability Score
DJF	December-January-February
DNN	deep neural network
DOY	day of year
ECMWF	European Centre for Medium-Range Weather Forecasts
ENSO	El Niño–Southern Oscillation
GCM	general circulation model
HRES	High Resolution
IFS	Integrated Forecasting System
JJA	June-July-August
MAE	mean absolute error
MAM	March-April-May
MJO	Madden-Julian Oscillation
ML	Machine Learning
MLR	Multiple linear regression
MLWP	ML-based Weather Prediction
MSE	mean squared error
NH	Northern Hemisphere
NWP	Numerical Weather Prediction
RMSE	root mean squared error
RNN	recurrent neural network
RSS	residual sum of squares
Sc	stratocumulus
SH	Southern Hemisphere
SON	September-October-November
SST	Sea Surface Temperature
WB2	WeatherBench2
XGBoost	Extreme Gradient Boosting

1 Introduction

In recent years, ML models have rapidly gained relevance in weather forecasting (Schultz et al., 2021; Dueben and Bauer, 2018). Unlike traditional NWP systems that rely on the explicit integration of physical equations describing the atmosphere (Bauer et al., 2015), MLWP models learn statistical relationships directly from historical data (Reichstein et al., 2019). This paradigm shift has been driven by the growing availability of high-resolution reanalysis datasets, advances in hardware acceleration and breakthroughs in deep learning architectures, especially those designed for spatiotemporal data. As a result, MLWP has emerged as a viable alternative to physics-based models.

Forecasts produced by MLWP models are often generated at significantly reduced computational cost and, in some cases, demonstrate forecast skill comparable to or exceeding that of operational NWP systems (Pathak et al., 2022; Bi et al., 2023). One of the most prominent examples of this new model class is Pangu-Weather, developed by Huawei Cloud, which leverages Earth-specific transformer architectures and has been trained on the ERA5 reanalysis dataset (Bi et al., 2023). By learning from past atmospheric states, Pangu-Weather is capable of producing global forecasts at high temporal and spatial resolution, including multiple vertical pressure levels, with lead times extending into the medium range.

Pangu-Weather has demonstrated considerable skill in the medium range (up to 10 days), outperforming traditional NWP systems on several key variables (Bi et al., 2023). However, like many MLWP models, Pangu-Weather exhibits systematic forecast biases, in particular in the lower troposphere, whose magnitude increases with lead time and whose spatial pattern varies regionally and seasonally (Bouallègue et al., 2024). These biases undermine forecast reliability and pose a major obstacle for the broader application of MLWP models in operational forecasting (Dueben and Bauer, 2018; Schultz et al., 2021). While the original evaluation focused on the medium range, the model itself can generate forecasts for longer lead times through autoregressive iteration, making it suitable for exploratory subseasonal forecasting.

The challenge of correcting forecast biases becomes especially relevant when considering forecasts on the subseasonal timescale, which spans lead times from roughly 2 weeks to 2 months (Robertson and Vitart, 2019; Vitart et al., 2017). This forecasting regime is of increasing societal relevance. Decisions in agriculture, water management, energy supply and disaster risk reduction all benefit from reliable guidance on this timescale (White et al., 2017; Mariotti et al., 2020). However, the subseasonal range is also one of the most difficult to forecast, due to the so-called “predictability desert” (Vitart et al., 2017). In this range, predictive signals from initial conditions decay rapidly, while slowly evolving boundary forcings such as sea surface temperature anomalies are often too weak to dominate forecast

skill (Mariotti et al., 2020; Pegion et al., 2019). Improving forecast accuracy under these conditions remains a central research challenge (White et al., 2017).

MLWP models like Pangu-Weather offer new possibilities for addressing this challenge. Since they do not rely on numerical time integration, they can, in principle, produce long-range forecasts more efficiently and with fewer accumulated numerical errors. However, the autoregressive nature of these models, in particular their tendency to use previous predictions as input for future steps, introduces a critical vulnerability. Small biases introduced early in the forecast can propagate and amplify, leading to significant errors at longer lead times (Watt-Meyer et al., 2021; Hamill and Whitaker, 2006). This problem underscores the need for effective and targeted bias correction strategies.

Various statistical techniques have been developed to mitigate forecast bias. Offline correction methods apply post-processing after the full forecast is completed, based on historical errors or climatological statistics (Wilks, 2019; Glahn and Lowry, 1972). These methods are relatively simple and computationally inexpensive, but they do not prevent the propagation of bias through the forecast chain. In contrast, online bias correction techniques aim to adjust the model output after each prediction step and feed the corrected field into the next iteration (Hamill and Whitaker, 2006; Watt-Meyer et al., 2021). This dynamic approach allows the model to stay closer to the desired trajectory and can significantly reduce cumulative error—particularly in autoregressive settings such as those used in many MLWP models.

Despite its promise, online bias correction has so far received limited attention in the context of MLWP. While several studies have applied such techniques to dynamical or hydrological models (e.g., Hamill and Whitaker, 2006; Watt-Meyer et al., 2021), systematic investigations in the ML domain are still rare (Rasp et al., 2020; Schultz et al., 2021). Recently, Bouallègue et al. (2024) have provided one of the first statistical assessments of MLWP in an operational-like context. Their findings underscore both the potential and the limitations of current MLWP systems and highlight the importance of post-processing and error correction for practical applications. These insights provide the central motivation for this thesis.

Against this background, the present work aims to investigate online bias correction for Pangu-Weather forecasts of 850-hPa temperature using ERA5 reanalysis data (Hersbach et al., 2020) as reference. The study focuses on evaluating the effectiveness of statistical correction methods applied in an online fashion in Pangu-Weather and seeks to answer the following research questions:

1. What are the underlying causes of systematic biases in Pangu-Weather forecasts, and how do they relate to atmospheric dynamics and model design?
2. Which statistical approaches are most suitable for correcting these biases in an online setting, and how do they compare in terms of accuracy and robustness?
3. How does bias correction affect medium-range forecast performance, and what are the implications for operational usability at subseasonal lead times?

The methodology is implemented within the standardized WB2 framework (Rasp et al., 2024), which offers reproducible tools and datasets for benchmarking MLWP models. Chapter 2 provides the theoretical foundations for NWP and MLWP and outlines the unique challenges of subseasonal forecasting. Chapter 3 outlines the datasets and statistical methods used for offline and online correction and provides an overview of relevant literature on online bias correction approaches. Results are presented and analyzed in Chapter 4, followed by a discussion of key findings, methodological limitations and future directions in Chapter 5.

By combining machine learning forecasts with statistical post-processing and real-time correction techniques, this thesis contributes to the development of hybrid forecasting systems that harness the strengths of both data-driven and physically consistent modeling. In doing so, it seeks to improve the usability, reliability and robustness of MLWP forecasts, in particular in the challenging subseasonal range where traditional forecasting approaches often struggle (Bouallègue et al., 2024; Schultz et al., 2021; Mariotti et al., 2020).

2 Theoretical Background

This chapter outlines the theoretical foundations relevant to this study. Section 2.1 introduces the principles of NWP, focused on the role of physical models, data assimilation and current limitations in forecast accuracy. Section 2.2 discusses data-driven approaches to weather forecasting, with an emphasis on recent developments in machine learning and their applications to meteorological data. Section 2.3 briefly introduces the field of subseasonal forecasting and its significance for extending the predictive horizon beyond the medium range.

2.1 Numerical Weather Prediction

NWP is the foundation of weather forecasting. It is based on the mathematical representation of physical processes governing the atmosphere. This concept was first introduced by Bjerknes (1904), who proposed that weather forecasting could be approached as a deterministic problem by solving the fundamental equations of hydrodynamics and thermodynamics. His vision laid the theoretical foundation for modern NWP. The governing processes are described by a set of nonlinear partial differential equations, primarily derived from the laws of fluid dynamics and thermodynamics. The equations include the Navier–Stokes equations for fluid motion, the mass continuity equation, the first law of thermodynamics and the ideal gas law. Since these equations cannot be solved analytically due to their complexity and the chaotic nature of the atmosphere, they are discretized and solved numerically on a computational grid covering the globe or a limited region.

Forecasts are generated by integrating the discretized equations forward in time from the analysis state. However, due to the inherent chaotic nature of the atmosphere, even small inaccuracies in the initial conditions can rapidly lead to growing forecast errors. This sensitivity to initial conditions is often referred to as the "butterfly effect" and is a well-known property of nonlinear dynamical systems such as the atmosphere (Lorenz, 1963). This suggests that minor perturbations, such as those arising from observational gaps, instrument noise or model assumptions, can intensify over time and significantly alter the predicted atmospheric evolution.

As a result, the accuracy of NWP models is highly dependent on the quality of the initial conditions used for the forecast. Therefore, precise estimation of the current atmospheric state is critical. However, atmospheric observations are incomplete and unevenly distributed in space and time. Data assimilation techniques are applied to optimally estimate the current state of the atmosphere. The process of data assimilation combines observations from various

sources, such as surface stations, weather balloons, aircraft and satellites, with a prior model forecast, known as the background. The result is a physically consistent estimate of the atmospheric state, called the analysis.

The spatial and temporal resolutions of NWP models are constrained by available computational power. Global models typically operate at resolutions ranging from 10 to 25 kilometers, while regional or limited-area models can resolve finer scales, often below 5 kilometers. Higher resolutions allow a better representation of small-scale processes such as convection, orographic effects and local circulations. However, many subgrid-scale processes, including cloud microphysics, radiation and turbulence, cannot be explicitly resolved and must therefore be parameterized. These parameterizations introduce approximations and are a significant source of model uncertainty.

Ensemble forecasting is widely used to account for the model's uncertainty. In this approach, multiple forecasts are run simultaneously with slightly different initial conditions or model physics. The resulting ensemble provides a probabilistic view of future weather states and is valuable for estimating forecast confidence and identifying potential extreme events.

Recent advances in computational resources, observational networks and data assimilation methods have led to significant improvements in forecast skill over the past decades. Forecasts up to about 7–10 days show now a high degree of reliability and models are increasingly able to predict extreme events such as storms and heatwaves. However, despite these continuing advancements in NWP, the accuracy of forecasts is still limited by several sources of uncertainty. These include imperfect representations of physical processes, approximations introduced through parameterizations and inaccuracies in the initial conditions. Even with significant improvements in observational networks and model resolutions, these challenges persist. Because the atmosphere behaves as a chaotic system, small errors in the initial state can grow rapidly and lead to large deviations in forecast outcomes. As a result, a single deterministic forecast is inherently limited in its capacity to provide reliable quantitative information about future atmospheric conditions.

To better address this uncertainty, numerical weather prediction has increasingly shifted toward probabilistic methods, as noted by Buizza and Leutbecher (2015). In particular, ensemble forecasting has become a central tool, providing a range of possible outcomes based on slightly varied initial conditions or model configurations. This approach has been essential in improving the reliability of weather forecasts and identifying potential extreme events over the past few decades.

While exact runtimes depend on the specific model configuration and computing architecture, it is common for operational global NWP forecasts to take several hours to compute forecasts up to several days ahead.

2.2 Data-driven weather models

Data-driven weather forecasting marks a significant shift in atmospheric science. It has emerged in a response to the growing availability of high-resolution datasets and the limitations of traditional NWP systems. Rather than solving the governing physical equations explicitly, data-driven approaches learn statistical or functional relationships between past and future atmospheric states from large historical datasets. These methods encompass a wide range of techniques, including classical statistical models and analog methods to modern machine learning and deep learning algorithms. The fundamental concept is to leverage large volumes of historical observations and reanalysis or model forecast data to train systems that can predict future atmospheric conditions.

In recent years, the combination of increasing amounts of available data and improved computational power has led to rapid progress in the field of data-driven weather forecasting. A major development are deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformers. These models are particularly effective at capturing complex spatial and temporal patterns in atmospheric data (Rasp et al., 2020; Schultz et al., 2021). They can be trained on previous observations, reanalysis data or NWP outputs to predict future weather variables at various time scales.

A key advantage of data-driven methods are their potential to complement or, in some cases, even outperform physics-based models, especially in scenarios where NWP systems face limitations such as coarse resolution, model biases or poorly parameterized processes. Machine learning techniques have proven strong performance in post-processing tasks, such as bias correction, downscaling and the generation of probabilistic forecasts (Vannitsem et al., 2021). Gradient boosting methods and neural networks have been successfully applied to refine the output of NWP models, reducing systematic errors and improving local forecast accuracy. Furthermore, analog-based techniques and ensemble learning approaches have demonstrated the ability to exploit patterns in historical weather situations to improve medium- to long-range forecasts.

Despite their promise, data-driven forecasts also face significant challenges. The atmosphere is a high-dimensional, nonlinear and chaotic system. Therefore, the generalization beyond the conditions represented in the training data is difficult for purely data-driven models. This limitation becomes especially relevant in extreme or previous unobserved situations. Moreover, many machine learning models require large amounts of labeled data and computational resources for training and they often lack interpretability compared to physics-based models, which are grounded in known laws of nature. To address these limitations, several strategies are being explored. These include physics-informed machine learning, hybrid models that combine physical NWP components with neural networks and regularization strategies that enforce physical consistency during training (Reichstein et al., 2019; Beucler et al., 2021).

Another area of active development is the use of data-driven approaches for long-range forecasts, in particular on the subseasonal timescale. Traditional NWP systems often struggle on this timescale due to the limited predictability of chaotic atmospheric dynamics.

Data-driven approaches can improve skill in the challenging subseasonal timescale. This is achieved by extracting low-frequency patterns from large climate datasets related to phenomena such as El Niño–Southern Oscillation (ENSO), the Madden-Julian Oscillation (MJO), or stratospheric variability. The forecast skill can further be improved when such approaches are combined with dynamical forecasts or employed as post-processing tools (Chattopadhyay et al., 2020; Díaz et al., 2023).

Recent breakthroughs highlight the potential of deep learning for global weather forecasting. For instance, Pangu-Weather (Bi et al., 2023), a transformer-based model, has demonstrated medium-range forecast skill comparable or exceeding operational NWP systems. Similarly, GraphCast (Lam et al., 2023), based on graph neural networks, has successfully modeled global atmospheric dynamics. These developments illustrate the growing capability of machine learning to enhance or partially replace certain aspects of traditional NWP systems.

2.3 Subseasonal forecasts

Subseasonal forecasts, typically defined as predictions with lead times ranging from two weeks to two months, occupy a critical but challenging timescale between medium-range weather prediction and seasonal climate forecasting. This range extends beyond the deterministic predictability limit of the atmosphere (approximately 10–14 days), yet is often too short for the slow-varying boundary forcings, such as sea surface temperatures, to provide strong predictive signals. Despite these challenges, the subseasonal scale is of immense societal value, supporting applications in agriculture, hydrology, energy management and disaster preparedness, in particular for anticipating the onset of heatwaves, cold spells, droughts or extended periods of heavy precipitation.

One of the central challenges in subseasonal forecasting arises from the limited sources of predictability on these timescales. The chaotic nature of the atmosphere quickly diminishes the skill of initial-condition-based forecasts, which are the backbone of NWP. Nevertheless, certain components of the climate system evolve more slowly and can serve as potential predictors. These include the MJO, quasi-biennial oscillation, stratosphere–troposphere interactions, soil moisture anomalies, sea ice extent and sea surface temperature patterns, including those related to the ENSO. The ability of forecast systems to detect and leverage these slowly varying signals affects the forecast skill on the subseasonal scale (Vitart, 2014; Mariotti et al., 2020).

Current operational subseasonal forecasts are typically generated by global coupled NWP systems, such as those maintained by European Centre for Medium-Range Weather Forecasts (ECMWF) or NCEP and are often they are issued in ensemble mode to capture inherent uncertainties. These systems attempt to simulate the fast-evolving synoptic weather and the slower climate drivers within a unified framework. Despite advances in model resolution and physical parameterizations, subseasonal forecast skill remains modest and highly dependent on region, variable and season. For example, forecasts of temperature tend to exhibit higher

skill than precipitation and periods of strong MJO activity often correspond with improved predictive capacity (Robertson and Vitart, 2019).

Recent years have seen the emergence of hybrid approaches that combine dynamical models with statistical or machine learning techniques to enhance forecast accuracy. These include post-processing methods for bias correction and downscaling, as well as more integrated strategies that fuse model output with data-driven components. Machine learning models have been used to predict the phase of the MJO to identify analogues of past weather patterns and to extract low-frequency signals from noisy data (Kim et al., 2021; Chattopadhyay et al., 2020). Some deep learning models such as Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023) where originally developed for medium-range forecasting. Nevertheless, they have shown potential to be extended toward longer lead times due to their ability to rapidly generate ensemble forecasts and capture global spatial dependencies.

The Subseasonal to Seasonal (S2S) prediction project, an initiative from the World Meteorological Organization (WMO), ECMWF and other partners, has played a crucial role in fostering research on this timescale by providing an open-access database of reforecasts and operational forecasts from multiple modeling centers (Vitart et al., 2017). This database has enabled systematic evaluation and intercomparison of forecast skill across different models, variables and lead times, revealing large potential for improvement. Especially through ensemble calibration, the use of multi-model ensembles and the integration of alternative prediction systems.

3 Data and Methods

This chapter provides an overview of the data sources and methodologies used in this thesis. Section 3.1 introduces the main datasets: the WeatherBench2 benchmark framework, the ERA5 reanalysis and forecasts from the deep learning model Pangu-Weather. Section 3.2 describes the online bias correction approach, which is used to improve the accuracy of raw forecasts. Finally, Section 3.3 presents the regression methods applied for bias correction, including multiple linear regression and the machine learning algorithm XGBoost.

3.1 Data

3.1.1 Weatherbench2

WB2 (Rasp et al., 2024) is a benchmark and evaluation framework designed to facilitate the standardized comparison of Artificial Intelligence (AI) models with traditional NWP systems. In general, a benchmark serves as a standardized reference for measuring progress in a specific task or comparing different approaches. In the case of WB2, it provides standardized datasets and evaluation metrics to enable fair and reproducible evaluation of data-driven weather models, helping to assess their strengths and limitations across different regions and forecast horizons.

The rapidly growing number of MLWPs forecasting models in recent years has made it increasingly difficult to evaluate their performance in a consistent and reproducible manner. Differences in training data, evaluation metrics, implementation, and experimental setups have often led to results that are not directly comparable. WB2 addresses this challenge by providing a unified framework that ensures fairness, transparency, and reproducibility across studies. The framework consists of several key components: It offers ground truth, like ERA5 and a limited number of station observations, and baseline datasets. Those are publicly available and stored in an easily accessible format in a public Google Cloud bucket, ensuring researchers to have access to high-quality data to train their models. Furthermore, WB2 offers an open-source evaluation code to compare different AI models against ground truth and obtain commonly used forecast skill scores. Additionally, WB2 is supplemented by a website, which shows the current headline scorecards. Those scorecards compare the skill, measured by the root mean squared error (RMSE), of different numerical and data-driven weather models in relation to ECMWF's Integrated Forecasting System (IFS) High Resolution (HRES) on eight different variables.

WB2 builds upon the foundation of its predecessor, WeatherBench1 (Rasp et al., 2020), introducing higher spatial and temporal resolution, a broader set of variables and an increased emphasis on transparency and reproducibility. A central difference between the two benchmarks lies in the performance range of the models considered. While WeatherBench1 primarily included models whose forecast skill remained significantly below that of state-of-the-art operational systems such as those of the ECMWF, WB2 is explicitly designed to evaluate models that reach, and in some cases even exceed, the skill of leading NWP systems. This shift reflects the rapid progress in data-driven weather forecasting and marks a transition from evaluating early-stage machine learning approaches to benchmarking models that are competitive with operational standards.

To ensure objective comparisons, WB2 offers open-source evaluation code that computes standardized metrics across multiple variables and lead times. The most commonly used headline scores are the RMSE for deterministic forecasts and the Continuous Ranked Probability Score (CRPS) for probabilistic forecasts. For precipitation, the Stable Equitable Error in Probability Space (SEEPS) is often employed as an alternative to RMSE due to the non-Gaussian nature of the variable. In addition to these primary metrics, WB2 supports a broader set of diagnostic scores, including the anomaly correlation coefficient (ACC), forecast bias, spread-skill ratio and spatial energy spectra, enabling a comprehensive evaluation of forecast quality across a range of model types and output formats. The results of these metrics are presented in public scorecards that benchmark the skill of different MLWP and physical models. In doing so, WB2 supports the identification of promising approaches and highlights areas where current models still fall short.

By combining open data, shared evaluation tools and a focus on reproducibility, WB2 plays a critical role in guiding the development of the next generation of MLWP systems.

3.1.2 ERA5

ERA5 (Hersbach et al., 2020) is the fifth generation reanalysis dataset of the global weather and climate by ECMWF under Copernicus Climate Change Service (C3S). In contrast to sparse and irregular observational data, ERA5 provides a spatially and temporally consistent estimate of the state of the atmosphere, land and ocean. Reanalysis refers to the process of combining historical observation data with a NWP model using data assimilation techniques to generate a physically coherent record over time.

ERA5 spans from 1940 to the present and is based on a 0.25° horizontal grid spacing, derived from ECMWF's IFS. It provides data at hourly intervals on 137 vertical levels, extending up to 1 hPa. The assimilation method used is four-dimensional variational assimilation (4D-Var) (Le Dimet and Talagrand, 1986), which ensures temporal consistency by assimilating observations over a time window rather than at single time steps. Compared to its predecessor ERA-Interim (Dee et al., 2011), ERA5 offers significantly improved spatial and temporal resolution, as well as a better representation of physical processes.

Thanks to its global coverage, high temporal and spatial resolution and physical consistency, ERA5 is widely used in atmospheric sciences, particularly as a reference for model

evaluation and in regions with sparse observational coverage. However, it is important to note that ERA5 is not a direct measurement of the atmosphere, but rather a model-based reconstruction constrained by observations. As such, it cannot be considered the absolute "truth" and its accuracy depends on the quality and availability of input observations as well as the performance of the data assimilation and forecast model.

In WB2 framework, the primary dataset for training and evaluating MLWP models is ERA5. To ensure accessibility and computational efficiency, the dataset has been downsampled to a 6-hourly temporal resolution and 13 pressure levels. Additionally, a higher-resolution version with hourly data and 37 pressure levels is also available. Due to its long temporal coverage from 1959 to 2023, ERA5 is particularly well-suited for climatological analyses and long-term model validation and it is widely used as a training dataset for MLWP weather forecasting models. In this thesis, the downsampled version is used with a regular grid of 64x32 gridpoints.

In the WB2 framework, the primary dataset for training and evaluating MLWP models is ERA5. To ensure accessibility and computational efficiency, the dataset has been downsampled to a 6-hourly temporal resolution and 13 pressure levels; a higher-resolution version with hourly data and 37 pressure levels is also available. Due to its long temporal coverage from 1959 to 2023, ERA5 is particularly well-suited for climatological analyses and long-term model validation, and it is widely used as a training dataset for MLWP weather forecasting models. In this study, we use Pangu-Weather forecasts from WB2 and compare them against the downsampled ERA5 reanalysis, which serves as the reference (“ground truth”) for bias correction and model evaluation.

3.1.3 Pangu-Weather

Pangu is a creature in Chinese mythology, considered to be the first living being of the universe. In Yang et al. (2008), the following version of this legend is described: The universe was formed inside a cosmic egg in which Pangu also slept. When he broke free, the heavy egg yolk (Yin) became the earth and the limpid and light part (Yang) became the heaven. To prevent the two parts from collapsing back together, Pangu stood between them. Each year, he grew ten feet taller for 18,000 years, until the earth and the sky were stable enough to remain apart. After his death, his body became geographical features: his limbs became mountains and his breath became the wind.

Based on this legend, the AI weather forecast model Pangu-Weather (Bi et al., 2022, 2023) is named. The model is trained on 39 years (1979 – 2017) of ERA5 reanalysis data. It consists of four deep neural networks, each designed for different lead times of 1 h, 3 h, 6 h and 24 h. Validation was conducted using data from 2019, while the test set comprises data from 2018. Each model operates on 13 pressure levels and includes five upper-air variables: temperature, geopotential, specific humidity, meridional wind and zonal wind. Additionally, four surface variables are evaluated: 2-meter temperature, mean sea level pressure and meridional and zonal 10-meter wind. All models operate at a spatial grid spacing of 0.25°.

To address biases related to the Earth's geometry, such as positional dependencies arising from latitude and longitude, Pangu-Weather incorporates Earth-relative features directly into its architecture. This ensures that forecasts are not skewed by regional imbalances in data representations, a phenomenon known as Earth positional bias.

Figure 3.1: Architecture of the Pangu-Weather model, a 3D Earth-specific transformer (3DEST) for NWP. The model encodes both upper-air and surface variables using patch embeddings, processes them through an encoder-decoder structure with Earth-specific transformer blocks and reconstructs the forecasts via patch recovery. The model operates on 3D spatiotemporal cubes and captures multiscale dependencies across atmospheric layers. Adapted from Bi et al. (2023).

The architecture used in the deep neural network (DNN) is known as 3D Earth-specific transformer (3DEST). It is specially designed to incorporate the Earth's geometry while keeping computational costs low. All variables together form the input of the DNN. As shown in Figure 3.1, the upper-air and surface variables are undergoing a patch embedding with a shifting window mechanism. In this architecture, upper-air and surface variables are divided into non-overlapping patches. For upper-air variables, the patch size is typically (2, 4, 4), representing the vertical pressure levels, latitude and longitude dimensions. Surface variables are processed with a patch size of (1, 4, 4). These patches are then linearly projected into a higher-dimensional feature space. Each resulting vector is referred to as a token, which encodes the meteorological information contained within a single patch. In essence, a token is a compact numerical representation of a small 3D region of the input data and serves as the basic unit on which the attention mechanism operates. To effectively capture local spatial dependencies and reduce artifacts at patch boundaries, a shifting window approach is applied, where the patch boundaries are periodically shifted in subsequent layers. This allows the network to model interactions across neighboring patches more effectively. Within each window, the model applies self-attention. This is a mechanism that dynamically weighs the importance of different spatial locations by learning relationships between them, enabling the network to focus on relevant features across the input data. As a result, the spatial resolution of the input can be reduced while preserving important structural information. After this embedding, the 3D cubes of the upper-air variables and the 2D squares of the surface variables are concatenated along the first dimension to form unified 3D cubes, which are then passed into the 3DEST network. This network follows

an encoder-decoder structure. The encoder comprises two initial layers followed by six layers in which the horizontal resolution is halved and the number of channels is doubled. The decoder mirrors this structure symmetrically.Each encoder or decoder layer is a 3DEST block, which builds on standard vision transformer principle but is adapted on the spherical geometry of the Earth's geometry. The self-attention mechanism operates similarly to that used in image processing. For each token, it identifies other relevant tokens from which to extract new features. Because self-attention is computationally expensive, windowed attention is applied. The grid is divided into smaller windows of size up to 2x12x6 tokens. To enable information exchange between these windows, a shifted window mechanism is introduced. Each window is shifted by half its width, leading to overlapping windows and ensuring continuity, especially across the longitudinal direction, which can thus be modeled periodically. After passing through the 3DEST, the cubes are split back into 3D upper-air and 2D surface components. Finally, a patch recovery step restores the original spatial structure of the data.

Figure 3.2: Hierarchical temporal aggregation in Pangu-Weather. From a given lead time, an algorithm determines the fewest possible steps needed to perform the forecasting. A_0 denotes the input weather state and $\hat{A}_t$ the predicted state after time t. FM1, FM3, FM6 and FM24 refer to the models with corresponding lead times of 1 h, 3 h, 6 h and 24 h. Adapted from Bi et al. (2023).

The process of hierarchical temporal aggregation in Pangu is shown in Figure 3.2. Since Pangu consists of four models, each trained for a specific lead time, any desired lead time can be achieved by iteratively combining these models, choosing the largest affordable lead time. FM1, FM3, FM6 and FM24 refer to models with lead times of 1 h, 3 h, 6 h and 24 h, respectively. Therefore, the forecast result of one step is used as the input for the next step. Figure 3.2 shows a lead time of 56 h which can be realized by executing FM24 twice, FM6 once, FM3 once and FM1 twice. This hierarchical composition allows for greater flexibility and improved accuracy compared to training a fixed-lead-time model.

Despite the advantages of flexibility and efficiency, the hierarchical temporal aggregation also introduces several limitations. First, the accuracy of the final forecast depends on the cumulative performance of all intermediate steps. Errors from short-term models (e.g., FM1 or FM3) can propagate and amplify over time, especially when long lead times require many sequential applications. Second, the aggregation scheme assumes that the output of one model is a suitable input for the next, which is not always the case, especially in regions with highly nonlinear dynamics or during extreme weather events.

Compared to other AI weather forecast models, Pangu-Weather is the first to outperform traditional NWP models. Considering a 5-day forecast of geopotential at 500 hPa, the

operational IFS reaches an RMSE of 333.7. Competing AI models such as FourCastNet (Pathak et al., 2022) report an RMSE of 462.5. In contrast, Pangu-Weather achieves an RMSE of only 296.7.

On WB2, Pangu-Weather outputs are publicly available, including global forecast fields at multiple lead times and different spatial resolutions. Furthermore, the pre-trained model weights and inference code have been made available through an official GitHub repository, enabling reproducibility of the original forecasts and facilitating their integration into downstream applications. However, the training code is not officially released. This limits the ability to retrain or fine-tune the model, and therefore constrains full reproducibility and extensibility.

Recent studies have shown that, despite its outstanding performance on many metrics, Pangu-Weather exhibits systematic biases. In particular, Bouallègue et al. (2024) and related evaluations on WB2 highlight a persistent cold bias in mid-tropospheric temperatures, especially at the 850-hPa level. This bias increases approximately linearly with forecast lead time, indicating that errors accumulate steadily as the model iterates through its hierarchical temporal aggregation scheme. Such behavior suggests that the model gradually loses atmospheric energy over time, potentially due to an insufficient representation of physical processes not fully captured by the data-driven architecture.

In contrast to Pangu-Weather, other MLWP models such as NeuralGCM (Kochkov et al., 2024) exhibit more stable error characteristics, likely due to their hybrid design that combines a physical core with machine-learned parameterizations. Similarly, traditional NWP systems like IFS HRES from ECMWF tend to show smaller and less systematic mid-tropospheric temperature biases, even though their forecast skill may not consistently surpass that of Pangu in other variables. Addressing these biases is crucial for ensuring the long-term reliability of AI-based forecasts, especially in operational settings.

3.2 Online Bias Correction

3.2.1 Methodological Background and Previous Research

To improve the quality of weather forecasts, we apply an online bias correction method. Unlike offline bias correction, which adjusts errors only after the forecast is fully generated, the online approach corrects systematic errors iteratively during the forecast process. This approach is particularly useful for long-range or iterative forecasts, where biases can accumulate and amplify over time. The fundamental idea is to apply bias corrections to each forecast step, which is then used as the initial condition for the next forecast step. As depicted in Figure 3.3, the forecasting process begins with initial conditions at time t, which are used to produce a 24-hour forecast. This forecast is then corrected using a bias correction model before being passed on as the starting point for the next 24-hour forecast step. This cycle continues iteratively for each lead time, ensuring that corrections are integrated at every stage of the forecasting chain.

Figure 3.3: Schematic of the online bias correction approach. Forecasts are produced iteratively in 24-hour steps starting from initial conditions at time t. After each forecast step i, a bias correction is applied and the corrected forecast serves as the initial condition for the next step.

Initial approaches to online bias correction often relied on linear methods due to their simplicity, transparency and low computational cost. One of the most basic techniques involves using a running mean or moving average of past forecast errors to continuously update predictions in real time. A more advanced approach was introduced by Hamill and Whitaker (2006). It uses linear regression based on statistical relationships derived from historical reforecast archives and led to statistically significant improvements in probabilistic skill scores, though no lead time extension was quantified. Similarly, Yuan and Wood (2012) demonstrated the effectiveness of recursive Kalman filters in sequentially correcting hydrological forecasts, showing improved streamflow forecasts with notable reductions in RMSE, particularly at short to medium lead times. This highlights the potential of linear filtering techniques for online applications. In operational ensemble prediction systems, regression-based real-time post-processing frameworks have also been implemented, such as the one proposed by Hagedorn et al. (2008). This enables the dynamic updating of statistical parameters as new observations become available and improves reliability and sharpness of ensemble forecasts, though again, lead time improvements were not explicitly stated. These linear approaches typically assume that forecast bias evolves gradually over time and can be captured using simple autoregressive models.

As computational resources and data availability have grown, online bias correction has increasingly moved toward machine learning-based approaches, which can model nonlinear and higher-dimensional dependencies. These methods allow more flexibility in capturing spatial and temporal patterns of model error. The first approach to integrate deep learning into the online post-processing of NWP models was given by Rasp and Lerch (2018). They explored the use of RNNs for online bias correction, treating the forecast error sequence as a time series and training the network to learn temporal error patterns. Their model improved RMSE and correlation skill over the baseline, particularly for temperature and geopotential height at mid-levels, though improvements in lead time were not quantified.

A deep learning approach, introduced by Laloyaux et al. (2022), updates a CNN online with the most recent observations. It post-processes ensemble mean forecasts and adapts better to changing weather regimes than statistical models. Their method outperformed static models by up to 15 % in Continous Ranked Probability Score (CRPS) and Brier Skill Score

across various lead times, especially during rapid regime shifts. Another method combines dynamical forecasts with machine learning corrections. The model from Zhou et al. (2023) updates every forecast cycle, allowing corrected outputs to serve as initial conditions for the next step. It improved precipitation and temperature forecasts by reducing RMSE by up to 10% in medium-range forecasts (3–5 days), though no explicit gain in forecast horizon was provided.

Another significant contribution came from Vannitsem et al. (2021), who evaluated both linear and nonlinear correction methods in a multi-model ensemble context. They showed that even simple machine learning models like decision trees or support vector machines, when trained online, could outperform traditional static methods in terms of reducing forecast bias and increasing reliability. Some models achieved up to 20% improvement in CRPS and increased correlation scores by up to 0.1 in ensemble-mean forecasts. Ji et al. (2022) implemented sliding-window training with convolutional neural networks, enabling their model to track non-stationary bias patterns over seasonal and interannual time scales. The model reduced mean absolute error (MAE) by up to 12% in monthly temperature forecasts and demonstrated better adaptability to regime changes, though no explicit lead time gain was reported.

Recent studies have also experimented with analog-based and ensemble learning methods. For instance, Lorenz et al. (2021) combines quantile mapping with analog correction techniques in an online setting, where the correction model is regularly updated using the latest available data. This hybrid approach reduced RMSE and bias for daily temperature forecasts, in particular improving extreme event prediction by 10–15%. Ensemble-based online learning was applied to hydrological forecasting, with Sun et al. (2020) finding that it improves both robustness and adaptability in bias correction. Their ensemble learning approach yielded up to 25% reduction in RMSE compared to static models and provided more stable performance across basins and lead times up to 7 days.

A notable advancement in the field is presented by Watt-Meyer et al. (2021), who proposed a method to perform online bias correction of a general circulation model (GCM) using machine learning of nudging tendencies from a hindcast simulation. A random forest is able to make reasonably skillful predictions of the nudging tendencies using only the atmospheric model state as input. When coupled back to the atmospheric model, the machine learning-corrected GCM extends its forecast skill horizon for 500 hPa geopotential height and surface pressure by about a day and for near-surface temperature by about half a day. Furthermore, the root mean square error of the time-mean pattern of precipitation is reduced by about 20%. These improvements come with only slight increase in computational cost. However, the machine learning correction does not improve all aspects of the simulated climate. It improves the intensity distribution of heavy daily surface precipitation greater than 50 mm day^{-1} but generates excessive light precipitation rates between 1 and 4 mm day^{-1}. It also induces significant temperature biases in the polar lower stratosphere after a number of weeks.

Among the reviewed studies, the work by Watt-Meyer et al. (2021) provides the most explicit quantitative improvement in forecast horizon, with a one-day lead time gain for key variables such as geopotential height and surface pressure. Most other studies reported

improvements in error metrics (e.g. RMSE, MAE, CRPS) or skill scores, but did not directly quantify the resulting extension in usable lead time. This highlights both the benefits and the current limitations in evaluating online bias correction methods consistently across different forecasting systems.

3.2.2 Implementation Strategy and Workflow

This section describes the detailed workflow employed to perform the bias correction on the weather forecast fields. The raw forecast data from the model are generated on a high-resolution grid of size 721×1440, which captures fine spatial structures and features. However, correcting biases directly on this fine grid is computationally expensive and potentially prone to overfitting due to the large number of grid points and spatial noise. Therefore, the forecast fields are first spatially aggregated by interpolation onto a coarser grid with dimensions 32×64. This coarser representation captures the larger-scale spatial variability of the forecast fields and associated systematic errors, enabling the bias correction models to focus on relevant spatial scales where biases are most pronounced and physically meaningful.

Once interpolated onto the coarser grid, the bias correction is performed independently for each grid point. This point-wise correction approach allows the model to learn and adjust for spatially varying systematic errors, including regional and seasonal differences, without imposing assumptions of spatial homogeneity. For every forecast initialization and lead time, the correction values are calculated and applied based on the model's learned mapping between forecast errors and predictor variables.

After the bias correction has been applied on the coarse grid, the corrected forecast fields are then interpolated back to the original fine grid. This reverse interpolation ensures that the final corrected forecasts retain the high spatial resolution necessary for detailed weather analysis and downstream applications such as impact modeling or data assimilation.

By combining these steps—downscaling the forecast to a coarser grid for correction, applying point-wise bias adjustments, and upscaling back to the original resolution—the workflow balances computational efficiency with spatial accuracy. It facilitates robust bias correction that respects the spatial complexity of forecast errors while delivering high-resolution corrected forecasts suitable for operational use.

3.3 Regression Methods

To achieve the correction of biases in Pangu, different regression approaches are used. Starting with the rather simple approach of Multiple linear regression (MLR) and the more advanced XGBoost, different types of regression are defined and explained.

3.3.1 Multiple Linear Regression

MLR (Yule, 1907) is a statistical concept to model the relationship between a dependent variable Y and multiple independent variables $X_1, X_2, \ldots X_n$. It is an extension of single linear regression, which considers only one independent variable. This model can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta 2 X_2 + \cdots + \beta_n X_n + \epsilon,$$

where

- Y is the outcome being predicted, also called predictands,
- $X_1, X_2, \ldots X_n$ are the independent variables, also called predictors,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the model, representing the impact of each variable on Y,
- ϵ is the error term, representing the difference between the observed values and the values predicted by the model.

The MLR model training aims to estimate the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ such that the model's predictions are as close as possible to the observed data. To estimate these coefficients, the least squares method is used. The objective of this method is to minimize the sum of the squared residuals. A residual is the difference between the observed value Y_i and the predicted value $\hat{Y}_i$:

$$\epsilon_i = Y_i - \hat{Y}_i, \tag{3.1}$$

where the predicted value $\hat{Y}_i$ is given by the linear model

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta 2 X_{i2} + \cdots + \beta_n X_{in} + \epsilon_i.$$

The total residual sum of squares (RSS) for all datapoints is given by

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

where n denotes the number of datapoints. The aim of the least squares method is now to find coefficients $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the RSS. This means that RSS serves as a loss-function.

Taking the average of the RSS calculated, one obtains the mean squared error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Taking the square root of the MSE yields the RMSE, which is relevant for later considerations.

For MLR, the `LinearRegression` class from the `scikit-learn` library (version 1.5.2) was used (Pedregosa et al., 2012). This library provides a well-established and efficient implementation of linear models suitable for regression tasks.

3.3.2 XGBoost

XGBoost (Chen and Guestrin, 2016) stands for "Extreme Gradient Boosting" and is a machine learning algorithm. It is based on the gradient boosting framework (Friedman, 2001) and widely used in regression and classification tasks, in particular in applications involving large datasets and structured data. XGBoost enhances traditional gradient boosting by incorporating regularization, parallel computation and optimized handling of missing data. This leads to improved performance and generalization.

The core idea behind XGBoost is the sequential training of decision trees, where each new tree corrects the residuals of the previous ensemble. A decision tree is a simple, tree-like model used to make predictions. It works by asking a series of yes/no questions about the input features. Each question splits the data into smaller groups that are more similar in terms of the target value. This process continues until the data is divided into small enough groups. At the end of each branch, which is called the leaf, the model assigns a prediction based on the average outcome in that group.

The prediction in iteration t is given by

$$\hat{Y}_i^{(t)} = \hat{Y}_i^{(t-1)} + \eta f_t(X_i)$$

where f_t is the prediction of the newly added tree at iteration t for input X_i and η is the learning rate, which controls the contribution of each new tree to the overall model.

In each iteration step the decision trees aims to minimize the residual errors (Equation (3.1)) from the previous predictions. The loss function to be minimized can, for instance, be the MSE,

$$\ell_{MSE} = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2,$$

or the MAE,

$$\ell_{MAE} = \sum_{i=1}^{N} |Y_i - \hat{Y}_i|, \tag{3.2}$$

where N is the number of data points, Y_i the true value and $\hat{Y}_i$ the model prediction.

To prevent overfitting and control model complexity, XGBoost introduces a regularization term to the objective function. This penalizes overly complex trees and improves generalization. The regularized objective function in iteration t becomes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} \ell(Y_i, \hat{Y}_i^{(t)}) + \sum_{k=1}^{M} \Omega(f_k),$$

where $\ell(Y_i, \hat{Y}_i^{(t)})$ is the chosen loss function (e.g., MAE or MSE) and the regularization term is defined as

$$\Omega(f_k) = \gamma T_m + \frac{1}{2}\lambda \sum_{j=1}^{T_m} \omega_j^2.$$

The components are:

- M the total number of trees,
- T_m the number of leaves in the m-th tree,
- f_k the prediction function of the k-th tree,
- γ, λ penalty parameters,
- ω_j the weight of the j-th leaf.

In the context of bias correction, a meaningful loss function is the MAE (3.2), as it minimizes the absolute difference between predictions and true values. To obtain a good balance between complexity and predictive performance, the hyperparameters of the XGBoost model, such as the maximum tree depth, learning rate and number of boosting rounds, must be chosen carefully. This helps prevent overfitting while ensuring that the model captures relevant patterns in the input features. Cross-validation is commonly used to identify optimal parameter settings and early stopping can help avoid overfitting by excessive training when no further improvement is observed. In this thesis, the focus lies on achieving a robust correction of systematic errors while maintaining model interpretability and computational efficiency.

For the gradient-boosted decision tree models, the `XGBRegressor` class from the `xgboost` library (version 2.1.1) was employed (Chen and Guestrin, 2016). XGBoost offers advanced features such as regularization, tree pruning, and efficient handling of missing data, which make it highly suitable for complex regression problems in machine learning applications.

3.4 Bias–Variance Decomposition

To assess the effectiveness of applied corrections, the MSE is decomposed into two components: the variance and the squared bias. This decomposition is a fundamental result in statistical learning theory and provides insights into the sources of prediction error. The following formulation is based on Hodson et al. (2021).

Theorem 3.4.1. *The MSE is decomposable into the variance and the square of the bias:*

$$MSE = Var + Bias^2. \tag{3.3}$$

Proof. Let $\hat{\theta}$ be an point estimator for a parameter θ. Then the bias of $\hat{\theta}$ is given as $Bias(\hat{\theta}, \theta) = E_\theta[\hat{\theta}] - \theta$. Then, one can write the MSE as $MSE(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$, the variance as $Var_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$ and the bias as .

With the linearity of the expected value yields

$$\begin{aligned}
MSE(\hat{\theta}) &= E_\theta[(\hat{\theta} - \theta)^2] \\
&= E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}] + E_\theta[\hat{\theta}] - \theta)^2] \\
&= E_\theta\left[(\hat{\theta} - E_\theta[\hat{\theta}])^2 + 2(\hat{\theta} - E_\theta[\hat{\theta}])(E_\theta[\hat{\theta}] - \theta) + (E_\theta[\hat{\theta}] - \theta)^2\right] \\
&= E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2] + 2E_\theta[\hat{\theta} - E_\theta[\hat{\theta}]](E_\theta[\hat{\theta}] - \theta) + (E_\theta[\hat{\theta}] - \theta)^2 \\
&= Var_\theta(\hat{\theta}) + 2(E_\theta[\hat{\theta}] - E_\theta[\hat{\theta}])(E_\theta[\hat{\theta}] - \theta) + (E_\theta[\hat{\theta}] - \theta)^2 \\
&= Var_\theta(\hat{\theta}) + (Bias(\hat{\theta}, \theta))^2.
\end{aligned}$$

This shows the desired relationship. □

In the proof above, the first step consists of adding and subtracting the expected value $E_\theta[\hat{\theta}]$ within the squared term, a technique often referred to as "adding zero". The binomial expansion then separates the expression into three components: the variance of the estimator, a cross-term, and the squared bias. By the linearity of expectation, the cross-term vanishes because the expected deviation from the mean is zero. Consequently, the mean squared error can be expressed as the sum of the variance and the squared bias, which clearly separates the error due to variability of the estimator and the error due to systematic deviation from the true parameter.

This decomposition is essential for understanding the trade-offs inherent in statistical estimation and model fitting. By distinguishing between variance and bias, it enables more informed decisions regarding model complexity and correction methods. Minimizing the MSE therefore involves balancing these two components to achieve the best possible predictive performance.

4 Results

The results of the bias correction applied to 850 hPa temperature forecasts from Pangu-Weather are presented and analyzed in this chapter. All models discussed in this chapter are trained on data from the years 2018 to 2021, with the year 2022 reserved for independent testing.

Section 4.1 begins by characterizing the systematic biases present in the Pangu-Weather forecasts. The analysis explores potential sources of these biases and investigates how they evolve with increasing lead time. In addition, the seasonal dependence and vertical structure of the biases are examined to provide a comprehensive understanding of their spatiotemporal behavior. In Section 4.2, different offline bias correction approaches are tested and compared. The focus lies first on multiple linear regression as a classical statistical method, followed by the use of gradient boosting via the XGBoost algorithm. For both methods, model performance is assessed using various configurations and predictor combinations, with particular attention to the effect of early stopping and loss function choice. Finally, Section 4.3 applies one of the offline models introduced in Section 4.2 in an online correction framework. The performance of the online bias correction is evaluated in terms of its ability to dynamically adjust Pangu-Weather forecasts and reduce systematic errors under operational-like conditions.

4.1 Systematic Biases in Pangu-Weather

4.1.1 Temporal Development of Forecast Biases

Pangu-Weather develops a systematic negative bias in the global 850-hPa temperature that increases with forecast lead time. As shown in Figure 4.1 and in line with Bouallègue et al. (2024), this bias follows an approximately linear trend, indicating that forecast errors accumulate steadily as the forecast horizon extends. In contrast, the biases observed in forecasts from the IFS HRES model developed by ECMWF are generally smaller in magnitude and show less pronounced variation over time. NeuralGCM (Kochkov et al., 2024), a recent developed forecasting model that combines a traditional dynamical core with a machine learning model representing non-resolved processes, demonstrates even greater consistency in its bias behavior, maintaining a comparatively stable and minimal bias in 850-hPa temperature across different lead times. This suggests that while Pangu-Weather performs competitively at short lead times, its forecast reliability in the mid-troposphere

diminishes more noticeably with increasing temporal range when compared to these other state-of-the-art models.

The observed cooling trend in Pangu-Weather forecasts may also hint at an underlying energy imbalance within the model. If the atmosphere is persistently cooling over time, this could imply a systematic loss of internal energy, unless it is counterbalanced by increases in other energy forms such as kinetic energy. However, investigating the full energy budget in Pangu-Weather is not straightforward, as it is a purely data-driven model that does not explicitly model or conserve physical quantities like energy. Unlike traditional NWP models, Pangu-Weather does not provide diagnostics for internal, kinetic or total energy, nor does it solve governing physical equations.

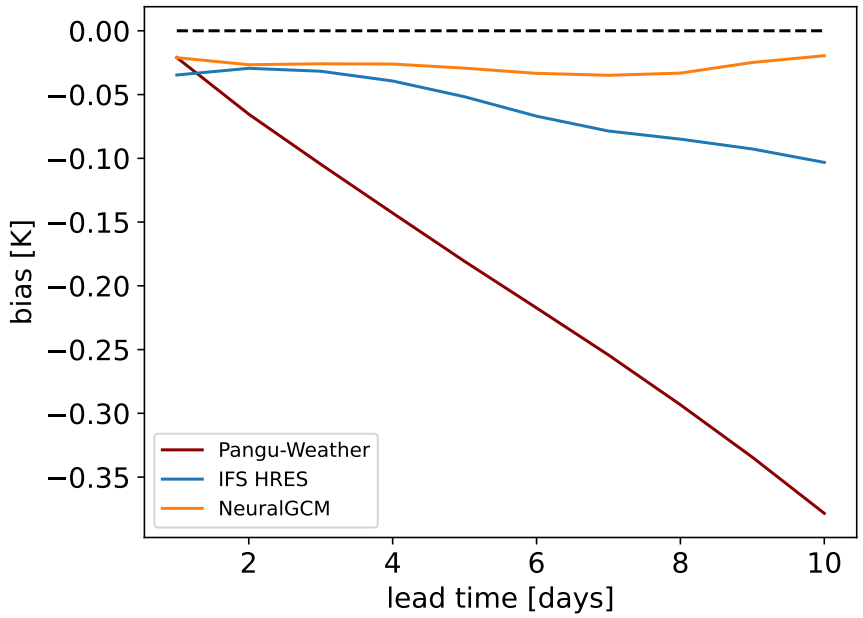
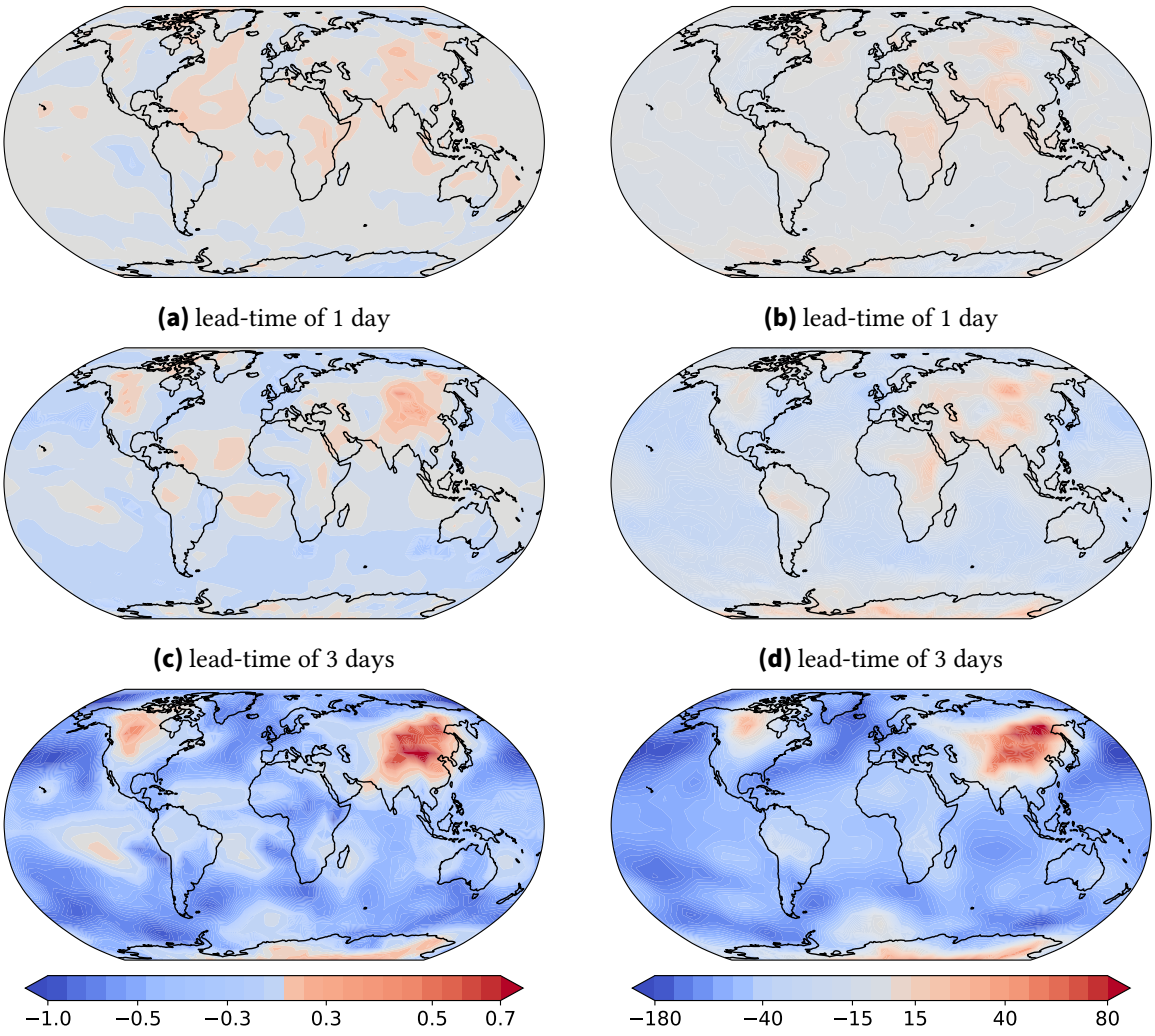


Figure 4.1: Global mean bias of 850-hPa temperature forecasts as a function of lead time (in days) for three different models: Pangu-Weather (dark red), IFS HRES (blue), and NeuralGCM (orange).

To further investigate the nature and potential causes of this growing negative bias in 850-hPa temperature in Pangu-Weather, the spatial distribution of the bias is analyzed at selected lead times of 1 day, 3 days, and 10 days. These snapshots are shown in Figure 4.2 in combination with the 500—1000 hPa geopotential thickness, which represents the 500–1000 hPa mean temperature following the hypsometric equation. Accordingly thickness and temperature biases should be correlated such that by comparing the spatial patterns of the temperature bias with the corresponding geopotential thickness fields, it is possible to assess whether the evolution of the bias is physically consistent.

As seen in Figure 4.2a, a temperature bias is already present at short lead times, although its magnitude remains relatively small. The order of magnitude of the temperature bias is ±0.2 K. Both positive and negative biases occur, with positive biases primarily found in the Northern Hemisphere. In particular, the positive biases appear over the northern Atlantic Ocean, eastern Asia and parts of Africa. The bias of 1000-500 hPa geopotential thickness



(a) lead-time of 1 day

(b) lead-time of 1 day

(c) lead-time of 3 days

(d) lead-time of 3 days

(e) lead-time of 10 days

(f) lead-time of 10 days

Figure 4.2: Spatial distribution of forecast bias at different lead times for two atmospheric variables: 850-hPa temperature (left column) and 500–1000 hPa geopotential thickness (right column). **(a)**, **(c)**, and **(e)** show the bias in 850-hPa temperature (in K) for lead times of 1, 3, and 10 days, respectively. **(b)**, **(d)**, and **(f)** depict the corresponding bias in 500–1000 hPa geopotential thickness.

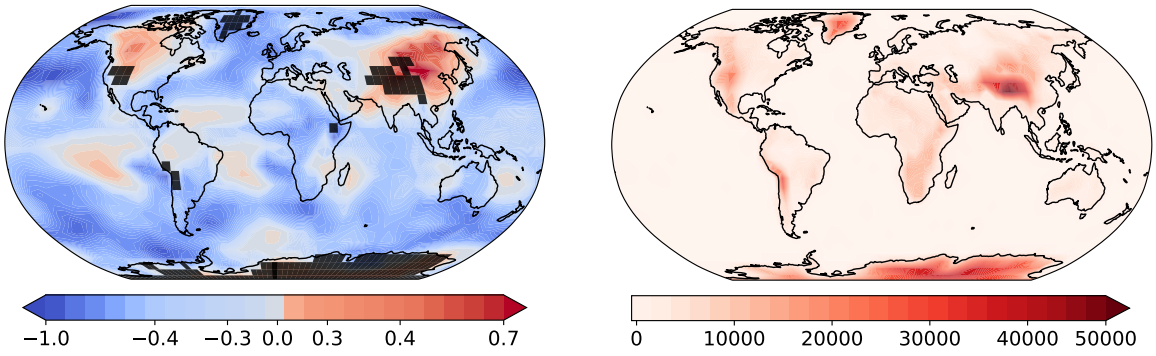
in Figure 4.2b shows, at least over land, a similar spatial pattern as the temperature bias. The order of magnitude is ±20 gpm. The geopotential thickness of a pressure layer refers to the average temperature of that layer. Therefore, a colder troposphere corresponds to a reduced vertical thickness. The agreement between both fields suggests that the cold bias reflects a broader error in the model's thermodynamic structure.

At a lead time of 3 days, as seen in Figure 4.2c and Figure 4.2d, the magnitude of the biases increases. The cold bias in 850-hPa temperature strengthens, in particular over the Southern Ocean. The warm biases also intensify, especially over East Asia and Canada. The positive biases over East Africa and the northern Atlantic Ocean in Figure 4.2a are reduced. These regions also exhibit a clear change in geopotential thickness, which is physically consistent

with the observed temperature anomalies. Nevertheless, differences in the temperature bias and geopotential thickness bias over ocean, especially in the tropics and subtropics, are visible. The order of magnitude for the temperature bias is ±0.4 K and for geopotential thickness ±40 gpm.

After ten days, as presented in Figure 4.2e and Figure 4.2f, the negative bias becomes the dominant signal in temperature and geopotential thickness bias. The 850 hPa temperature shows widespread cold anomalies across much of the mid- and high latitudes of the Southern Hemisphere and the seas of the Northern Hemisphere. Similarly, the geopotential thickness is substantially underestimated in the same regions. At the same time, the warm biases over East Asia, parts of Canada and Antarctica intensify considerably. The order of magnitude for the temperature bias is between −1 K and 0.7 K. For geopotential thickness it is between −180 gpm and 80 gpm.

The spatial alignment of cold temperature and reduced thickness indicates a persistent and physically consistent cold bias throughout the lower troposphere. This suggests a systematic drift in the model forecasts, which becomes increasingly pronounced with lead time.



(a) Mask **(b)** Surface geopotential [gpm]

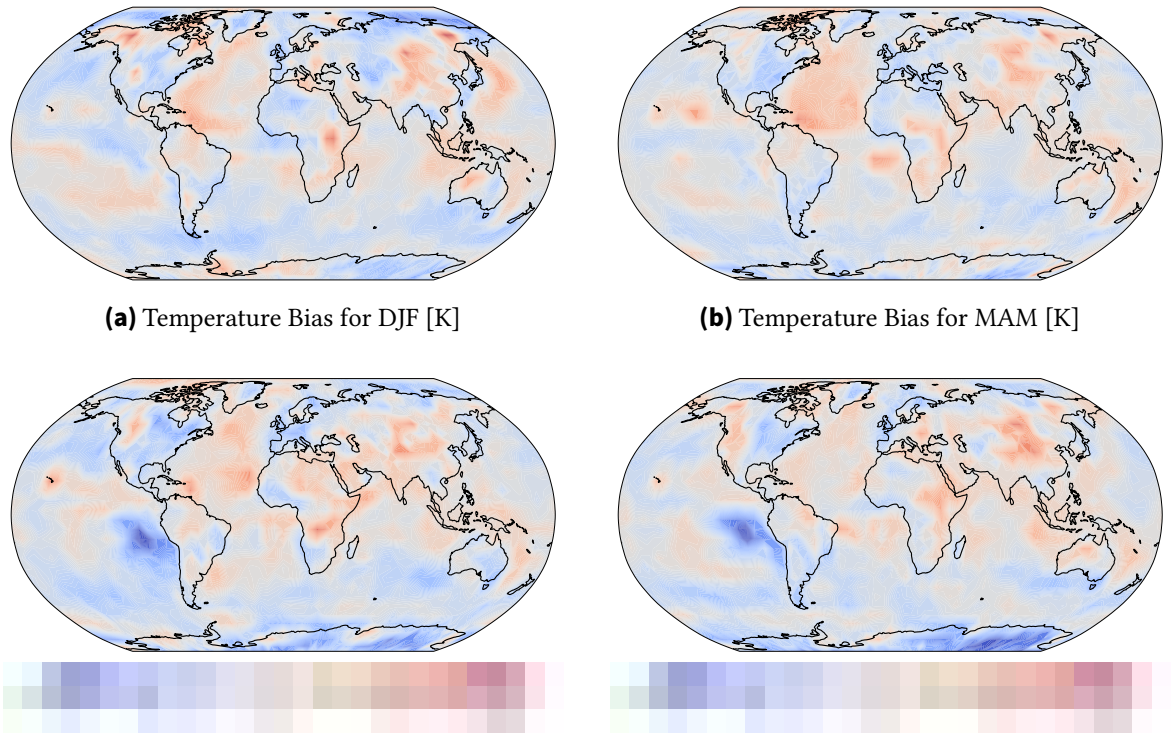
Figure 4.3: Illustration of masked regions based on orographic constraints. **(a)** shows a mask highlighting grid points where the 850 hPa geopotential height is below the surface geopotential height. This indicates locations where the 850 hPa level lies beneath the terrain. **(b)** shows the surface geopotential height (in geopotential meters, gpm), providing a reference for global topography and its influence on the mask.

It is important to note that in regions with high topography, such as the Himalayas, Andes, Antarctica and parts of the Rocky Mountains, the 850 hPa pressure level lies below the surface. In these areas, ERA5 provides values through interpolation. Since Pangu-Weather is trained on ERA5 data, the model adopts this interpolation. As a result, regions with higher altitude should be interpreted with caution. To identify affected areas, a mask is applied to all grid points where the geopotential at 850 hPa exceeds the surface geopotential. This mask is shown in Figure 4.3a, overlaid on the 850 hPa temperature bias at a lead time of 10 days. For reference, Figure 4.3b displays the geopotential at surface. Since the geopotential increases with height in a gravitational field, the surface geopotential provides a proxy for elevation. Higher values indicate regions of higher altitude. Overall, the mask aligns well to the topography and helps explain parts of the large positive biases. However, not

all positive biases can be attributed to elevation effects. In particular, over East Asia and North America, strong biases occur on the northern flanks of high terrain. This suggests that additional mechanisms contribute to the model error in these regions.

4.1.2 Seasonal Variation of Forecast Biases

While analyzing the overall global bias provides a first understanding of systematic errors in the model, it is important to recognize that these biases are not necessarily constant throughout the year. Seasonal variations in atmospheric circulation can cause the magnitude and spatial structure of the bias to change significantly between seasons. Therefore, separating the bias by season allows for a more detailed and accurate assessment, helping to uncover seasonal dependencies that would otherwise be hidden in the global annual mean. The seasons are defined according to the meteorological convention: December-January-February (DJF) for winter, March-April-May (MAM) for spring, June-July-August (JJA) for summer and September-October-November (SON) for autumn. Figure 4.4 displays the 850 hPa temperature biases at a lead time of one day, categorized by seasons.



(a) Temperature Bias for DJF [K]

(b) Temperature Bias for MAM [K]

(c) Temperature Bias for JJA [K]

(d) Temperature Bias for SON [K]

Figure 4.4: Spatial distribution of temperature bias (in K) at the 850 hPa level after a lead time of 24 hours, averaged over the four meteorological seasons: **(a)** Winter (DJF), **(b)** Spring (MAM), **(c)** Summer (JJA) and **(d)** Autumn (SON).

Overall, the general pattern of the global bias is relatively stable across the different seasons, especially in the Northern Hemisphere. But important differences in regional structures and

magnitudes can be observed. Throughout all seasons, a tendency toward positive biases over eastern Asia, Canada and western Africa is apparent, whereas negative biases, in particular over parts of North America are consistently present.

In boreal winter, shown in Figure 4.4a, negative temperature biases are most pronounced over tropical oceans and parts of North America and West Asia. Positive biases appear over eastern Asia, western Africa and Canada. Since these positive biases occur mainly on the Northern Hemisphere, where winter conditions prevail, this suggests that factors such as extensive snow cover and persistent cold air masses likely influence the temperature forecast errors in these regions.

During spring, presented in Figure 4.4b, the general bias pattern remain similar but tend to weaken slightly compared to winter. The positive biases over East Asia, Canada and West Africa are still visible but less pronounced. The cold biases observed in winter also tend to diminish. This transitional season, characterized by changing large-scale circulation patterns, appears to moderate some of the stronger biases seen during the winter months.

Figure 4.4c shows that the general pattern of the spring biases persists into summer. However, a notable strengthening of cold biases over the Southern Hemisphere is evident. In particular, a strong negative bias develops in the Southeast Pacific region. This increase in bias over southern oceans is likely linked to the austral winter and related processes, such as enhanced surface cooling and weaker boundary layer mixing. Meanwhile, positive biases over East Asia and Canada persist, with magnitudes similar to previous seasons. These observations suggest that the model tends to overestimate temperatures in some continental regions during summer, while underestimating temperatures over the Southern Hemisphere during its winter.

Autumn, as shown in Figure 4.4d, again reveals a bias structure similar to MAM, with generally weaker biases compared to JJA. The Southern Hemisphere cold biases remain present but decrease in magnitude, consistent with the transition from austral winter to spring. Warm biases over East Asia and northeastern Canada are still observable but less intense than in DJF.

Overall, while the large-scale structure of the temperature bias remains broadly consistent across seasons, there are clear modulations in the intensity and regional focus of the biases. In particular, the amplification of cold biases in the Southeast Pacific during JJA and the persistent warm biases over East Asia, Canada and Antarctica across all seasons highlight the importance of considering seasonal variations when evaluating and correcting model errors. These findings suggest that systematic errors in Pangu-Weather are not uniform throughout the year and season-specific bias correction approaches could be beneficial for improving forecast skill.

4.1.3 Vertical Structure of Forecast Biases

To understand, where and why large biases after a lead time of 24 h develop, vertical profiles of these regions are interesting. In Figure 4.5 two of those regions are marked. The region

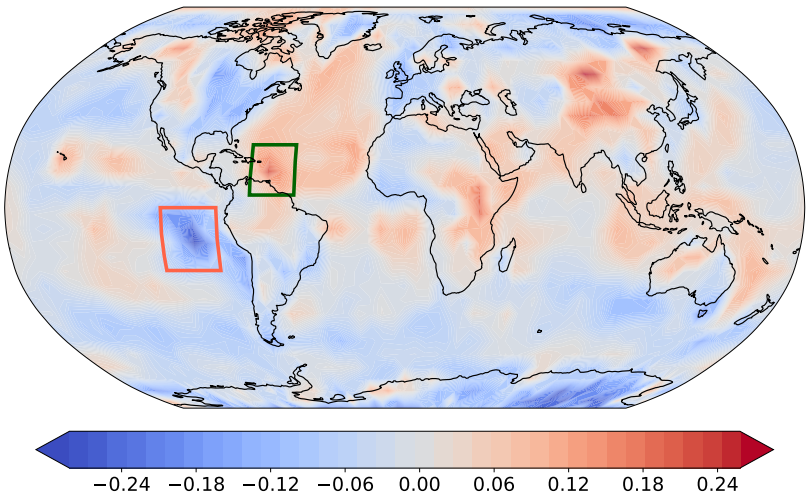


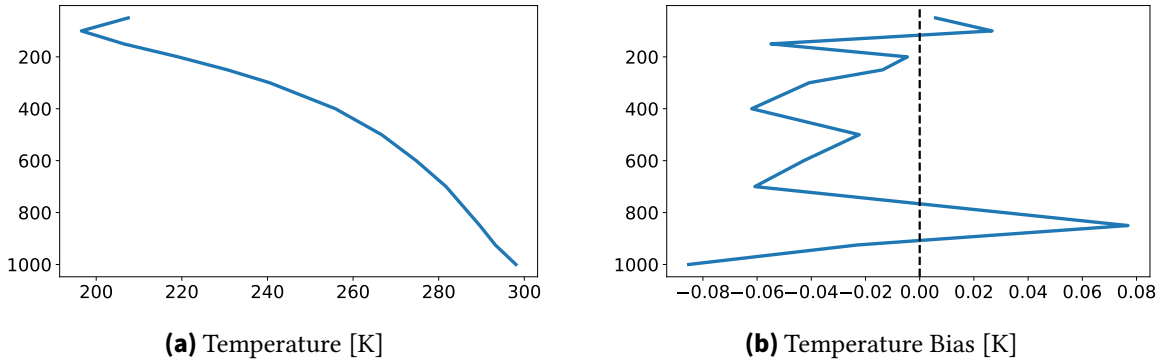
Figure 4.5: Spatial distribution of the 850 hPa temperature bias after a 24-hour lead time, highlighting two regions with pronounced systematic errors. The orange box marks the Southeast Pacific stratocumulus region, characterized by a strong cold bias. The green box indicates the Caribbean region, where the model exhibits a pronounced warm bias. These areas are selected for further analysis due to the persistence and magnitude of their seasonal temperature biases.

green box in the western Atlantic Ocean, next to the Caribbean, shows a large positive bias. In contrast, the orange box in the eastern Pacific Ocean shows a dominant negative bias. In Figure 4.6 and Figure 4.7 the vertical profiles for both regions of the temperature, the temperature bias, the geopotential bias and the specific humidity bias are given.

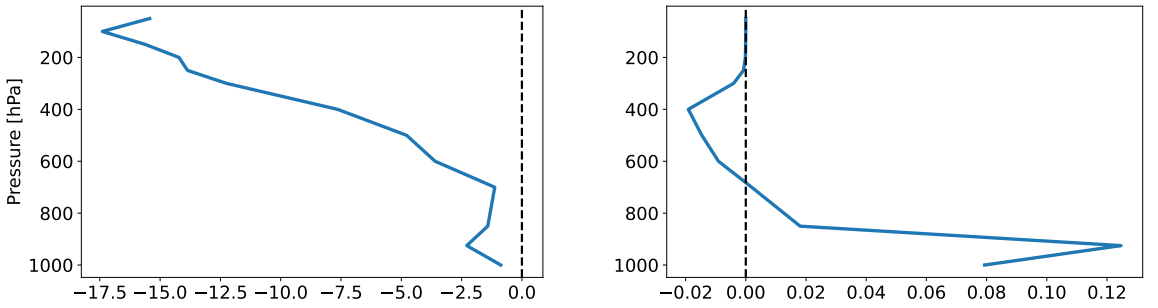
Figure 4.6a shows the vertical profile of the temperature from 1000 hPa to 50 hPa in the Caribbean region over the western Atlantic Ocean. As expected for a tropical oceanic atmosphere, the temperature decreases steadily with height, reflecting a typical moist adiabatic lapse rate. The profile appears smooth and stable, indicative of a well-stratified troposphere and a warm, moist boundary layer near the surface.

The corresponding temperature bias, shown in Figure 4.6b, is mostly negative throughout the column. This indicates a general underestimation of temperature by the model. A notable exception is a narrow layer around 850 hPa, where a positive bias occurs. Below 850 hPa, in the near-surface layers, the bias turns negative again. This pattern suggests that the model underestimates temperatures at the surface and aloft, with a warm anomaly confined to a relatively shallow mid-level. Such vertical inconsistencies may point to limitations in how Pangu-Weather represents the vertical temperature structure and balances radiative cooling and convective heating.

A likely explanation for these vertical bias patterns lies in unresolved or misrepresented cloud processes. In tropical maritime environments, cloud-related processes, such as shallow cumulus convection, deep convective updrafts, stratiform cloud layers and the associated latent heat release, are key drivers of vertical temperature and humidity profiles. If these processes are not adequately captured, especially those on subgrid scales, systematic model biases are to be expected.



(a) Temperature [K] **(b)** Temperature Bias [K]



(c) Geopotential Bias [gpm] **(d)** Specific Humidity Bias [g/kg]

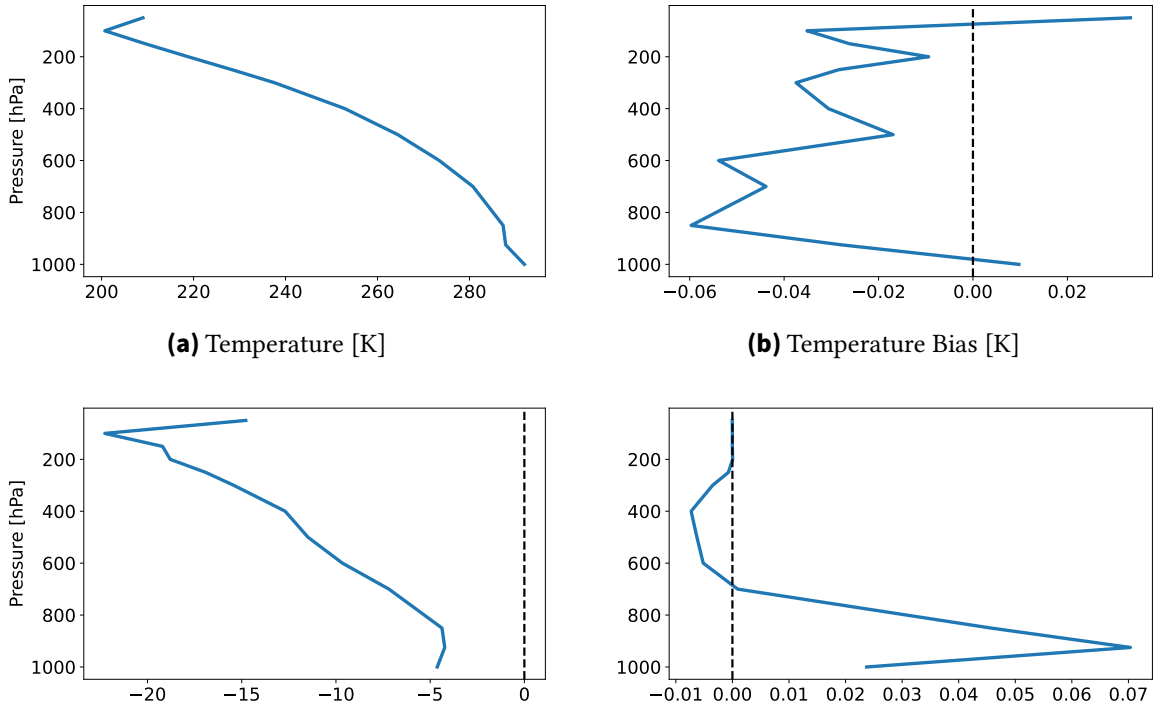
Figure 4.6: Vertical profiles of key atmospheric variables in the Caribbean region, averaged over the area marked in Figure 4.5. Shown are: **(a)** the mean temperature profile, **(b)** the temperature bias, **(c)** the geopotential height bias and **(d)** the specific humidity bias.

This limitation is particularly relevant for Pangu-Weather, as it is a transformer-based, data-driven forecast model that generates future atmospheric states based solely on learned statistical relationships from past data, without simulating the underlying physical processes. It does not explicitly resolve convection, cloud microphysics, or radiative transfer. Therefore, the complex thermodynamic interactions between moisture, clouds, and radiation, especially in cloud-rich tropical environments, cannot be dynamically represented. Instead, their effects must be inferred implicitly from training data. While this architecture enables extremely fast and often skillful large-scale forecasts, it limits the model's ability to reproduce vertical structures that depend on unresolved physics.

The positive temperature bias at 850 hPa may reflect an overestimation of latent heat release due to excessive convective activity, likely linked to the surplus of low-level moisture seen in the specific humidity bias. Conversely, the negative temperature bias near the surface could be caused by missing evaporative cooling below cloud base or by excessive cloud shading, reducing surface heating. Additionally, cloud-radiative effects, such as longwave warming in the upper troposphere and shortwave cooling near the surface, are not explicitly represented in Pangu-Weather, which may further contribute to the observed vertical inconsistencies. These findings highlight the inherent architectural limitations of data-driven models like Pangu-Weather in reproducing physically consistent thermodynamic profiles, particularly in the tropics.

In Figure 4.6c, the bias of the geopotential height shows a consistent negative offset throughout most of the troposphere, reaching values below −15 gpm in the mid-to-upper levels. Since geopotential height is directly influenced by the integrated temperature profile, a systematic cold bias, such as that seen here, naturally results in a compressed vertical structure and lower geopotential heights. This bias is therefore consistent with the temperature underestimations described earlier.

Figure 4.6d illustrates the bias in specific humidity. A pronounced positive bias is observed in the lower troposphere, especially below 700 hPa, with peak deviations of up to +0.12 g/kg. This indicates that Pangu-Weather overestimates near-surface moisture over the ocean. Such overestimations are common in both data-driven and traditional NWP models over tropical oceans, where moisture fluxes from the surface are strong and challenging to parameterize correctly. Above 600 hPa, the bias decreases and eventually becomes slightly negative, indicating better agreement with ERA5 or possibly an underestimation of upper-tropospheric moisture.



(a) Temperature [K]

(b) Temperature Bias [K]

(c) Geopotential Bias [gpm]

(d) Specific Humidity Bias [g/kg]

Figure 4.7: Vertical profiles of key atmospheric variables in the Southeast Pacific stratocumulus region, averaged over the area marked in Figure 4.5. Shown are: **(a)** the mean temperature profile, **(b)** the temperature bias, **(c)** the geopotential height bias and **(d)** the specific humidity bias.

In contrast to the previously discussed region, the annually averaged vertical profiles in the stratocumulus (Sc) region off the west coast of South America (Figure 4.7) reveal characteristic model biases associated with the representation of low marine clouds. Although the profiles represent means over the entire year, their structure and magnitude reflect

the dominant physical processes in this region, many of which are strongly seasonal in nature.

The annual mean temperature profile is shown in Figure 4.7a. In the lower troposphere, a nearly isothermal layer is apparent between approximately 950 and 800 hPa, suggesting a weak thermal stratification. Although not a sharp inversion, this structure indicates a region where vertical mixing is suppressed, which is favorable for stratocumulus formation. Such conditions are typically found in regions with persistent marine boundary layer clouds, which are common in this part of the southeastern Pacific due to the influence of cold sea surface temperatures associated with the Humboldt Current. In this region, a pronounced temperature inversion is generally expected as a defining feature of the stratocumulus-topped boundary layer. The absence of a clear inversion in the annual mean profile may be attributed to temporal averaging or insufficient vertical resolution, which can obscure the typically sharp transition in temperature.

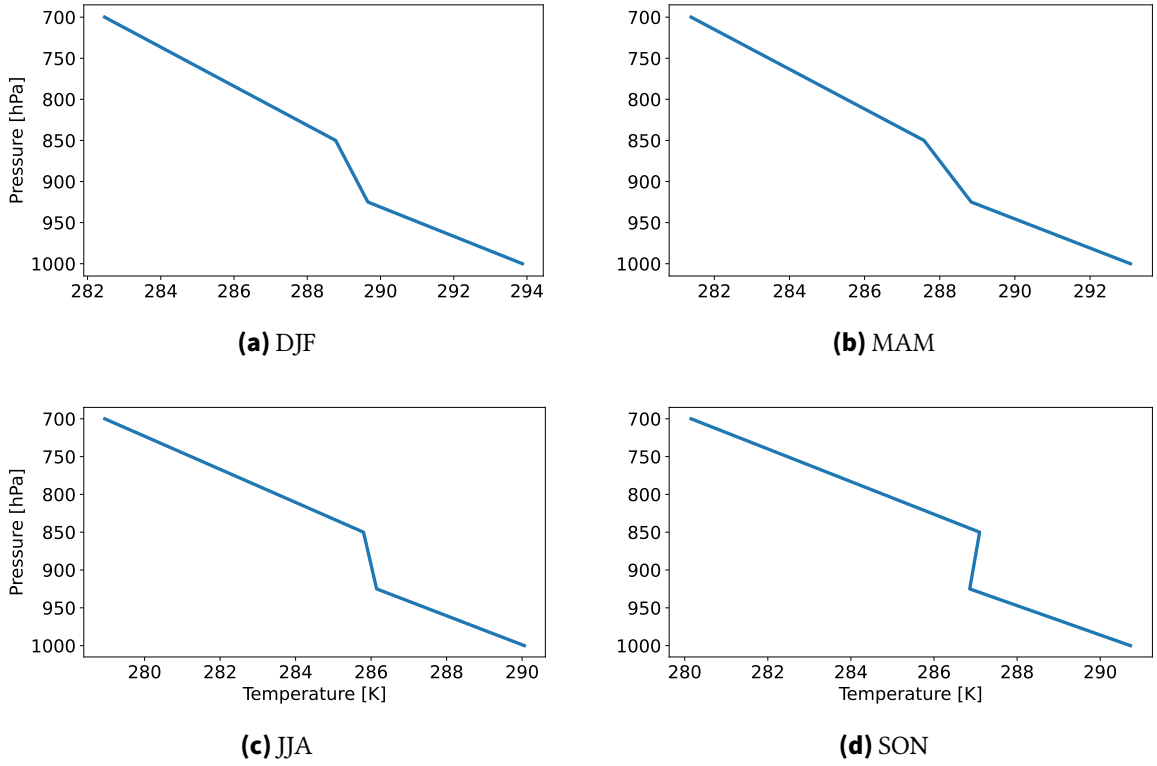
The vertical structure of the mean temperature bias is illustrated in Figure 4.7b. The profile reveals predominantly negative temperature bias extending from the lower troposphere up to approximately 100 hPa. This cold bias indicates that the model systematically underestimates temperatures throughout much of the troposphere, with the largest deviations occurring in the lower and middle levels. These cold biases likely reflect an overestimation of low-level cloud cover in the model. Excessive stratocumulus clouds can enhance longwave radiative cooling near the cloud top, resulting in systematically lower temperatures within the boundary layer. This suggests that the model tends to simulate too frequent or too persistent stratocumulus clouds in this region.

The geopotential height bias, as shown in Figure 4.7c, exhibits predominantly negative values throughout the lower and middle troposphere. This pattern aligns with the cold temperature bias in the same region, since lower temperatures reduce the vertical thickness of the atmosphere and consequently lead to underestimated geopotential heights. In addition, the proximity of the Andes mountain range may further influence this bias: terrain-induced blocking of the airflow can lead to a buildup of air mass on the western slopes, altering the local pressure distribution and enhancing the negative anomaly in geopotential height.

A positive bias in specific humidity near the surface is evident in Figure 4.7d. This suggests that the model tends to simulate a boundary layer that is moister than observed. When considered alongside the concurrent cold bias in temperature, this pattern implies that the model retains excessive moisture within the weakly stratified lower troposphere. Such behavior is characteristic of models that overestimate low-level cloud cover, particularly in persistent stratocumulus regimes. In these environments, overpredicted cloudiness is frequently associated with a combination of negative temperature and positive humidity biases in the near-surface layers.

In summary, although Figure 4.7 shows annual mean profiles, the characteristic signatures of overestimated stratocumulus conditions with a cold and moist lower troposphere and a negative geopotential bias are clearly present. These features are particularly associated with the seasonal maximum of the Humboldt Current during JJA, even if not directly resolved in the figure. The orographic influence of the Andes further amplifies the geopotential

bias near the coast. The seasonal evolution of the thermal inversion and its impact on cloudiness will be explored in more detail in the following Figure 4.8. This figure shows the temperature profiles in the lower troposphere between 1000 hPa and 700 hPa for the four meteorological seasons.



(a) DJF **(b)** MAM **(c)** JJA **(d)** SON

Figure 4.8: Vertical profile of the temperature between 1000 hPa and 700 hPa for different seasons: **(a)** DJF, **(b)** MAM, **(c)** JJA and **(d)** SON. The profiles highlight seasonal variations in the vertical structure of the bias in the lower troposphere.

During SON, as shown in Figure 4.8d, a pronounced temperature inversion is evident in the vertical profile. This inversion, which refers to an increase in temperature with height near the surface, is a characteristic feature of stratocumulus-topped boundary layers and is well developed in this season. In contrast, in the other seasons, the mean profiles show no such inversion on average. JJA and DJF can be considered transitional periods. While inversions can occur during individual episodes, they are not sufficiently frequent or persistent to appear clearly in the seasonal mean. In MAM, the inversion signal is weakest and the vertical structure is generally closer to a well-mixed boundary layer.

A similar seasonal pattern is observed in Sea Surface Temperature (SST). Figure 4.9 shows the SSTs for the seasons. The SSTs are highest in MAM and lowest in SON. This variation is largely driven by the seasonality of the Humboldt Current, a cold eastern boundary current along the west coast of South America. The upwelling of cold subsurface waters is strongest during SON, leading to reduced SSTs. In contrast, during MAM, the upwelling is weaker, resulting in warmer surface waters.

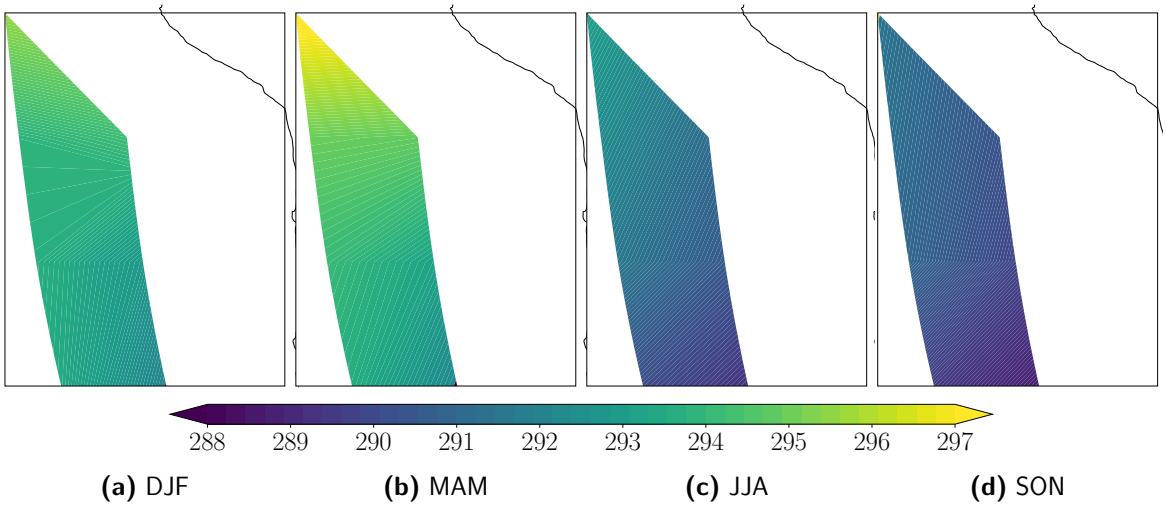


Figure 4.9: Seasonal distribution of the Sea Surface Temperature in Kelvin for different seasons: **(a)** DJF, **(b)** MAM, **(c)** JJA and **(d)** SON. The spatial patterns indicate seasonal variations in the Humboldt current.

The combination of colder SSTs and a well-established inversion layer in SON provides favorable conditions for the development of extensive stratocumulus cloud decks. These clouds are known to be persistent and widespread in regions with strong marine boundary layer inversions, especially under the influence of cold ocean currents. The model appears to simulate excessive cloud cover under these conditions, which leads to a systematic overestimation of cloud-related cooling. As a result, a pronounced negative temperature bias is observed near the surface in this region during SON.

In summary, the seasonal variations in bias are closely linked to the underlying atmospheric and oceanic conditions. In particular, the strength of the temperature inversion and the SSTs, both influenced by the Humboldt Current, play a key role in modulating cloud formation processes and the resulting forecast errors.

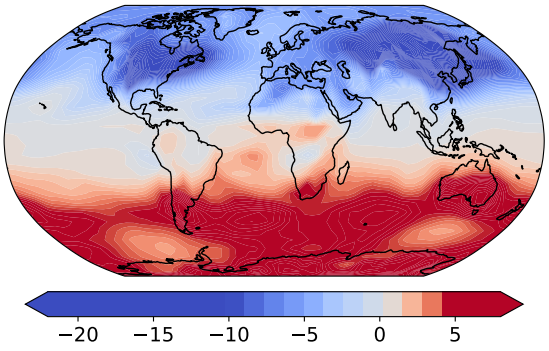
4.2 Offline Bias Correction

In the offline bias correction, the methods described in Section 3.3 are applied on a forecast from Pangu-Weather with a lead time of 24 h. In Section 4.2.1 the correction of the forecast result with MLR. As an alternative, a correction with XGBoost is applied in Section 4.2.2. The results are compared with ERA5 ground truth data.

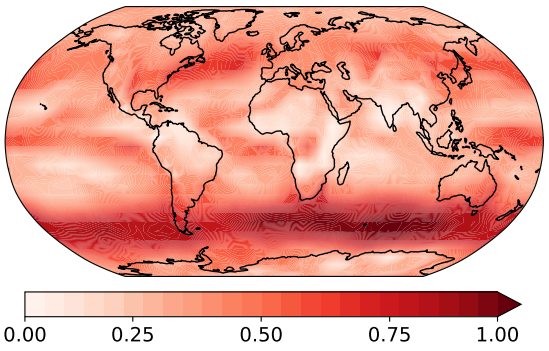
4.2.1 Multiple Linear Regression

To correct biases occurring in 850 hPa temperature forecasts from Pangu-Weather, a MLR model is applied for each individual grid point. The target variable is the model bias, defined as the difference between Pangu-Weather temperature forecasts and ERA5 reanalysis

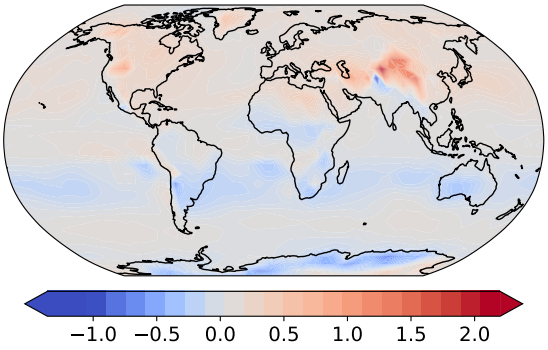
data. The model was trained using data from the years 2018 to 2021 and evaluated on the independent test year 2022. As predictor variables, the day of the year, the predicted temperature and the meridional wind component were selected. The selection of these predictors is based on physical intuition and preliminary analyses. The day of the year serves as a proxy for seasonal variations that influence systematic errors. It is defined as a periodic function with a maximum during summer and a minimum during winter. This periodic representation ensures that days with similar climatological characteristics, such as late spring and early autumn, are treated similarly. These periods often share comparable temperature regimes despite occurring at different times of the year. The forecast temperature allows the model to account for biases that scale with the intensity of the predicted values, such as consistent overestimation during warmer periods. The inclusion of meridional wind reflects the idea that large-scale advective processes can modulate local temperature anomalies and contribute to regional forecast errors. This constitutes a deliberately simple, first-step approach aimed at testing whether physically interpretable variables can already explain part of the systematic forecast error through a purely linear relationship.



(a) Day of the year



(b) Temperature



(c) Meridional Wind

Figure 4.10: Regression coefficients of the MLR model for the predictors: **(a)** day of the year, **(b)** temperature and **(c)** meridional wind (v-component). The maps illustrate the spatial variability in the influence of each predictor on the bias, highlighting regionally distinct sensitivities in the model.

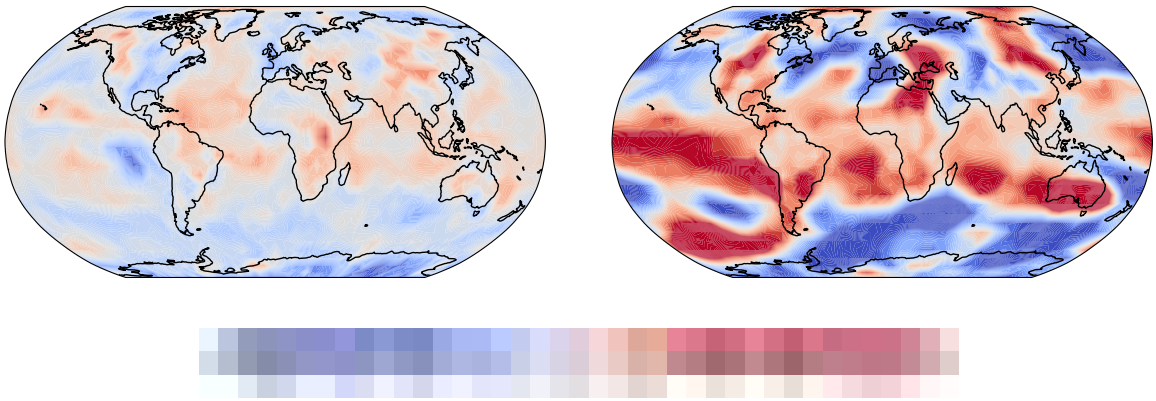
Figure 4.10 shows the spatial distribution of the regression coefficients for the three predictors. These coefficients represent how strongly each variable influences the local forecast bias and provide valuable insight into the underlying physical structures of the errors.

The coefficients for the day of the year are displayed in Figure 4.10a. A strong hemispheric pattern is evident. Coefficients are negative in most of the Northern Hemisphere and positive in the Southern Hemisphere. This implies that, on average, the bias becomes more negative in northern summer and more positive in southern summer. In other words, the model tends to overpredict temperatures during the summer months of each hemisphere. This highlights that a large component of the systematic error is linked to the annual cycle, and that a linear seasonal correction term can partially compensate for it. While the coefficients in the Southern Hemisphere are relatively uniform, larger values are observed in the Northern Hemisphere, in particular over land areas. This indicates stronger seasonal dependencies in these regions. Furthermore, the negative coefficients in the Northern Hemisphere tend to be substantially larger in magnitude than the positive ones in the Southern Hemisphere, suggesting a more pronounced summer overestimation bias in the north.

A clear pattern emerges in the coefficients for the forecast temperature, as shown in Figure 4.10b. In all regions, the coefficients are strictly positive, indicating a robust linear relationship between the forecasted temperature and the model bias. As the predicted temperature increases, the bias becomes more positive. This reflects a systematic tendency of Pangu-Weather to overestimate especially warm conditions. The strength of this effect varies regionally. The highest coefficients are found in the southern storm track regions, suggesting that temperature-dependent biases are particularly pronounced in dynamically active midlatitude environments of the Southern Hemisphere. In contrast, tropical and subtropical regions generally exhibit smaller coefficients despite their frequent occurrence of high absolute temperatures. This suggests that while warm conditions are common there, the temperature-dependent bias is less pronounced compared to midlatitude storm track areas. These results underscore that part of the systematic error scales with the magnitude of the predicted temperature.

The coefficients for the meridional wind, shown in Figure 4.10c, exhibit greater spatial variability and are generally weaker in magnitude compared to the other predictors. Most values are close to zero, indicating that the meridional wind has only a limited influence on the temperature bias in many regions. A particularly notable feature is the strong positive coefficient pattern over large parts of Central and Eastern Asia. This suggests that in these regions, increased poleward wind is strongly correlated with a warm bias in the forecast. However, this signal may not solely reflect atmospheric dynamics but could instead be related to unresolved orography in the model. In contrast, slightly negative coefficients are found in parts of the Southern Hemisphere subtropics. While the meridional wind is not the dominant predictor, its spatially varying influence contributes to regional refinement of the bias correction, especially in dynamically complex areas.

The impact of applying the correction model on a 24 h forecast with Pangu-Weather is illustrated in Figure 4.11. The raw forecast bias of the Pangu-Weather model for the year 2022, shown in Section 4.2.1, reveals relatively small magnitudes but pronounced and consistent spatial patterns. Following the application of the MLR-based correction (Section 4.2.1), the



(a) Bias in Pangu-Weather **(b)** Bias after correction with MLR

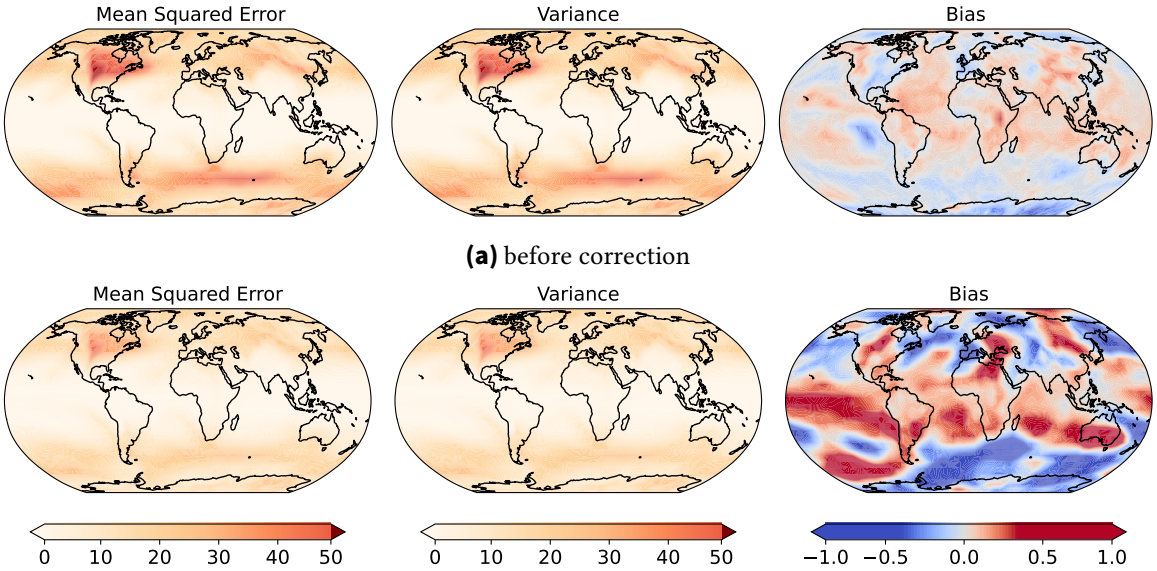
Figure 4.11: Spatial distribution of the 850-hPa temperature bias (in K) before and after correction using MLR. **(a)** shows the systematic bias in Pangu-Weather forecasts, while **(b)** illustrates the residual bias after the application of the MLR-based correction.

spatial bias structure undergoes substantial changes. The correction successfully reduces the global mean bias from −0.028 K to −0.011 K, which corresponds to a reduction of almost 61 %.This apparent improvement masks the fact that the spatial bias pattern has in many regions worsened. In particular, the correction tends to overcompensate in several areas, introducing new, often stronger biases.

This outcome illustrates a key limitation of the simple linear correction approach. Although it is capable of adjusting the global mean bias through large-scale, linear relationships, the model lacks the flexibility to adequately capture the complex, spatially heterogeneous structure of forecast errors. The emergence of pronounced local errors suggests possible overfitting to regional patterns in the training data or a limited generalization capability to the test year 2022. Therefore, despite the apparent improvement in the global mean metric, the correction quality remains questionable when evaluated in terms of spatial consistency and local reliability. These findings point to the necessity of more advanced correction methods that can incorporate nonlinear relationships or spatial dependencies to better constrain regional forecast errors.

To further assess the performance of the bias correction, the forecast errors were decomposed into MSE, variance and bias components, as shown in Figure 4.12. This decomposition is based on the error identity introduced in Equation (3.3), where the MSE is expressed as the sum of the variance and the squared bias. It provides deeper insight into the nature of the errors before and after applying the MLR-based correction. The top row shows the error characteristics of the uncorrected Pangu-Weather forecasts. The MSE is particularly high over North America and in the southern storm track region. The pattern in the MSE is largely driven by increased variance, as the values of the biases are comparably small.

After the correction (Figure 4.12b), the MSE is visibly reduced in many regions, especially where it was originally high. This reduction is primarily attributable to a decrease in the variance component. That is an indicator that the MLR model effectively smooths local



(a) before correction

(b) after correction with Multiple Linear Regression

Figure 4.12: Decomposition of the 850-hPa temperature forecast errors into MSE, variance and bias components before and after correction using multiple linear regression (MLR). **(a)** shows the decomposition in Pangu-Weather forecasts, while **(b)** illustrates the decomposition after the application of the MLR-based correction.

fluctuations and reduces the amplitude of the forecast errors. However, the bias component shows a more problematic behavior. Consistent with the spatial analysis discussed earlier, many regions develop new systematic biases or see pre-existing ones intensified. This confirms that, although the correction reduces random variability, it can introduce spatially coherent errors that affect the overall reliability of the forecast.

This behavior can be explained by the fact that classical MLR relies on the least squares criterion, which minimizes the average squared error across all training samples. As a consequence, the model tends to prioritize reducing the total MSE, which includes both variance and squared bias, but has no explicit incentive to minimize the bias alone. Instead, it balances bias and variance to reach the lowest overall MSE, even if this means introducing strong local biases in regions where variance can be reduced more effectively. As a result, the reduction in variance comes at the expense of increased biases in most areas. While this trade-off is mathematically optimal under the least squares loss, it is not necessarily desirable from a physical or forecast quality perspective, especially when spatial consistency is important.

4.2.2 XGBoost

Linear models are limited in their ability to capture the nonlinear structure of forecast errors. To address this, more flexible models were applied that can represent complex dependencies between variables. The focus lies on the correction of 850 hPa temperature forecasts using

models that differ in loss function, predictor configuration and regularization strategy. Nonlinear methods are particularly suited for this task as forecast errors may depend on interactions between temperature, wind, humidity and seasonal factors. These relationships are often not additive and cannot be fully described by linear combinations of predictors.

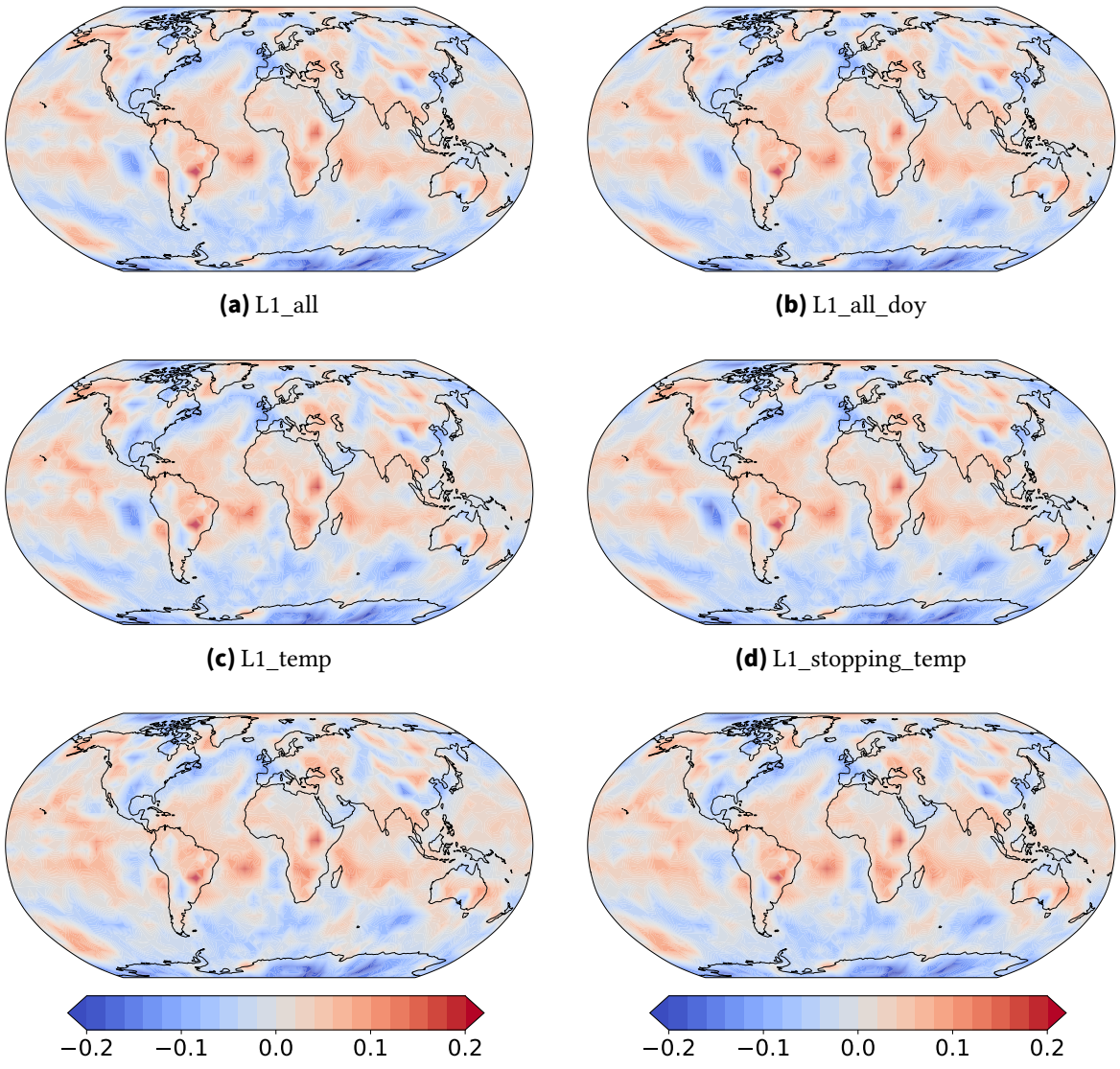
A total of six bias correction models were developed and are summarized in Table 4.1. Four models use the MAE (Equation (3.2)), also known as L1 loss, as a loss function. The MAE minimizes the average absolute difference between predictions and observations, making it more robust to outliers. Two models use the MSE, or L2 loss, which penalizes larger errors more strongly by squaring the differences. This often leads to smoother fits but potentially increased sensitivity to outliers. The L1-based models differ in their predictor sets and use of early stopping. L1_temp and L1_stopping_temp employ temperature as the sole predictor, without and with early stopping. L1_all adds the zonal and meridional wind components and specific humidity. L1_all_doy further includes the day of year (DOY) to represent seasonal effects. The two L2 models, L2_temp and L2_stopping_temp, use only temperature as input without and with early stopping. This model set allows a systematic comparison of loss functions, feature complexity and regularization strategies in correcting biases of the 850 hPa temperature forecasts with Pangu-Weather.

Model Name	**Loss Function**	**Predictors**	**Early Stopping**
L1_temp	MAE	temperature	No
L1_stopping_temp	MAE	temperature	Yes
L1_all	MAE	temperature, u- & v-wind, specific humidity	Yes
L1_all_doy	MAE	temperature, u- & v-wind, specific humidity, DOY	Yes
L2_temp	MSE	temperature	No
L2_stopping_temp	MSE	temperature	Yes

Table 4.1: Configuration of the six bias correction models developed for bias correction of 850 hPa temperature forecasts. The models differ in loss function (MAE and MSE), input features and the application of early stopping as a regularization strategy.

Early stopping (Prechelt, 1998) is a regularization method implemented during model training to avoid overfitting. The training process is terminated when the performance on an independent validation dataset no longer shows improvement. This prevents the model from excessively fitting noise or special features in the training data. This technique promotes better generalization to unseen data by selecting model parameters at the point of optimal validation performance.

The spatial pattern of the bias after the correction, using ERA5 as the reference dataset, is shown in Figure 4.13. In addition, the global mean biases and the bias reduction for each model are summarized in Table 4.2.



(a) L1_all **(b)** L1_all_doy

(c) L1_temp **(d)** L1_stopping_temp

(e) L2_temp **(f)** L2_stopping_temp

Figure 4.13: Spatial distribution of temperature forecast errors at 850 hPa for different bias correction methods and settings. Panels (a) to (f) show the error patterns for the following approaches: **(a)** linear model with L1 loss, using all predictors and early stopping (L1_all), **(b)** linear model with L1 loss, using all predictors plus day of year and early stopping (L1_all_doy), **(c)** linear model with L1 loss, using temperature only (L1_temp), **(d)** linear model with L1 loss, using temperature only and early stopping (L1_stopping_temp), **(e)** nonlinear model with L2 loss, using only temperature (L2_temp) and **(f)** nonlinear model with L2 loss, using only temperature and early stopping (L2_stopping_temp). ERA5 reanalysis data serve as the reference. The color scale indicates the magnitude and sign of the errors, with red representing positive biases and blue representing negative biases.

Model	Global Mean Bias (K)	Bias Reduction (%)
Uncorrected	−0.0280	
L1_temp	−0.0116	58.5
L1_stopping_temp	−0.0123	56.1
L1_all	−0.0123	56.3
L1_all_doy	−0.0119	57.4
L2_temp	−0.0096	65.6
L2_stopping_temp	−0.0098	65.0

Table 4.2: Global mean bias and relative bias reduction for each correction model, using ERA5 as reference.

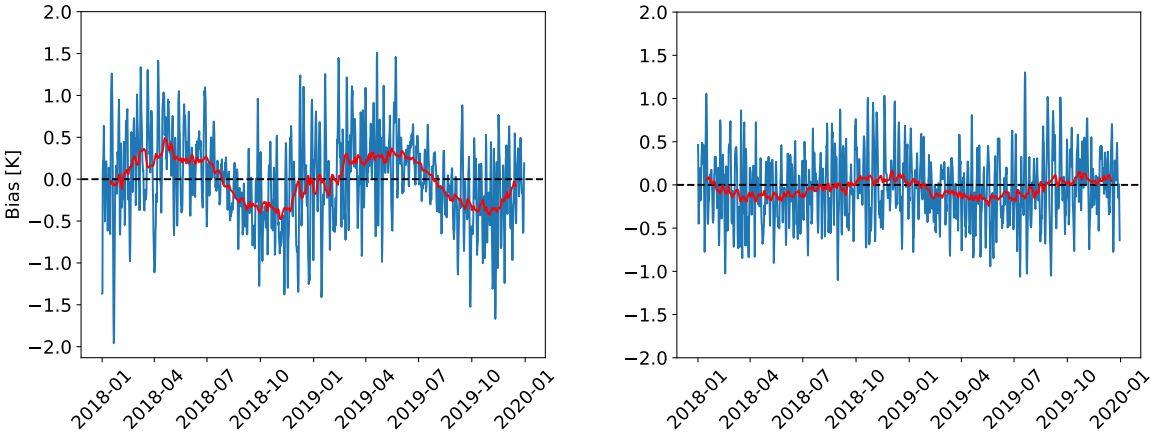
The results for the L1_all model, which uses temperature, horizontal wind components and specific humidity as predictors, are shown in Figure 4.13a. The global mean bias is reduced to -0.01226 K, corresponding to a bias reduction of 56.3 %. While a notable warm bias remains over northern South America, especially Brazil, the Sc-region shows reduced biases compared to the uncorrected data. Some cold biases emerge in the Southern Ocean and northern high latitudes. A particularly strong warm bias over East Asia, prominent in the uncorrected data, is substantially reduced and even turns slightly negative in some areas. An extension of the L1_all model through inclusion of the DOY as a seasonal predictor is shown in Figure 4.13b. This model achieves a slightly improved global mean bias of −0.01194 K, corresponding to a 57.4 % bias reduction. However, the overall spatial structure remains nearly identical to that of Figure 4.13a, suggesting that incorporating seasonal information provides limited added value. Residual bias patterns, particularly over Brazil and East Asia, remain largely unchanged.

A simpler approach is represented by the L1_temp model in Figure 4.13c, which relies solely on temperature as predictor and omits early stopping. Despite its reduced complexity, the global bias improves slightly to −0.01163 K, which corresponds to a reduction of 58.5 %. This is marginally outperforming the more elaborate variants. The spatial distribution of biases closely resembles those of the previous models, with somewhat stronger cold biases over the Sc-region. The impact of early stopping is evaluated in Figure 4.13d, which shows the results for the L1_stopping_temp model. Using the same predictor as L1_temp, it achieves a slightly lower performance with a global bias of −0.01229 K, corresponding to a bias reduction of 56.1 %. This shows a small performance drop compared to the variant without stopping. Nevertheless, residual biases are slightly reduced in magnitude over oceanic regions, suggesting a beneficial regularization effect on spatial consistency.

Figure 4.13e corresponds to the L2_temp model, based on the MSE loss and trained without early stopping. This model achieves a global mean bias of −0.00964 K, the lowest among all tested models, with a 65.6 % bias reduction. However, the spatial structure of the bias is mostly the same as in the other models except of the Sc-region. There, the bias is strongly reduced. Finally, Figure 4.13f illustrates the L2_stopping_temp model, which combines the MSE loss with only the temperature as predictor and early stopping. The global mean bias is −0.00980 K, corresponding to a 65.0 % reduction. While performance is slightly worse

than in Figure 4.13e, the spatial pattern is similar. As in all models, a persistent warm bias remains over Brazil.

The observation that the bias remains larger in models trained with early stopping compared to those trained without warrants careful consideration. Several factors may contribute to this phenomenon. Early stopping may cause the training process to terminate prematurely before the model has sufficiently learned to reduce the bias, resulting in underfitting. Additionally, model complexity and training parameters such as learning rate may interact with early stopping. A model that is too simple or trained with an unsuitable learning rate might benefit from longer training without early stopping to better fit the data and reduce bias. Overall, tuning the early stopping parameters, ensuring a representative validation set and selecting an appropriate metric are essential to achieving optimal bias reduction in predictive models.



(a) Northern Hemisphere **(b)** Southern Hemisphere

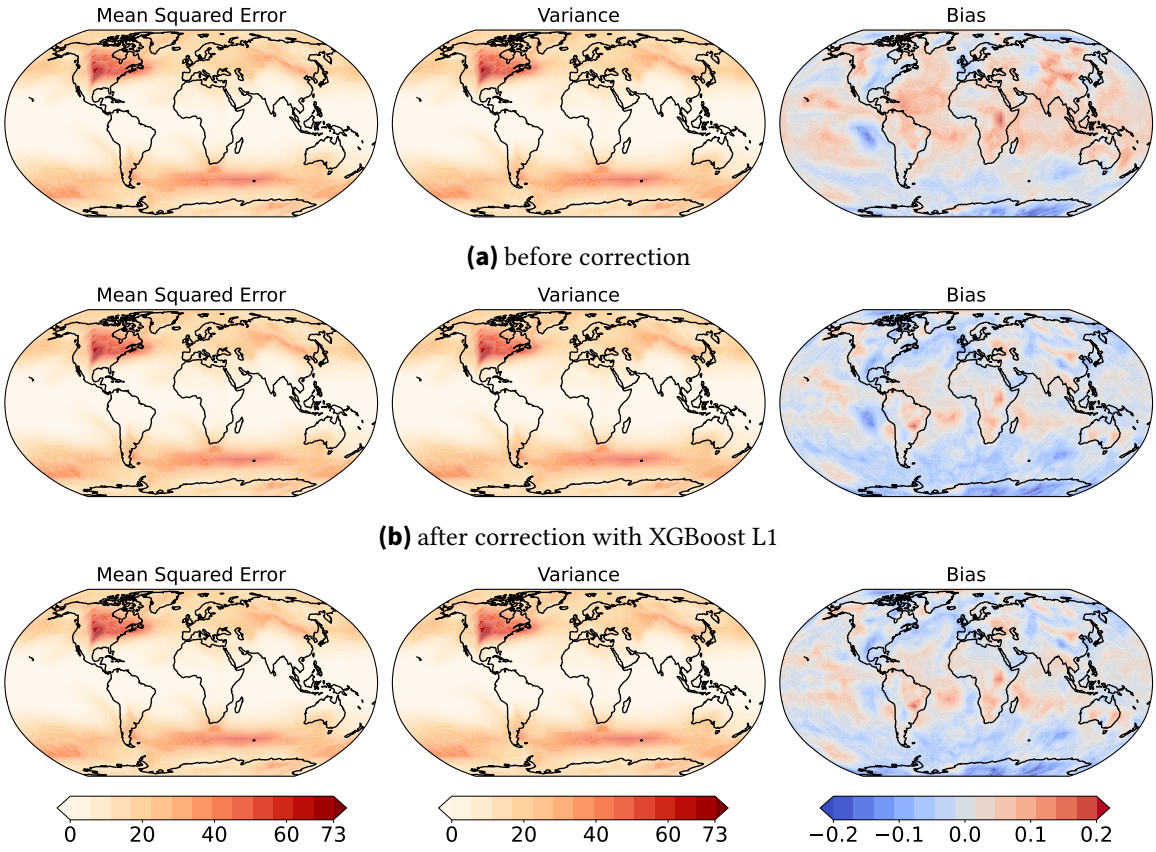
Figure 4.14: Time series of temperature bias at 850 hPa (in K), shown separately for the Northern and Southern Midlatitudes. Blue lines represent the bias at each forecast time step, while the red lines indicate the 14-day running mean to highlight seasonal patterns.

The DOY variable is often considered a useful predictor in weather-related models because it can capture seasonal cycles and temporal patterns. However, in the context of our bias correction models, the inclusion of DOY does not significantly improve performance. There are several reasons why DOY may not be as relevant and why the model is still able to learn effectively without it.

To better understand the limited contribution of DOY, the temporal evolution of the 850 hPa temperature bias was analyzed separately for the Northern Hemisphere (NH) and Southern Hemisphere (SH) midlatitudes. A 14-day running mean was applied to spatially averaged bias values to highlight seasonal patterns. The results are shown in Figure 4.14. In the NH, a pronounced annual cycle is evident, with bias peaking around May and reaching a minimum near October. This pattern suggests that forecasted temperatures are systematically overestimated during the warm season and underestimated during the colder months. In the SH, however, the seasonal cycle is notably weaker, with a phase shift

relative to the NH. The observed asymmetry between hemispheres may be attributed to the unequal distribution of land and ocean. The NH, dominated by landmasses, exhibits stronger seasonal variability. In contrast, the SH, with its extensive ocean coverage, displays more subdued thermal responses.

The presence of a clearly defined seasonal cycle in the bias, in particular in the NH, indicates that much of the seasonal information is already embedded in the systematic forecast errors. As a result, DOY provides limited additional predictive value, since the model is capable of implicitly learning the seasonal variation directly from the structure of the temperature bias itself. This explains the negligible differences in performance between models trained with and without DOY as a predictor.



(a) before correction

(b) after correction with XGBoost L1

(c) after correction with XGBoost L2

Figure 4.15: Decomposition of the 850-hPa temperature forecast errors into MSE, variance and bias components before and after correction using XGBoost. **(a)** shows the decomposition in Pangu-Weather forecasts, while **(b)** illustrates the decomposition after the application of the correction using XGBoost with L1 loss (L1_stopping_temp). **(c)** shows the decomposition after the applicatio of the correction using XGboost with L2 loss (L2_stopping_temp).

To investigate the differences between the models trained with L1 and L2 loss functions, the decomposition of the MSE into its variance and bias components (Equation (3.3)) are analyzed. The decomposition in Figure 4.15 reveals that both models exhibit similar patterns and spatial structures in terms of MSE and variance when compared to the uncorrected

forecasts. This suggests that the variability of the predictions remains largely unchanged by the choice of loss function. In contrast, the bias component differs between the two models, indicating that the primary effect of the correction lies in bias reduction. Since the primary goal of the correction process is to minimize bias, this outcome is of particular significance.

It is important to note that the L2 loss implicitly combines bias and variance in a single objective function without explicitly allowing control over which component to reduce. Therefore, one could expect the correction to affect both bias and variance. However, the results show that variance remains stable while bias decreases, implying that the model primarily focuses on correcting systematic errors.

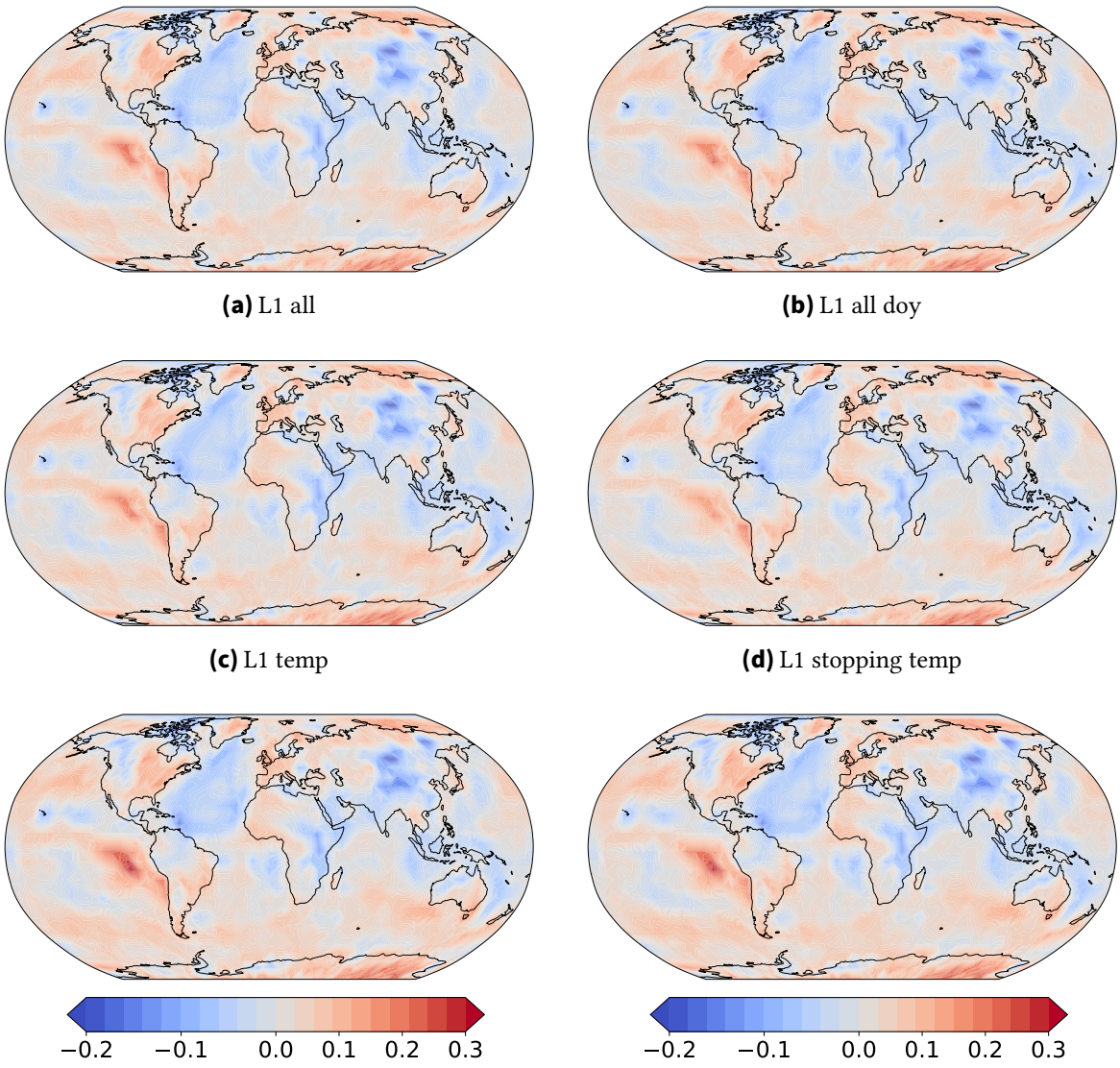
The question arises why the model trained with L2 loss still performs better than the one trained with L1 loss despite this similarity in variance patterns. One explanation is that the L2 loss penalizes larger errors more strongly due to its quadratic nature, which effectively drives the model to reduce larger bias components more aggressively. In contrast, the L1 loss treats all errors linearly, which may lead to less emphasis on correcting larger systematic deviations. Consequently, the L2-trained model achieves better overall error reduction even though it does not explicitly distinguish between variance and bias in the optimization process.

To evaluate the effect of the bias correction models relative to the uncorrected Pangu-Weather forecasts, difference plots of corrected minus uncorrected predictions were generated for each model. The results are shown in Figure 4.16. Across all models, a consistent pattern emerges. Regions that were originally too cold tend to be adjusted towards warmer temperatures, while regions that were initially too warm are corrected towards cooler values. This indicates that the bias correction effectively counteracts systematic temperature deviations in the uncorrected forecasts. Notably, the area over Brazil shows a distinct behavior where all correction models predict slightly higher temperatures compared to the uncorrected Pangu-Weather data.

In addition to the evaluation of correction performance, computational efficiency plays a crucial role in model selection, especially when considering operational applications. It is well known that the training time of statistical correction models increases with the number of predictors employed. This is due to the higher dimensionality of the input space, which requires more complex calculations and longer optimization procedures. Therefore, models incorporating a larger set of predictors consequently demand substantially longer training times.

Furthermore, the implementation of early stopping has a significant impact on training duration. Models trained with early stopping converge more quickly because the training process terminates once the validation metric no longer improves, thereby preventing unnecessary iterations. This mechanism not only reduces the overall training time but also helps to avoid overfitting. In contrast, models trained without early stopping continue to iterate for a fixed number of epochs, often resulting in longer training times.

The computational demands also extend to the application phase, where the time required for bias correction of temperature varies in accordance with model complexity and the



(a) L1 all **(b)** L1 all doy

(c) L1 temp **(d)** L1 stopping temp

(e) L2 temp **(f)** L2 stopping temp

Figure 4.16: Spatial distribution of temperature forecast errors at 850 hPa for different bias correction methods and settings. Panels (a) to (f) show the error patterns for the following approaches: **(a)** linear model with L1 loss, using all predictors and early stopping (L1_all), **(b)** linear model with L1 loss, using all predictors plus day of year and early stopping (L1_all_doy), **(c)** linear model with L1 loss, using temperature only (L1_temp), **(d)** linear model with L1 loss, using temperature only and early stopping (L1_stopping_temp), **(e)** nonlinear model with L2 loss, using only temperature (L2_temp) and **(f)** nonlinear model with L2 loss, using only temperature and early stopping (L2_stopping_temp). Uncorrected Pangu-Weather data from WB2 serve as the reference. The color scale indicates the magnitude and sign of the errors, with red representing positive biases and blue representing negative biases.

number of predictors used. Therefore, the selection of an appropriate model requires a balanced consideration of both correction quality and computational cost. Given that the performance differences among the tested models are relatively minor in this analysis, the L1_stopping_temp model is adopted for further use. This choice reflects a compromise that optimizes training efficiency while maintaining satisfactory correction accuracy.

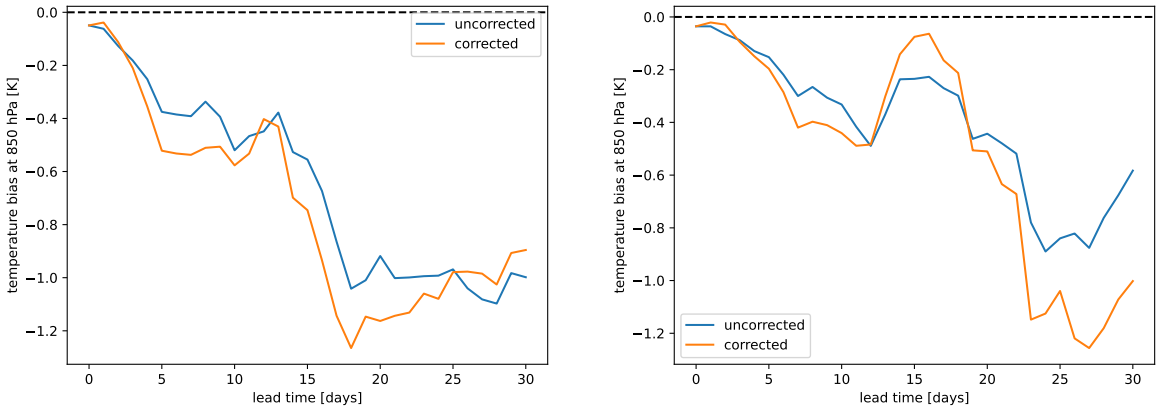
4.3 Online Bias Correction

In the framework of online bias correction, the model L1_stopping_temp is integrated into the Pangu-Weather forecast system. The correction is applied iteratively every 24 hours, using the most recent forecast to update the prediction. This section evaluates the performance of the approach using two case studies of extreme precipitation events in California, one of which is the high-impact event in December, characterized by a sequence of atmospheric rivers that brought exceptional rainfall to the region (DeFlorio et al., 2023). For both cases, a 10-member ensemble is used with lead times up to 30 days, and the evaluation is based on the ensemble mean. The ensemble members are based on the Ensemble of Data Assimilations (EDA) of the ECMWF (Isaksen et al., 2010), which represents uncertainty in the initial conditions by perturbing observations and model physics during the data assimilation process. The iterative correction is computationally efficient, with each update step requiring roughly 30 seconds to execute on the bwUniCluster 3.0. This enables integration into extended forecast chains without introducing substantial computational cost.

Figure 4.17 illustrates the results of the online bias correction for both case studies. Figure 4.17a shows the temperature bias at 850 hPa for a case in December 2022, while Figure 4.17b corresponds to a case in February 2023. The temperature bias is defined as the difference between the Pangu-Weather ensemble mean forecast and the ERA5 reanalysis, with negative values indicating a cold bias. For each lead time (in days), the bias is shown for both the uncorrected forecast (blue line) and the corrected forecast (orange line), allowing for a direct comparison of the two approaches.

In both case studies, the bias after 24 hours is clearly closer to zero in the corrected forecast than in the uncorrected one. This aligns well with the results from the offline bias correction and confirms that the statistical correction model performs reliably at the early stage of the forecast. The improved proximity to zero indicates that the correction successfully removes a significant portion of the systematic error after just one update cycle.

As the forecast progresses, however, the difference between the corrected and uncorrected forecasts becomes more variable. At certain lead times, the corrected forecast still shows a bias closer to zero, while at others, the bias is slightly more negative than in the uncorrected version. This indicates that the effectiveness of the online correction depends on the evolving atmospheric state and potentially on accumulated forecast errors that are harder to correct using a static correction model.



(a) December case

(b) February Case

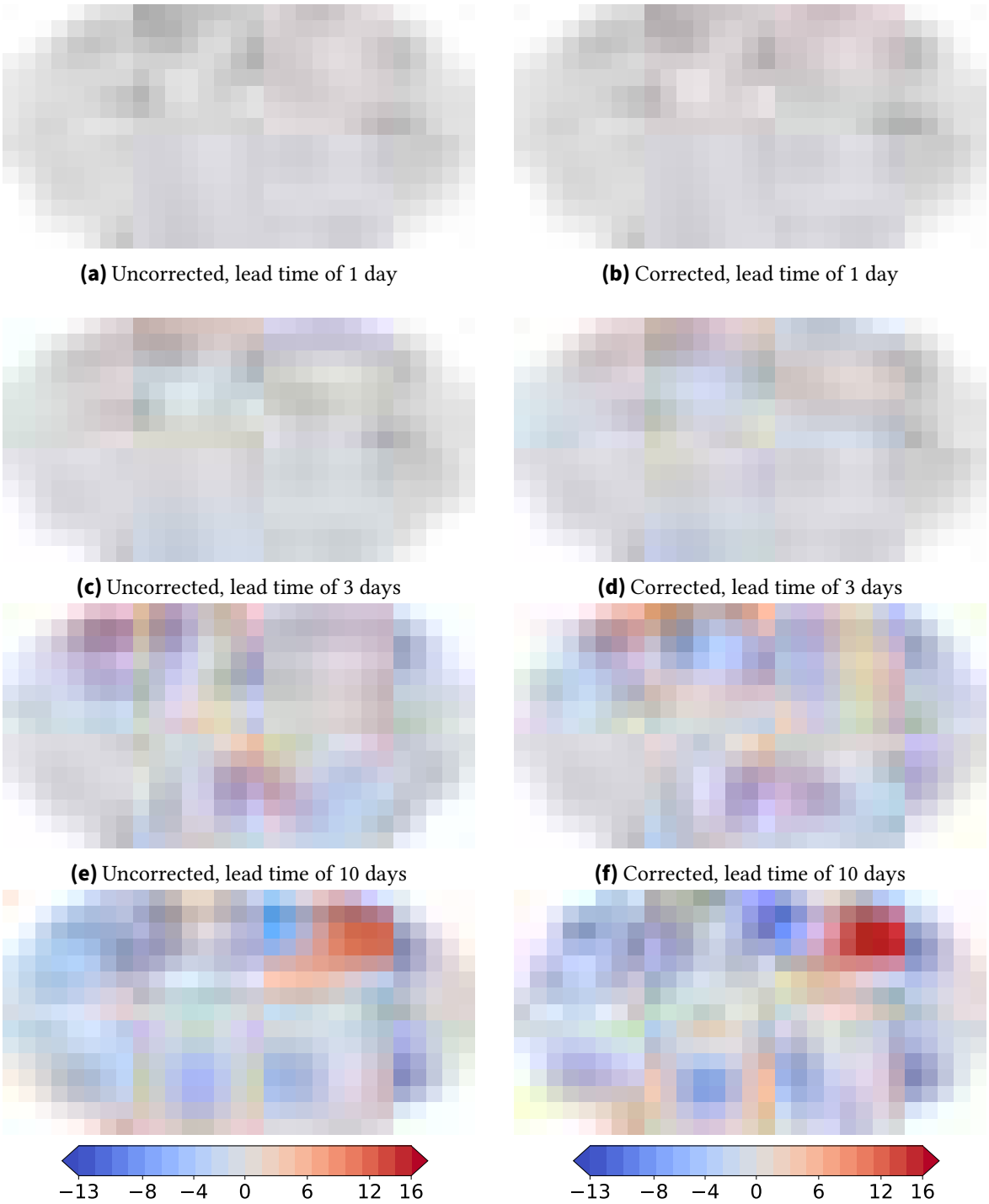
Figure 4.17: Time series of temperature bias at 850 hPa over a 30-day lead time for two case studies: **(a)** initialized at 15 December 2022 and **(b)** initialized at 02 February 2023. The comparison shows the evolution of uncorrected (blue) and online bias corrected (orange) forecasts, highlighting the effectiveness of the bias correction method in reducing systematic errors over time.

In the December case (Figure 4.17a), which was initialized on 15 December 2022, the bias of the uncorrected forecast steadily increases in magnitude, reaching a peak around day 18. This indicates a growing systematic error in the raw Pangu-Weather forecast as lead time increases. This behavior is consistent with the findings from Section 4.1, where the evaluation showed that the bias in Pangu-Weather forecasts tends to grow with lead time. It also aligns with results from Bouallègue et al. (2024). The corrected forecast shows a slight reduction in bias around day 13, suggesting a temporary improvement due to the application of the online correction. However, in two longer phases between days 2 and 12, and again from day 14 to day 24, the corrected forecast consistently exhibits a more negative bias compared to the uncorrected version.

Despite this, a clear improvement becomes evident after day 25. The bias of the corrected forecast decreases substantially in magnitude and moves significantly closer to zero. The temporal variation in correction effectiveness highlights the dynamic nature of forecast errors and suggests that the static correction model may struggle to fully capture the evolving structure of biases beyond the short range.

In the February case (Figure 4.17b), initialized on 2 February 2023, the uncorrected forecast generally exhibits a clear negative bias trend over the lead time. However, the corrected forecast shows especially between days 12 and 19 a significantly improvement compared to the uncorrected forecast. Nevertheless, outside of this interval, the corrected forecast show more negative biases compared to the uncorrected forecast.

To assess the spatial characteristics of the forecast bias and the impact of the applied bias correction method, Figure 4.18 displays two-column maps for the December case study at forecast lead times of 1, 3, 10 and 30 days. Each row corresponds to one lead time, with the left column showing the uncorrected bias and the right column showing the bias after applying the online correction. All maps focus on the 850 hPa temperature field.



(a) Uncorrected, lead time of 1 day

(b) Corrected, lead time of 1 day

(c) Uncorrected, lead time of 3 days

(d) Corrected, lead time of 3 days

(e) Uncorrected, lead time of 10 days

(f) Corrected, lead time of 10 days

(g) Uncorrected, lead time of 30 days

(h) Corrected, lead time of 30 days

Figure 4.18: Spatial distribution of forecast bias of the 850-hPa temperature at different lead times for the December case. The left column shows the bias of the uncorrected Pangu-Weather forecast, the right column shows the bias of the online bias corrected forecast. **(a)** and **(b)** show the bias for a lead time of 1 day, **(c)** and **(d)** for a lead time of 3 days, **(e)** and **(f)** for a lead time of 10 days and **(g)** and **(h)** for a lead time of 30 days.

At a 1-day lead time (Figures 4.18a and 4.18b), the overall forecast bias is small in magnitude across most regions. Accordingly, the differences between the uncorrected and corrected fields are minor. Nevertheless, some subtle structures emerge in the corrected version that are absent or less pronounced in the original forecast. In particular, the southern midlatitudes exhibit localized regions of slightly more negative bias after correction. Additionally, a region of slightly increased positive bias develops over Central Asia. This feature appears only after correction and may again be related to topographic influences, as previously discussed. The persistent challenges in this region underscore the difficulties the correction model faces in areas with complex orography.

After 3 days (Figures 4.18c and 4.18d), the overall bias remains modest in magnitude, but spatial patterns become more distinct. The correction introduces a notable area of enhanced negative bias over Iceland. Meanwhile, the persistent positive bias over the Central Asia continues to stand out, with little change in intensity compared to the 1-day lead time.

At a 10-day lead time (Figures 4.18e and 4.18f), forecast errors become more pronounced and the benefits and limitations of the correction approach become clearer. In the uncorrected forecast, several regions exhibit strong biases. Pronounced positive biases appear over Alaska and South Africa, as well as in adjacent areas to the south. Conversely, large-scale negative biases are observed over the western United States, northern Africa, and parts of the southern midlatitudes. After correction, the southern hemisphere bias structures remain relatively similar in both extent and intensity, indicating that the correction has limited impact in those areas. However, a new region of negative bias appears over Australia. The positive bias over Alaska is considerably reduced in the corrected version, highlighting a successful adjustment by the model in this region. In contrast, the negative bias over the western United States becomes even more pronounced after correction, suggesting a potential overcompensation or misestimation by the model in that region.

At a 30-day lead time (Figures 4.18g and 4.18h), the forecast bias becomes very pronounced, reflecting the considerable loss of skill at longer forecast horizons. Large-scale and spatially coherent biases dominate the midlatitudes, while the tropics remain comparatively less affected. In the uncorrected forecast, an extensive region of strong positive bias spans much of Asia. Conversely, a marked negative bias emerges over North America, particularly affecting the continental interior. These patterns are consistent with the expected increase in dynamical model uncertainty at longer lead times, especially in regions with strong synoptic variability.

In the corrected forecast, several notable differences appear. A more negative bias develops north of Europe. Over Asia, the positive bias is not only retained but appears even more pronounced compared to the uncorrected version. This intensified warm bias in the corrected field can be seen as a continuation of the smaller anomaly already present at shorter lead times. Its consistent amplification over time suggests that the correction model introduces or reinforces a systematic regional bias in this area. As discussed previously, the complex topography of the Himalayan region likely contributes to this behavior. The correction model may struggle to represent such terrain-induced biases accurately, especially if these effects are underrepresented or smoothed out in the training data.

To further investigate the spatial characteristics of the forecast bias, Figure 4.19 presents the results for the February case. As before, spatial maps are shown for lead times of 1, 3, 10, and 30 days, with the uncorrected forecast on the left and the bias-corrected version on the right.

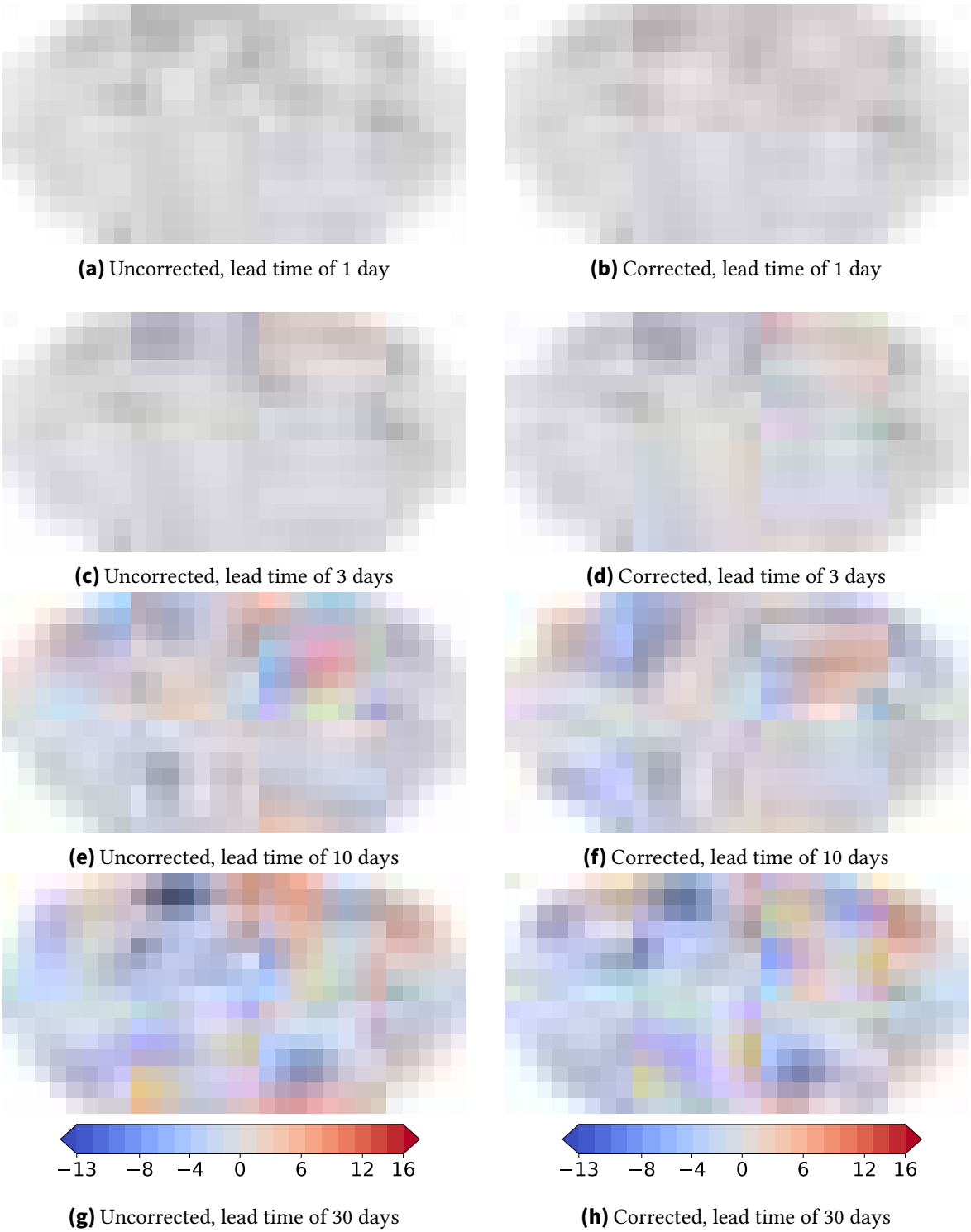
At a 1-day lead time (Figures 4.19a and 4.19b), the bias magnitude remains generally low, similar to the December case. Both uncorrected and corrected forecasts exhibit weak, spatially incoherent anomalies. The correction introduces minor local changes, particularly in the southern midlatitudes, where a slight negative tendency emerges. Again, a small positive anomaly emerges over Central Asia in the corrected forecast. This suggests that the correction model consistently reinforces a warm bias in this area at medium lead times, possibly due to persistent issues related to orography, as previously discussed. These repeated structures across seasons point toward a systematic limitation of the correction approach in handling terrain-induced biases.

At 3 days lead time (Figures 4.19c and 4.19d), biases become slightly more pronounced. In the corrected field, a clear negative bias develops over the North Atlantic, especially around Iceland. Over Canada, a positive bias emerges, which, however, is noticeably reduced in the corrected version.

After 10 days (Figures 4.19e and 4.19f), the bias fields become more coherent. The uncorrected forecast shows positive biases over Canada and southern Africa, along with negative biases over the western U.S., the tropical Atlantic, and parts of East Asia. The corrected forecast partially reduces the positive bias over Canada and introduces a more defined warm anomaly over Central Asia, particularly along the Himalayan arc—again a direct parallel to the December case. This recurring pattern reinforces the notion that the correction model may struggle to resolve biases in regions with complex topography, potentially due to limited representation in the training data or insufficient spatial generalization.

At a 30-day lead time (Figures 4.19g and 4.19h), the forecast biases reach their highest magnitude. As in December, the most pronounced anomalies are found in the northern and southern midlatitudes, while tropical regions remain relatively unbiased. The uncorrected field shows a strong positive bias across central and eastern Asia and a broad negative bias over eastern North America, which are both features already present at shorter lead times. In the corrected version, the positive bias over Asia intensifies further, once again peaking in the Himalayan region, thereby continuing the structure that emerged at days 3 and 10. This consistent growth of the warm anomaly in the corrected field mirrors the December development and suggests a structural issue in the correction process. In contrast, the cold bias over North America is slightly weakened, indicating some success of the correction model in that region.

A closer inspection of the Sc region over the eastern Pacific Ocean, which previously exhibits a relatively large cold bias, reveals a notable improvement in the December case at longer lead times. Specifically, at day 30, the corrected forecast shows a substantial reduction of the bias in this region, suggesting that the correction model can successfully mitigate persistent systematic errors in certain oceanic regimes. In contrast, the February case does not display a similar improvement. However, it is important to consider that both case



(a) Uncorrected, lead time of 1 day

(b) Corrected, lead time of 1 day

(c) Uncorrected, lead time of 3 days

(d) Corrected, lead time of 3 days

(e) Uncorrected, lead time of 10 days

(f) Corrected, lead time of 10 days

(g) Uncorrected, lead time of 30 days

(h) Corrected, lead time of 30 days

Figure 4.19: Spatial distribution of forecast bias of the 850-hPa temperature at different lead times for the February case. The left column shows the bias of the uncorrected Pangu-Weather forecast, the right column shows the bias of the online bias corrected forecast. **(a)** and **(b)** show the bias for a lead time of 1 day, **(c)** and **(d)** for a lead time of 3 days, **(e)** and **(f)** for a lead time of 10 days and **(g)** and **(h)** for a lead time of 30 days.

studies fall within the DJF season, during which the mean bias in the Sc region has already been found to be relatively small in the earlier seasonal analysis. This limits the potential for visible improvements through bias correction and highlights the need to interpret case study results in the context of broader climatological conditions.

The results of the online bias correction applied within the Pangu-Weather forecasting system show a complex and spatially heterogeneous impact on the temperature bias at 850 hPa. On a global scale, a reduction of the bias is evident at certain forecast lead times, which demonstrates the potential of the correction method to improve overall forecast accuracy. However, this apparent global improvement is not uniform and is accompanied by the emergence or amplification of strong positive biases in particular regions. In some areas, particularly those with complex topography such as the Himalayan region, the correction model tends to reinforce or even increase warm biases. These persistent anomalies suggest structural limitations of the static correction approach, which may stem from insufficient representation of orographic effects in the training data or limited spatial generalization capabilities of the statistical model. Conversely, at several specific locations, the bias correction achieves a reduction of systematic errors, even if these improvements are often relatively small in magnitude. These localized positive effects confirm that the correction approach can successfully mitigate forecast biases under certain atmospheric conditions or in less complex terrain.

5 Conclusion and outlook

The increasing use of MLWP models such as Pangu-Weather introduces new challenges for weather forecasting. Though these models offer significant reductions of computational costs and improvements global forecast skill compared to NWP models, they are not free from systematic errors. These biases, which represent persistent and non-random deviations from reference data, can compromise the reliability of forecasts, in particular in sensitive applications such as climate monitoring, energy planning or early warning systems. The present thesis systematically investigates the temperature bias at the 850 hPa level in forecasts from Pangu-Weather compared to ERA5 reanalysis data, with the aim of both understanding the origin and structure of the bias and developing suitable offline and online correction strategies.

The first part of the work focuses on analyzing the spatial and temporal characteristics of the temperature bias, with particular attention paid to seasonal variability and physical plausibility. In the second part, various statistical and machine learning models are implemented and tested in an offline and online correction framework. The correction models are designed to operate point-wise for each grid cell and forecast initialization time, making them adaptable for operational use. Among the models tested are MLR and several configurations of XGBoost, trained with different loss functions and stopping criteria. These models are evaluated not only in terms of RMSE reduction but also with a particular emphasis on bias minimization.

The following section addresses the research questions posed in the introduction:

1. What are the underlying causes of biases in Pangu-Weather forecasts and how are they related to atmospheric variables and model characteristics?

   The analysis demonstrates that Pangu-Weather exhibits a generally negative temperature bias at 850 hPa, which tends to increase towards longer lead times. This negative bias is widespread and largely global in nature, with only a few regions showing weak positive biases. These exceptions can often be attributed to unresolved orographic features that are smoothed out in the model's representation of the Earth's surface. Furthermore, particularly pronounced cold biases are identified in stratocumulus-dominated regions, such as the eastern Pacific and the western Atlantic. These biases are likely related to unresolved cloud microphysics and an overestimation of cloud cover by the model, pointing to structural limitations in the physical representation within the forecasting system. Additionally, the temperature profiles in these regions exhibit much weaker inversions than would typically be expected, which further suggests deficiencies in the model's ability to realistically represent boundary layer

processes and the vertical structure of the atmosphere. The seasonal analysis reveals that the magnitude and distribution of the bias are not static. Certain regions, such as the eastern Pacific Ocean, exhibit significantly stronger biases in boreal summer compared to winter, suggesting seasonally modulated errors likely tied to differences in surface-atmosphere coupling and cloud dynamics.

2. What methods can be applied to correct the bias in Pangu-Weather effectively?

   Among the correction methods tested in this thesis, the MLR model proves to be insufficient for the main objective of bias correction. Although it is able to reduce the RMSE to some extent, it fails to significantly correct the systematic error itself. This indicates that MLR may improve the overall variance-based accuracy of the forecasts, but does not effectively address the mean offset that defines the bias. In contrast, machine learning approaches based on the XGBoost algorithm perform considerably better. Several model variants are evaluated, each differing in terms of input features, loss functions and stopping criteria.

   The model variant referred to as L2_temp, which is trained using the L2 loss, temperature as the only predictor and without an early stopping criterion, achieves the largest reduction in mean bias. Specifically, it lowers the global mean bias after 24 hours lead time from -0.028 K to -0.0096 K, which corresponds to a relative improvement by approximately 65.6%. However, models optimized using the L2 loss function do not explicitly target the MAE and can therefore fail to consistently minimize the bias across different situations. This is because the L2 loss gives disproportionate weight to large individual errors, which may distract the model from correcting smaller but systematic deviations.

   Models trained with the L1 loss function, on the other hand, are more robust with respect to outliers and inherently focus more on reducing the median error, which is more closely related to the bias in many cases. The L1_stopping_temp model, despite showing comparatively weaker results in standard evaluation metrics such as RMSE and MSE, offers a more consistent reduction of the bias across different regions and seasons. Furthermore, it requires less computational effort, which makes it a promising candidate for operational implementation. Taken together, the results suggest that while L2-based models may achieve stronger numerical performance in specific cases, L1-based models provide a more reliable and generalizable solution for bias correction in temperature forecasts.

3. To what extent does bias correction improve medium-range weather predictions and how does it impact forecast accuracy?

   The evaluation of the online bias correction applied to Pangu-Weather forecasts reveals a somewhat inconclusive picture regarding its effectiveness in improving medium-range temperature predictions at 850 hPa. While the correction model shows clear benefits in reducing the temperature bias shortly after forecast initialization, in particular within the first 24 hours, these improvements are generally not sustained over longer lead times. In two test cases, the bias correction fails to reliably reduce systematic errors.

The results indicate that the static correction approach, which applies a fixed adjustment learned from past data, struggles to capture the evolving and complex nature of forecast errors, especially as lead times increase and atmospheric states become more uncertain. This limitation leads to variable performance. Some time periods and spatial regions experience minor bias reductions, whereas others show negligible changes or even degradation compared to the uncorrected forecasts. A spatial analysis further confirms this heterogeneous behavior. In areas with complex terrain or pronounced synoptic variability, such as the Himalayan region or parts of North America, the correction model sometimes exacerbates existing biases or introduces systematic warm or cold anomalies not present in the original forecast. These artifacts point to structural deficiencies in the correction model, possibly stemming from insufficient representation of orographic influences and dynamic atmospheric processes in the training data.

Overall, the findings suggest that while online bias correction holds promise for partially mitigating systematic errors in MLWP, the current implementation is not yet capable of providing consistent and robust bias reduction across all lead times and regions. The limited and sometimes counterproductive effects observed highlight the need for more advanced, adaptive correction methods that can dynamically respond to changing forecast conditions and incorporate physical constraints to ensure realistic adjustments.

While this thesis focuses on the online bias correction of temperature forecasts at the 850 hPa level, it is important to emphasize that temperature is not the only variable subject to systematic error in MLWP systems. In particular, the geopotential height is closely linked to temperature through the hypsometric equation and plays a critical role in large-scale circulation patterns and dynamical diagnostics. Future work should therefore extend the correction framework to include geopotential height in a joint modeling approach. This would not only improve the physical consistency of the corrected forecasts, but also enhance their usefulness in downstream applications. In particular, more accurate and physically consistent forecasts can contribute to more reliable early warning systems, which depend on robust detection of atmospheric patterns associated with high-impact weather events. Improved representation of geopotential fields can help better anticipate the development and movement of synoptic-scale systems, thereby supporting timely and informed decision-making in disaster risk management and public safety planning.

Another important direction for future research concerns the temporal robustness and generalizability of the online correction models. In this thesis, the evaluation of the correction performance is limited to two isolated extreme weather events. Though this setting allows for testing the models under high-impact conditions, it does not reflect the full range of synoptic variability present in typical forecast scenarios. In particular, it remains unclear how well the correction models perform under more moderate or climatologically neutral conditions. To address this limitation, future evaluations should be conducted over longer and more representative time periods, ideally covering a full calendar year as done in the bias characterization stage of this thesis. Such extended testing would allow for a

more comprehensive assessment of seasonal dependencies, model stability, and operational applicability.

Furthermore, previous studies demonstrate that online bias correction can be effective in traditional NWP systems. For instance, Watt-Meyer et al. (2021) show that online learning techniques can improve forecast accuracy by dynamically adapting to evolving error characteristics in ensemble prediction systems. These findings highlight the potential of online correction methods when integrated into physics-based models. In contrast, the comparatively limited effectiveness observed in this thesis for Pangu-Weather raises the question of whether this is a general limitation of MLWP systems or a model-specific shortcoming. Clarifying this issue will require further comparative analyses across different machine learning–based forecasting systems and correction approaches. Such investigations are essential to evaluate the general applicability and performance of online bias correction in data-driven forecasting frameworks.

Overall, the results of this thesis provide a solid foundation for further development of bias correction techniques in MLWP. Incorporating additional variables, extending the temporal scope of evaluation and ensuring physical consistency between corrected fields represent the next key steps toward operational applicability and scientific reliability.

Bibliography

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

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, 2021: Enforcing Analytic Constraints in Neural-Networks Emulating Physical Systems. arXiv, arXiv:1909.00912, DOI: https://doi.org/10.48550/arXiv.1909.00912.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. arXiv, DOI: https://doi.org/10.48550/arXiv.2211.02556.

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, DOI: https://doi.org/10.1038/s41586-023-06185-3, publisher: Nature Publishing Group.

Bjerknes, V., 1904: *Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik*. No. 21, Meteorologische Zeitschrift, Ed. Hölzel.

Bouallègue, Z. B., and Coauthors, 2024: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, **105 (6)**, E864–E883, DOI: https://doi.org/10.1175/BAMS-D-23-0162.1.

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

Chattopadhyay, A., E. Nabizadeh, and P. Hassanzadeh, 2020: Analog Forecasting of Extreme-Causing Weather Patterns Using Deep Learning. *Journal of Advances in Modeling Earth Systems*, **12 (2)**, e2019MS001 958, DOI: https://doi.org/10.1029/2019MS001958.

Chen, T., and C. Guestrin, 2016: XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794, DOI: https://doi.org/10.1145/2939672.2939785.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137 (656)**, 553–597, DOI: https://doi.org/10.1002/qj.828.

DeFlorio, M. J., and Coauthors, 2023: From California's extreme drought to major flooding: Evaluating and synthesizing experimental seasonal and subseasonal forecasts of

landfalling atmospheric rivers and extreme precipitation during Winter 2022 - 2023. URL https://repository.library.noaa.gov/view/noaa/60219.

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, DOI: https://doi.org/10.5194/gmd-11-3999-2018.

Díaz, N., M. Barreiro, and N. Rubido, 2023: Data driven models of the Madden-Julian Oscillation: understanding its evolution and ENSO modulation. *npj Climate and Atmospheric Science*, **6 (1)**, 1–12, DOI: https://doi.org/10.1038/s41612-023-00527-8.

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29 (5)**, 1189–1232, DOI: https://doi.org/10.1214/aos/1013203451, publisher: Institute of Mathematical Statistics.

Glahn, H. R., and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology and Climatology*, **11 (8)**, 1203–1211, DOI: https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Monthly Weather Review*, **136 (7)**, 2608–2619, DOI: https://doi.org/10.1175/2007MWR2410.1.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, **134 (11)**, 3209–3229, DOI: https://doi.org/10.1175/MWR3237.1.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, DOI: https://doi.org/10.1002/qj.3803.

Hodson, T. O., T. M. Over, and S. S. Foks, 2021: Mean Squared Error, Deconstructed. *Journal of Advances in Modeling Earth Systems*, **13 (12)**, e2021MS002 681, DOI: https://doi.org/10.1029/2021MS002681.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of data assimilations at ECMWF. URL https://www.ecmwf.int/en/elibrary/74969-ensemble-data-assimilations-ecmwf.

Ji, Y., X. Zhi, L. Ji, Y. Zhang, C. Hao, and T. Peng, 2022: Deep-learning-based post-processing for probabilistic precipitation forecasting. *Frontiers in Earth Science*, **10**, DOI: https://doi.org/10.3389/feart.2022.978041.

Kim, H., Y. G. Ham, Y. S. Joo, and S. W. Son, 2021: Deep learning for bias correction of MJO prediction. *Nature Communications*, **12 (1)**, 3087, DOI: https://doi.org/10.1038/s41467-021-23406-3.

Kochkov, D., and Coauthors, 2024: Neural General Circulation Models for Weather and Climate. arXiv, arXiv:2311.07222, DOI: https://doi.org/10.48550/arXiv.2311.07222.

Laloyaux, P., T. Kurth, P. D. Dueben, and D. Hall, 2022: Deep Learning to Estimate Model Biases in an Operational NWP Assimilation System. *Journal of Advances in Modeling Earth Systems*, **14 (6)**, e2022MS003 016, DOI: https://doi.org/10.1029/2022MS003016.

Lam, R., and Coauthors, 2023: GraphCast: Learning skillful medium-range global weather forecasting. arXiv, arXiv:2212.12794, DOI: https://doi.org/10.48550/arXiv.2212.12794.

Le Dimet, F.-X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**, 97–110, DOI: https://doi.org/10.3402/tellusa.v38i2.11706.

Lorenz, C., T. C. Portele, P. Laux, and H. Kunstmann, 2021: Bias-corrected and spatially disaggregated seasonal forecasts: a long-term reference forecast product for the water sector in semi-arid regions. *Earth System Science Data*, **13 (6)**, 2701–2722, DOI: https://doi.org/10.5194/essd-13-2701-2021.

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

Mariotti, A., and Coauthors, 2020: Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. *Bulletin of the American Meteorological Society*, **101 (5)**, E608–E625, DOI: https://doi.org/10.1175/BAMS-D-18-0326.1.

Pathak, J., and Coauthors, 2022: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. arXiv, arXiv:2202.11214, DOI: https://doi.org/10.48550/arXiv.2202.11214.

Pedregosa, F., and Coauthors, 2012: Scikit-learn: Machine Learning in Python. URL https://arxiv.org/abs/1201.0490v4.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A Multimodel Subseasonal Prediction Experiment. *Bulletin of the American Meteorological Society*, **100 (10)**, 2043–2060, DOI: https://doi.org/10.1175/BAMS-D-18-0270.1.

Prechelt, L., 1998: Early Stopping - But When? *Neural Networks: Tricks of the Trade*, G. B. Orr, and K.-R. Müller, Eds., Springer, Berlin, Heidelberg, 55–69, DOI: https://doi.org/10.1007/3-540-49430-8_3, URL https://link.springer.com/chapter/10.1007/3-540-49430-8_3.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: WeatherBench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, **12 (11)**, e2020MS002 203, DOI: https://doi.org/10.1029/2020MS002203.

Rasp, S., and S. Lerch, 2018: Neural networks for post-processing ensemble weather forecasts. arXiv, arXiv:1805.09091, DOI: https://doi.org/10.48550/arXiv.1805.09091.

Rasp, S., and Coauthors, 2024: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, **16 (6)**, e2023MS004 019, DOI: https://doi.org/10.1029/2023MS004019.

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, DOI: https://doi.org/10.1038/s41586-019-0912-1.

Robertson, A. W., and F. Vitart, 2019: *Sub-Seasonal to Seasonal Prediction*. Elsevier, DOI: https://doi.org/10.1016/C2016-0-01594-2, URL https://linkinghub.elsevier.com/retrieve/pii/C20160015942.

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, DOI: https://doi.org/10.1098/rsta.2020.0097.

Sun, Y., W. Bao, K. Valk, C. C. Brauer, J. Sumihar, and A. H. Weerts, 2020: Improving Forecast Skill of Lowland Hydrological Models Using Ensemble Kalman Filter and Unscented Kalman Filter. *Water Resources Research*, **56 (8)**, e2020WR027 468, DOI: https://doi.org/10.1029/2020WR027468.

Vannitsem, S., and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, **102 (3)**, E681–E699, DOI: https://doi.org/10.1175/BAMS-D-19-0308.1.

Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores: Evolution of the ECMWF Sub-Seasonal Forecast Skill. *Quarterly Journal of the Royal Meteorological Society*, **140 (683)**, 1889–1899, DOI: https://doi.org/10.1002/qj.2256.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, **98 (1)**, 163–173, DOI: https://doi.org/10.1175/BAMS-D-16-0017.1.

Watt-Meyer, O., N. D. Brenowitz, S. K. Clark, B. Henn, A. Kwa, J. McGibbon, W. A. Perkins, and C. S. Bretherton, 2021: Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, **48 (15)**, e2021GL092 555, DOI: https://doi.org/10.1029/2021GL092555.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, **24 (3)**, 315–325, DOI: https://doi.org/10.1002/met.1654.

Wilks, D., 2019: *Statistical Methods in the Atmospheric Sciences*. Elsevier, DOI: https://doi.org/10.1016/C2017-0-03921-6, URL https://linkinghub.elsevier.com/retrieve/pii/C20170039216.

Yang, L., D. An, and J. A. Turner, 2008: *Handbook of Chinese Mythology*. 1st ed., OUP USA, iSBN: 9798400661174.

Yuan, X., and E. F. Wood, 2012: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resources Research*, **48 (12)**, 2012WR012 256, DOI: https://doi.org/10.1029/2012WR012256.

Yule, G. U., 1907: On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **79 (529)**, 182–193, URL http://www.jstor.org/stable/92723.

Zhou, S., C. Y. Gao, Z. Duan, X. Xi, and Y. Li, 2023: A robust error correction method for numerical weather prediction wind speed based on Bayesian optimization, variational mode decomposition, principal component analysis, and random forest: VMD-PCA-RF (version 1.0.0). *Geoscientific Model Development*, **16 (21)**, 6247–6266, DOI: https://doi.org/10.5194/gmd-16-6247-2023.

Acknowledgment

First and foremost, I would like to thank Dr. Julian Quinting for the excellent supervision and for always taking the time to answer my questions (no matter how detailed or chaotic they were) and for guiding me through this thesis with patience and genuine encouragement. I am also grateful to Prof. Dr. Peter Knippertz for his valuable input following my midterm presentation.

A big thank you goes to Siyu Li, who kindly took care of running the simulations and saved me hours of processing time and stress. I'm also thankful to everyone in the research group "Meteorological Data Science" for creating such a supportive atmosphere and for always being willing to lend an ear when I needed advice or simply someone to talk to and for the many enjoyable lunch breaks we spent together.

To my wonderful office mates, thank you for the many jokes, solidarity in times of stress and the shared snacks (especially the emergency chocolate!). Special thanks to my cycling crew, Sarah and Jasmin, who cycled with me almost every day and turned the daily commute into a joyride, even in the rain. I'm pretty sure the mornings would have felt twice as long without you. Vanessa, thank you for our weekly lunch-dates and for being a constant source of motivation, laughter and perspective. I truly don't know what I would've done without you.

I am deeply grateful to my parents, who have supported me unconditionally throughout every stage of my life and education. Your belief in me has been the foundation for everything I've achieved. And Maike, thank you for always being there for me, even from afar. Your advice, encouragement and calm presence were constant sources of strength.

I also want to thank my friends from the KIT symphony orchestra who not only opened my heart with music every week, but also never missed a chance to gently tease me about my unique way of using Vim :D

Finally, to my trampoline team, thank you for helping me bounce back every week, both physically and mentally. Whether it was during training or just chatting afterwards, you never failed to lift my spirits and celebrate the ups (and ignore the occasional crash landings). And Laura, thank you for always being there, for your open heart and for always listening with care and making a real effort to understand my thoughts, no matter how technical or tangled they were. Your support gave me the kind of boost no trampoline ever could.