Outsmarting the Storm: Evaluating AI in Weather Forecasting

A Comparative Analysis of the AI-Driven GraphCast and Pangu-Weather Models Against HRES and Aspire in Operational Context, Evaluated with Observational Data

K.L. Reith

Delft University of Technology

TU Delft

Outsmarting the Storm: Evaluating AI in Weather Forecasting

A Comparative Analysis of the AI-Driven GraphCast and Pangu-Weather Models Against HRES and Aspire in Operational Context, Evaluated with Observational Data

by

K.L. Reith

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Friday November 15, 2024 at 15:00.

Student number:	4667247	
Project duration:	March 14, 2024 – November 15, 2024	
Thesis committee:	Prof. dr. ir. R. A. Verzijlbergh	TU Delft, Supervisor
	Dr. J. Sun	TU Delft, Second Supervisor
	Dr. ir. J. M. Maljaars	Whiffle b.v., Company Supervisor
	Ir. A. Julian	Whiffle b.v.

Cover: Hurricane Florence is pictured from the International Space Station as a category 1 storm as it was making landfall near Wrightsville Beach, North Carolina, Sept. 14, 2018 by NASA under CC BY-NC 2.0 (Modified)

TU Delft

Abstract

This thesis evaluates AI-based weather forecasting models—specifically GraphCast and Pangu-Weather—against traditional numerical weather prediction models like the ECMWF High-Resolution Model (HRES) and the Aspire Meso and LES models, focusing on the Netherlands due to its dense observational network and wind energy infrastructure. The study examines operational considerations, including computational costs and hardware requirements, highlighting that while AI models offer significant computational efficiency during inference, they require substantial resources for training and have limitations in local adaptability and variable inclusion (e.g., Pangu-Weather lacks precipitation data). Performance analysis demonstrates that GraphCast consistently outperforms HRES, Pangu-Weather, and Aspire Meso across various meteorological variables and lead times. Despite the advantages, the models exhibit baseline errors due to shared data sources like ERA5, leading to correlated errors that limit the effectiveness of ensemble forecasting. The research underscores the potential of AI models to enhance forecast accuracy and reduce imbalance costs in wind energy production but also emphasizes challenges related to the black-box nature of AI models, data bottlenecks, and limitations imposed by neural scaling laws, where larger models do not necessarily yield better performance. Recommendations for future work include incorporating more observational data, enhancing model modularity and adaptability, improving temporal resolution, expanding variable ranges, emphasizing validation against observational data, and exploring advanced ensemble techniques. The findings suggest that while AI-based models like GraphCast and Pangu-Weather hold significant promise for advancing weather forecasting, addressing limitations in data quality, model architecture, and operational flexibility is crucial for realizing their full potential in operational settings.

Contents

Summary i

1 Introduction 1
1.1 Advancements in Weather Forecasting Technologies . . . 1
1.2 The Role of Weather Forecasting in Renewable Energy . . . 2
1.3 Types of Weather Forecasting Models . . . 2
1.3.1 Numerical Weather Prediction (NWP) Models . . . 2
1.3.2 Large Eddy Simulation (LES) Models . . . 2
1.3.3 AI-Based Weather Forecasting Models . . . 2
1.4 Evolution and Challenges of AI-Based Models . . . 3
1.4.1 AI Architectures . . . 4
1.4.2 Challenges in AI-based Weather Forecasting . . . 5
1.5 Gap in the Literature . . . 6
1.6 Problem Statement . . . 6
1.7 Research Objectives and Questions . . . 6
1.8 Scope of the Study . . . 7
1.9 Thesis Structure . . . 8

2 Theoretical framework of Forecasting Models 9
2.1 GraphCast . . . 9
2.1.1 Graph Neural Networks: Learning from Relationships . . . 9
2.1.2 Model Architecture . . . 10
2.1.3 Data Sources and Inputs . . . 12
2.1.4 Training Methodology . . . 12
2.1.5 Forecasting Methodology . . . 13
2.1.6 Model Performance and Evaluation . . . 13
2.1.7 Limitations and Future Work . . . 13
2.1.8 Conclusion . . . 14
2.2 Pangu-Weather . . . 14
2.2.1 Transformer Architecture in Weather Forecasting . . . 14
2.2.2 3D Earth-Specific Transformers . . . 15
2.2.3 Model Architecture . . . 15
2.2.4 Hierarchical Temporal Aggregation . . . 15
2.2.5 Data Sources and Inputs . . . 16
2.2.6 Training Methodology . . . 16
2.2.7 Forecasting Methodology . . . 17
2.2.8 Model Performance and Evaluation . . . 17
2.2.9 Limitations and Future Work . . . 17
2.2.10 Conclusion . . . 17
2.3 HRES . . . 17
2.3.1 Model Architecture . . . 18
2.3.2 Physical Parameterization Schemes . . . 18
2.3.3 Data Assimilation System . . . 19
2.3.4 Forecast Generation Process . . . 19
2.3.5 Model Performance and Evaluation . . . 19
2.3.6 Limitations and Future Work . . . 20
2.4 Aspire: LES and Mesoscale Models . . . 20
2.4.1 Atmospheric Boundary Layer (ABL) . . . 21
2.4.2 Large Eddy Simulation (LES) . . . 21

2.4.3 Model Architecture and GPU Parallelization . . . 22
2.4.4 Boundary Conditions and Data Coupling . . . 23
2.4.5 Aspire Mesoscale Model . . . 23
2.4.6 Applications to Renewable Energy . . . 24
2.4.7 Limitations and Future Developments . . . 25
2.5 Summary . . . 25

3 Research Design, Datasets and Methodology 26
3.1 Research Design . . . 26
3.2 Datasets . . . 27
3.2.1 SYNOP Observation Data . . . 28
3.2.2 SYNOP Processing . . . 28
3.2.3 ERA5 Reanalysis Dataset . . . 28
3.2.4 ECMWF Analysis Dataset . . . 29
3.2.5 Comparison of Datasets . . . 29
3.2.6 Understanding the Limitations . . . 30
3.3 Baseline Error Assessment . . . 30
3.3.1 Analysis of Mean Absolute Error . . . 30
3.3.2 Analysis of Bias . . . 33
3.3.3 Spatial Patterns . . . 33
3.3.4 Critical Assessment and Conclusions . . . 33
3.4 Model Setup and Configuration . . . 34
3.4.1 GraphCast . . . 34
3.4.2 Pangu-Weather . . . 36
3.4.3 HRES . . . 37
3.4.4 Aspire . . . 38
3.4.5 Computational Cost Summary . . . 41
3.5 Ensemble Forecasting . . . 42
3.5.1 Implementation in This Study . . . 42
3.5.2 Summary . . . 43
3.6 Evaluation of Results . . . 43
3.6.1 Dimensions of Analysis: Aggregating and Averaging . . . 43
3.6.2 Methods . . . 44

4 Analysis of Forecast Performance 47
4.1 Visualization of Model Forecasts . . . 47
4.2 Time Series Analysis of Forecasts . . . 48
4.3 Lead Time Analysis . . . 51
4.4 Spatial Analysis of Forecast Errors . . . 51
4.5 Scatter plots . . . 58
4.6 Inter-model Similarity . . . 60
4.7 Assessment of Ensemble Model Performance . . . 61
4.8 Analysis of Large Eddy Simulation Model Performance . . . 62
4.9 Summary and Conclusions . . . 64

5 Discussion 66
5.1 Operational Considerations . . . 66
5.1.1 Computational Costs and Hardware Requirements . . . 66
5.1.2 Variable Selection and Relevance . . . 66
5.2 Model Performance and Economic Implications . . . 67
5.2.1 Performance Analysis of Models . . . 67
5.2.2 Performance of Ensemble Forecasts . . . 69
5.2.3 Discussion on Imbalance Costs . . . 69
5.3 Implications of Neural Scaling Laws for Weather Forecasting Models . . . 69
5.3.1 Scaling Laws and Model Performance . . . 70
5.3.2 Model Size and Training Compute . . . 70
5.3.3 Data Quality and Model Architecture . . . 70
5.3.4 Intermodel Correlation and Data Bottlenecks . . . 70

5.3.5 Limitations of Scaling in Weather Forecasting . . . 70
5.3.6 Implications for Future Model Development . . . 71
5.3.7 Conclusion . . . 71
5.4 Model Diversity . . . 71
5.5 Limitations . . . 72
5.5.1 Limitations of AI Forecasting . . . 72
5.5.2 Limitations of This Study . . . 72
5.6 Implications for Weather Forecasting . . . 73
5.7 Recommendations and Future Work . . . 73
5.7.1 Incorporate More Observational Data . . . 73
5.7.2 Enhance Model Modularity and Adaptability . . . 74
5.7.3 Improve Temporal Resolution . . . 74
5.7.4 Expand Variable Range . . . 74
5.7.5 Emphasize Validation Against Observational Data . . . 74
5.7.6 Explore Advanced Ensemble Techniques . . . 74
5.8 Conclusions . . . 74

6 Conclusions **75**
6.1 Summary of Research . . . 75
6.1.1 Theoretical Framework of Forecasting Models . . . 75
6.1.2 Research Design and Methodology . . . 75
6.1.3 Analysis of Forecast Performance . . . 76
6.2 Implications for Weather Forecasting . . . 76
6.2.1 Validation of AI-Based Forecasting . . . 76
6.2.2 Challenges and Limitations . . . 77
6.2.3 Implications of Neural Scaling Laws and Model Architecture . . . 77
6.2.4 Model Diversity and Ensemble Forecasting . . . 77
6.2.5 Economic Implications . . . 78
6.2.6 Operational Benefits and Challenges . . . 78
6.3 Limitations of the Study . . . 78
6.4 Recommendations for Future Work . . . 78
6.4.1 Incorporate More Observational Data . . . 78
6.4.2 Enhance Model Modularity and Adaptability . . . 78
6.4.3 Improve Temporal and Spatial Resolution . . . 79
6.4.4 Expand Variable Range . . . 79
6.4.5 Emphasize Validation Against Observations . . . 79
6.4.6 Explore Advanced Ensemble Forecasting Techniques . . . 79
6.5 Final Conclusion . . . 79

References **81**

A Supplementary Figures and Maps **85**
A.1 Introduction . . . 85
A.2 Supplementary Lead Time Analysis . . . 85
A.2.1 Lead Time Plots by Metric . . . 85
A.3 Supplementary Spatial Analysis . . . 90
A.3.1 Geographic Map Plots . . . 90
A.4 Supplementary Taylor Diagrams . . . 108
A.4.1 Taylor Diagrams . . . 108
A.5 Supplementary Inter-model Similarity Analysis . . . 109
A.5.1 Model Similarity Matrices . . . 109
A.6 Supplementary Scatter Plots . . . 113
A.7 Supplementary Time Series Plots . . . 117
A.7.1 Time Series Plots . . . 117

B Detailed Analysis of LES Model Performance **120**

C Imbalance Cost Calculation for Example Wind Farm **124**
C.1 Introduction . . . 124

C.2 Assumptions . . . 124
C.3 Annual Energy Production . . . 125
C.4 Baseline Imbalance Cost Calculation . . . 125
C.5 Adjusted Imbalance Costs for Model Performance Changes . . . 125
C.5.1 GraphCast Model . . . 125
C.5.2 Pangu-Weather Model . . . 125
C.5.3 Aspire Meso Model . . . 126
C.5.4 Four Model Ensemble . . . 126
C.6 Cost Impact Assessment . . . 126
C.6.1 GraphCast Model Cost Savings . . . 126
C.6.2 Pangu-Weather Model Additional Cost . . . 126
C.6.3 Aspire Meso Model Additional Cost . . . 126
C.6.4 Four Model Ensemble Cost Savings . . . 127

1

Introduction

Weather forecasting is a cornerstone of modern society, with implications spanning from everyday decisions to critical operations in sectors like agriculture, energy, and disaster management. The increasing reliance on renewable energy sources, particularly wind and solar power, has further underscored the need for precise and reliable weather predictions (Sweeney et al., 2020).

1.1. Advancements in Weather Forecasting Technologies

The journey of weather forecasting has been marked by continuous innovation. Vilhelm Bjerknes laid the groundwork for Numerical Weather Prediction (NWP) in 1904, proposing that weather forecasting could be approached as a deterministic initial-value problem using the laws of physics. He recognized that this approach required both an accurate knowledge of the atmosphere's initial state and the laws governing its evolution, yet he also realized the complexity involved in solving the necessary nonlinear partial differential equations analytically (Bjerknes, 1904).

Between 1916 and 1922, Lewis Fry Richardson attempted to solve weather forecast equations using numerical methods. Although his hand-calculated 6-hour forecast proved unrealistic, his visionary work *Weather Prediction by Numerical Process* laid the foundation for future computational forecasting efforts. He famously envisioned a large-scale "forecast factory," where numerous human calculators would work in tandem to compute weather predictions faster than real-time, highlighting the immense computational challenge of the time (Richardson, 1922).

The development of electronic computers provided the necessary computational power to overcome the hurdles faced by Richardson. In 1950, the first successful computerized weather forecast was produced by a team led by John von Neumann and Jule Charney using the ENIAC computer. Charney's simplification of the general equations of atmospheric motion, known as the quasi-geostrophic approximation, was pivotal in making numerical weather prediction feasible (Charney, 1948).

Subsequent advancements in computational capabilities and data assimilation techniques throughout the 20th century significantly enhanced forecast accuracy. The launch of the TIROS-1 satellite in 1960 marked the beginning of the integration of satellite data into weather models, providing crucial global observations that significantly improved model initialization and forecast skill (Coiffier, 2011). Ensemble forecasting emerged later as a strategy to address inherent uncertainties in weather prediction by running multiple simulations with varied initial conditions (Gneiting & Raftery, 2005).

Recent years have seen the rise of machine learning (ML) and artificial intelligence (AI) in weather forecasting. AI models are now capable of processing vast datasets and discerning complex patterns, offering the potential to further enhance prediction accuracy and computational efficiency (De Burgh-Day & Leeuwenburg, 2023).

1.2. The Role of Weather Forecasting in Renewable Energy

The global shift towards renewable energy sources has amplified the importance of accurate weather forecasts. Wind and solar power generation are intrinsically linked to weather conditions, and their variability presents challenges for grid management. Precise weather forecasts are indispensable for optimizing operational strategies for power producers, ensuring grid stability, and predicting energy demand, thereby facilitating the seamless integration of renewable energy into the power grid (International Renewable Energy Agency, 2020).

Fabbri et al. (2005) show that forecast errors can have a very large impact on operational cost for wind farms. This is due to three main reasons: imbalance cost of not producing the energy you forecasted (positive wind errors), the opportunity cost of selling the energy you are producing at a lower price (negative wind error), and general operational costs due to less than optimal planning of maintenance. While Fabbri et al. (2005) is dated, its probabilistic approach to modeling forecast errors is still relevant. Their estimate that the error prediction cost can reach up to 10% of the wind farm operator's total cost shows the importance of accurate forecasting.

1.3. Types of Weather Forecasting Models

This section provides a brief introduction to three fundamental approaches to weather forecasting: Numerical Weather Prediction (NWP), Large Eddy Simulations (LES), and AI-Based Weather Forecasting. The first two methods, NWP and LES, are rooted in physical principles and rely on mathematical representations of atmospheric processes, while the third approach leverages data and machine learning techniques to forecast weather patterns.

1.3.1. Numerical Weather Prediction (NWP) Models

Numerical Weather Prediction models are the cornerstone of traditional weather forecasting. They simulate atmospheric processes by numerically solving mathematical equations derived from the fundamental laws of physics, such as the conservation of mass, momentum, and energy. Exemplified by models like the ECMWF IFS HRES (Integrated Forecasting System High Resolution), NWP models use initial conditions obtained from observations to predict the future state of the atmosphere. They provide high-resolution global predictions and are considered the gold standard in operational forecasting (Lang et al., 2024). NWP systems such as the IFS offer both deterministic forecasts, like HRES, and probabilistic ensemble forecasts through systems like the IFS ENS (Ensemble Prediction System).

1.3.2. Large Eddy Simulation (LES) Models

Large Eddy Simulation models are an extension of physics-based modeling, focusing on resolving turbulent flows explicitly. LES models solve the Navier-Stokes equations, filtering out the smaller-scale turbulence while directly simulating the larger eddies. This approach provides detailed, high-resolution insights into atmospheric dynamics, particularly valuable for applications requiring hyper-local forecasts, such as wind farm operations (Siebesma, Wyszogrodzki, et al., 2010). However, the explicit resolution of turbulence in LES models comes at a significant computational cost. Moreover, due to the chaotic nature of turbulence, the predictive accuracy of LES models diminishes more rapidly over time compared to models that parameterize turbulence.

1.3.3. AI-Based Weather Forecasting Models

AI-based models represent a paradigm shift from traditional physics-based forecasting to data-driven approaches. These models employ machine learning techniques to learn patterns and relationships in vast amounts of historical weather data. Instead of relying on explicit physical equations, AI models implicitly capture atmospheric behaviors through statistical learning (McNally et al., 2024).

Key distinctions and advantages of AI-based models include:

- **Data-Driven Approach**: AI models learn from data, enabling them to identify complex, non-linear relationships in atmospheric phenomena that may be challenging to represent explicitly in physical models.
- **Computational Efficiency**: Once trained, AI models can generate forecasts much faster and with lower computational resources than traditional NWP models, as they avoid the computationally

intensive process of solving differential equations (Dueben & Bauer, 2018).

- **Scalability and Adaptability**: AI models can easily incorporate new data sources and scale to handle large datasets, enhancing their ability to capture intricate patterns across different spatial and temporal scales (De Burgh-Day & Leeuwenburg, 2023).
- **Facilitated Ensemble Forecasting**: The computational efficiency of AI models makes it feasible to generate large ensembles of forecasts, improving the quantification of uncertainty in weather predictions (Bremnes et al., 2023).

To better illustrate the distinctions between these three types of weather forecasting models, Table 1.1 compares them across key dimensions.

Table 1.1: Comparison of Weather Forecasting Models

Dimension	NWP Models	LES Models	AI-Based Models
Basis	Physics-based	Physics-based	Data-driven
Equations Solved	Primitive equations (e.g., Navier-Stokes, thermodynamics)	Navier-Stokes with explicit turbulence resolution	None (learned from data)
Resolution	Medium to high	Very high	Variable (constrained by data)
Computational Cost	High	Very high	Low (once trained)
Forecast Range	Global, medium to long-term	Local, short-term	Variable
Scalability	Limited by computational resources	Limited	High
Interpretability	High	High	Moderate to low
Applications	General weather forecasting, climate studies	Hyper-local forecasting, wind farm operations	Rapid forecasting, large ensemble generation
Strengths	Well-understood physics, global coverage	Detailed local turbulence modeling	Fast, scalable, handles large datasets
Limitations	Computationally intensive	Extremely high computational cost, short forecast range	Requires large datasets, may lack physical interpretability

In summary, NWP and LES models rely on the explicit numerical solution of physical equations governing atmospheric processes, offering detailed insights but often at high computational costs. In contrast, AI-based models bypass the need to solve these equations directly, instead learning atmospheric behaviors from historical data. This approach offers computational efficiency and scalability but may lack the interpretability and physical constraints inherent in physics-based models.

1.4. Evolution and Challenges of AI-Based Models

The field of AI-based weather modeling has witnessed remarkable progress, with several notable models contributing to its advancement:

- **Dueben and Bauer (2018)**: Pioneered the use of neural networks for global weather prediction, laying the foundation for subsequent research.
- **Weyn et al. (2020, 2021)**: Developed Convolutional Neural Network (CNN)-based models for medium-range forecasting, showcasing the feasibility of AI approaches (Keisler, 2022).
- **Rasp and Thuerey (2021)**: Achieved improved benchmark results with a deep residual CNN model capable of predicting multiple variables (Mardani et al., 2023).

- **Pathak et al. (2022)**: Introduced Adaptive Fourier Neural Operators, enhancing the capture of long-range dependencies in weather patterns.
- **Keisler (2022) and Lam et al. (2023a)**: Leveraged Graph Neural Networks (GNNs), with GraphCast demonstrating performance surpassing traditional NWP models in certain aspects.
- **Bi et al. (2023)**: Developed Pangu-Weather, showing improved skill compared to HRES, utilizing an Earth-specific Transformer architecture.
- **Chen et al. (2023)**: Introduced FuXi, the first to rival the IFS ENS ensemble forecast at medium to long-range forecasting. It does this by training different models for different forecast horizons.
- **Mardani et al. (2023)**: Addressed the resolution problem that arises from using ERA5 as the main data source. It does this by first predicting the mean and then performing a diffusion step to recover the variation over kilometer-scale grid boxes.

In the next section we will zoom in on four of the main models to highlight the wide variety in architectures.

1.4.1. AI Architectures

AI-based weather forecasting models employ a variety of architectures, each leveraging different machine learning techniques to predict atmospheric conditions. These architectures reflect the diversity and innovation in the field, offering unique approaches to capturing the complex patterns of weather data. This section highlights some of the prominent AI architectures used in weather forecasting, providing an overview of how they work and their contributions to the field.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks are a class of deep learning models particularly effective for processing grid-like data structures, such as images. In weather forecasting, CNNs are utilized to capture spatial features from meteorological data, treating weather maps similarly to images. By applying convolutional filters, CNNs can learn local patterns and spatial hierarchies, making them suitable for capturing localized weather phenomena. Models like those developed by Weyn et al. (2020, 2021) have successfully applied CNNs to medium-range forecasting, demonstrating the potential of CNNs in modeling atmospheric dynamics.

Graph Neural Networks (GNNs)

Graph Neural Networks extend deep learning to graph-structured data, which is beneficial for modeling the irregular and interconnected nature of geographical information. In the context of weather forecasting, GNNs represent the Earth's surface as a graph, where nodes correspond to spatial locations with associated weather variables, and edges represent the relationships between these locations. This allows GNNs to capture both local and global interactions in atmospheric systems. Models like GraphCast (Lam et al., 2023a) have utilized GNNs to process high-dimensional meteorological data efficiently, improving predictive accuracy and computational efficiency.

Transformer-Based Models

Transformers, originally developed for natural language processing, have been adapted for weather forecasting due to their ability to model long-range dependencies in data. By using attention mechanisms (Vaswani et al., 2017), transformers can weigh the importance of different inputs, capturing complex spatiotemporal patterns in atmospheric data. Pangu-Weather (Bi et al., 2023) leverages this capability by introducing Earth-specific transformers that account for geographical variations, enhancing the model's ability to generalize across diverse climatic regions.

Fourier Neural Operators (FNOs)

Fourier Neural Operators combine neural networks with Fourier transforms to efficiently model global dependencies in data. In weather forecasting, FNOs operate in the frequency domain to capture large-scale atmospheric patterns. The Adaptive Fourier Neural Operator (Pathak et al., 2022) leverages this approach to reduce computational complexity while maintaining the ability to model long-range interactions, enabling high-resolution global forecasts.

Ensemble and Cascading Models

Some AI models utilize ensemble and cascading approaches to improve forecasting accuracy over different time horizons. By training specialized models for short-range, medium-range, and long-range forecasts, these approaches aim to minimize cumulative errors. FuXi (Chen et al., 2023), for instance, employs a cascading model with a U-Transformer architecture, integrating convolutional and transformer components to capture both local and global patterns in atmospheric data.

These diverse AI architectures illustrate the innovative ways machine learning techniques are being applied to weather forecasting. Each offers unique advantages—whether it's the spatial feature extraction of CNNs, the relational modeling of GNNs, the long-range dependency capture of transformers, or the computational efficiency of FNOs. Together, they contribute to the advancement of AI-based weather forecasting, providing faster and potentially more accurate predictions by leveraging data-driven methods.

In the subsequent chapters, we will delve deeper into some of these architectures to explore their methodologies and assess their performance in weather prediction tasks.

1.4.2. Challenges in AI-based Weather Forecasting

Despite these advancements, several challenges persist in the operational implementation of AI-based weather models (De Burgh-Day & Leeuwenburg, 2023; Dueben & Bauer, 2018):

- **Limited Resolution**: Current AI-based models often have limited temporal and spatial resolution. While models like GraphCast excel at global-scale predictions, they struggle with localized, short-term forecasts crucial for applications such as renewable energy management and severe weather alerts. This is partly due to the reliance on coarse-resolution data, which fails to capture the complexities of local weather phenomena.
- **Reliance on Reanalysis Data**: Most studies rely on reanalysis datasets for training and assessment, which may not fully encapsulate the complexities of actual atmospheric conditions. This reliance, coupled with potential overfitting and resolution limitations, poses challenges for the operational deployment of AI models (McNally et al., 2024).
- **Lack of Model Diversity**: There is a noticeable lack of diversity in the modeling approaches used in AI-based weather prediction. Most models are trained on similar datasets, such as ERA5 reanalysis, leading to high inter-model similarity and reducing the robustness of ensemble forecasting systems. This uniformity highlights the need for incorporating different data sources and architectural innovations to enhance model diversity and improve forecast reliability (De Burgh-Day & Leeuwenburg, 2023).
- **Computational Demands**: While AI models can offer fast predictions at inference, their training process is computationally intensive, requiring significant resources. This poses a barrier to their widespread adoption, especially in operational settings where computational efficiency is a key consideration (De Burgh-Day & Leeuwenburg, 2023).
- **Integration of Additional Variables**: There is difficulty in integrating additional atmospheric variables beyond the commonly used surface parameters. Variables like solar radiation, cloud cover, and upper-atmospheric dynamics are critical for specific applications but are often omitted in current AI models. This limitation reduces the applicability of AI models in domains that require a more comprehensive set of forecast variables (Dueben & Bauer, 2018).
- **Transparency and Trust**: The "black-box" nature of many AI approaches raises concerns about transparency and trust, especially in critical decision-making scenarios. Addressing these concerns through the development of interpretable AI models and robust validation against real-world observational data is essential for gaining wider acceptance in the meteorological community (De Burgh-Day & Leeuwenburg, 2023).

Some initiatives have begun defining alternative ways of assessing AI-based models but remain tied to gridded datasets instead of direct observations. Notably, recent studies have assessed the performance of leading global models—GraphCast, Pangu-Weather (Olivetti & Messori, 2024; Pasche et al., 2024), and FourCastNet (Pasche et al., 2024)—in extreme weather scenarios. Both studies found that

performance was highly dependent on specific scenarios; some models perform better in certain geographical regions or atmospheric conditions. They conclude that these models can serve as useful additions to current physics-based models, but challenges need to be overcome before widespread operationalization.

1.5. Gap in the Literature

Despite significant advancements in AI-based weather forecasting models, a critical gap remains in their comprehensive evaluation against traditional numerical weather prediction (NWP) models using real-world observational data, especially within operational contexts. Most existing studies focus on model development and performance metrics derived from reanalysis datasets or controlled environments, which may not fully capture the complexities and variabilities of actual atmospheric conditions. Additionally, there is a paucity of research examining the operational considerations, such as computational efficiency, hardware requirements, and economic implications, associated with deploying AI-based models in real-world settings.

Furthermore, previous evaluations often overlook region-specific assessments that consider the unique climatic conditions and data availability of areas like the Netherlands, which is of particular interest due to its dense observational network and growing reliance on renewable energy sources. The integration of AI models into operational workflows, especially for applications with significant economic impacts like wind energy forecasting, necessitates a thorough understanding of their performance, limitations, and practical feasibility compared to established NWP models.

This gap underscores the need for research that not only compares the predictive accuracy of AI-based models to traditional models using observational data but also explores their operational viability and potential economic benefits in specific regional contexts.

1.6. Problem Statement

Although AI-based weather models like GraphCast and Pangu-Weather have shown considerable promise in enhancing forecasting accuracy, several critical issues hinder their widespread operational adoption. Their reliance on extensive training data, often derived from existing NWP models like HRES, raises concerns about the diversity of error patterns and potential asymptotic improvements in forecast skill. The black-box nature of AI models poses challenges for interpretability and trustworthiness, making it difficult for forecasters to diagnose errors or biases. Additionally, these models may lack flexibility in adapting to local scales or specific regional phenomena due to their global design, and they require substantial computational resources for both training and operational deployment.

Moreover, there is limited research assessing the economic implications of employing AI-based models in sectors sensitive to forecast accuracy, such as renewable energy. Understanding how these models perform in operational settings, particularly in regions with dense observational networks and significant economic stakes like the Netherlands, is crucial.

Ensembling models is a common practice in operational forecasting to improve accuracy and reliability. However, the benefits of combining AI-based models with traditional NWP models have not been thoroughly investigated.

To address these issues, a comprehensive evaluation of GraphCast, Pangu-Weather, traditional NWP models like HRES, and state-of-the-art LES models such as Aspire, including the assessment of model ensembles, is essential. Such an evaluation will help identify the specific strengths and limitations of AI models and their ensembles in operational contexts, assess their economic impact, and determine the necessary improvements for their effective integration into existing forecasting systems.

1.7. Research Objectives and Questions

The primary objective of this research is to conduct a comprehensive evaluation of AI-based models GraphCast and Pangu-Weather, comparing their performance with that of the IFS HRES model by ECMWF and the Aspire model by Whiffle, using observational data from the Netherlands. This evaluation includes assessing the benefits of model ensembles and the performance of Aspire in both mesoscale and Large Eddy Simulation (LES) settings at two specific station locations. The specific

objectives are:

- **Assess Forecast Accuracy**: Quantitatively evaluate the accuracy of each model, including ensembles, in predicting key atmospheric variables across varying lead times and spatial locations, using observations from the Royal Netherlands Meteorological Institute (KNMI) SYNOP stations.
- **Analyze Operational Viability**: Examine the operational considerations, including computational efficiency, hardware requirements, and ease of integration into existing workflows, associated with each model and their ensembles.
- **Evaluate Economic Implications**: Assess the potential economic impacts of forecast accuracy differences among the models and their ensembles, particularly in the context of renewable energy operations like wind farms.
- **Identify Strengths and Limitations**: Analyze the unique advantages and limitations of each model, including their ability to capture localized weather phenomena, short-term variability, dependence on training data, and performance at different scales.

To achieve these objectives, the study addresses the following research questions:

1. How does the forecast accuracy of GraphCast, Pangu-Weather, and their ensembles compare to HRES and Aspire in predicting key atmospheric variables (e.g., 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total precipitation) across varying lead times and spatial locations?
2. What are the operational considerations, such as computational efficiency and hardware requirements, associated with deploying AI-based models like GraphCast and Pangu-Weather, and their ensembles, compared to traditional NWP models and LES models?
3. What are the potential economic implications of using AI-based models and their ensembles over HRES and Aspire, particularly in the context of wind energy forecasting in the Netherlands?
4. What are the specific strengths and limitations of AI-based models and their ensembles in terms of model interpretability, dependence on training data, flexibility in local adaptation, and performance at different scales, compared to traditional NWP models?

1.8. Scope of the Study

This study focuses on the Netherlands, leveraging its dense observational network of 47 KNMI SYNOP stations, which provide high-resolution measurements of atmospheric variables essential for accurate evaluation. The models under evaluation are GraphCast and Pangu-Weather (AI-based global weather forecasting models utilizing advanced machine learning techniques), HRES (ECMWF's NWP model based on physical principles), and Aspire (A high-resolution turbulence resolving LES model). Aspire is assessed in both mesoscale settings across all stations and at even finer scales using LES at two specific station locations to evaluate its performance in capturing localized weather phenomena.

In addition to individual model assessments, the study also explores the benefits of model ensembles, specifically evaluating how combinations of models perform compared to the models individually. Ensembling is crucial in operational settings, as models are rarely used in isolation.

The evaluation encompasses daily forecasts for the entire year of 2022, up to 96 hours ahead, at 6-hour intervals. The year 2022 was chosen because it is not included in GraphCast's and Pangu-Weather's training data, ensuring an unbiased assessment of their forecasting capabilities. The study focuses on surface variables, constrained by the variables available from the AI-based models and the observational data, including 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total precipitation over the last 6 hours.

In addition to assessing forecast accuracy, the study examines operational considerations such as computational efficiency, hardware requirements, and the models' adaptability to operational settings. Furthermore, it evaluates the economic implications of forecast accuracy differences among the models and their ensembles, particularly their impact on wind energy operations in the Netherlands.

1.9. Thesis Structure

The thesis is organized into six chapters, each serving a distinct purpose in presenting the research and findings:

- **Chapter 1: Introduction** — Establishes the context and importance of accurate weather forecasting, particularly in the realm of renewable energy and wind power generation. It highlights the emerging role of AI in weather prediction, identifies gaps in current research, and sets forth the problem statement, research objectives, and key research questions addressed in the study.
- **Chapter 2: Theoretical Framework of Forecasting Models** — Provides an in-depth exploration of the four weather forecasting models evaluated: GraphCast, Pangu-Weather, IFS HRES, and Aspire. It discusses their theoretical foundations, architectures, data inputs, and operational mechanisms, setting the stage for understanding their comparative performance and operational considerations.
- **Chapter 3: Research Design, Methodology, and Evaluation** — Details the research design, including the selection of observational data from KNMI SYNOP stations and the rationale for focusing on the Netherlands. It describes the configuration and setup of each model, the specific variables analyzed, the evaluation metrics used to assess forecast performance, and the methods for analyzing operational and economic implications. The chapter also explains the approach for creating and evaluating model ensembles and the assessment of Aspire in LES settings at two specific station locations.
- **Chapter 4: Analysis of Forecast Performance** — Presents a comprehensive analysis of the models' predictive capabilities, including statistical evaluations of forecast accuracy across different lead times, spatial locations, and atmospheric variables. It compares individual models as well as ensembles. The analysis also includes the performance of Aspire in LES setting.
- **Chapter 5: Discussion** — Interprets the key findings from the analysis, discussing their implications for weather forecasting, operational deployment, and economic considerations. It addresses the strengths and limitations of each model and ensemble, considering factors like accuracy, computational efficiency, interpretability, and adaptability at different scales. The chapter also acknowledges the study's limitations and provides recommendations for future research and model development.
- **Chapter 6: Conclusions** — Summarizes the main findings, emphasizing the potential of AI-based models like GraphCast and Pangu-Weather to advance weather forecasting. It highlights the importance of integrating diverse data sources, enhancing model adaptability, and balancing the strengths of AI models with traditional NWP models and ensembles. The chapter concludes with final remarks on the evolving landscape of weather prediction and the role of AI and ensemble methods in meeting the needs of society amid changing climatic conditions.

2

Theoretical framework of Forecasting Models

In this chapter, we explore the theoretical frameworks underlying the weather forecasting models utilized in this study. The chapter aims to provide an understanding of the diverse methodologies used to generate weather predictions, offering insights into their architectures, data inputs, and operational mechanisms.

Weather forecasting has evolved from basic empirical methods to sophisticated numerical and machine learning-based models, each designed to address specific challenges in predicting atmospheric conditions. This chapter delves into the core concepts behind the models compared, highlighting the differences in their approaches to simulating and predicting weather phenomena.

The four distinct weather forecasting models are: GraphCast, Pangu-Weather, IFS HRES (which we will call HRES), and Aspire, each representing different approaches to atmospheric prediction. GraphCast and Pangu-Weather are cutting-edge, data-driven models that leverage artificial intelligence and machine learning techniques to forecast weather based on historical datasets, making it highly adaptable to complex meteorological patterns. In contrast, HRES is a numerical weather prediction (NWP) model that relies on the fundamental laws of physics, utilizing equations governing atmospheric dynamics and thermodynamics to simulate weather conditions. Lastly, Aspire is a large eddy simulation (LES) model, designed for high-resolution, turbulence-resolving simulations, providing detailed insights into small-scale atmospheric phenomena. In this study Aspire is used both in turbulence resolving LES setting and in mesoscale turbulence parameterizing setting, more on this in subsection 3.4.4.

2.1. GraphCast

GraphCast represents a significant advancement in the application of machine learning to medium-range global weather forecasting. It leverages a Graph Neural Network (GNN) architecture to model the intricate and dynamic interactions inherent in atmospheric systems, delivering accurate predictions across a 10-day forecast horizon. In this section, we delve into the technical details of GraphCast, encompassing its architecture, training process and data sources.

2.1.1. Graph Neural Networks: Learning from Relationships

Before diving into the specific design of GraphCast, it is important to start by laying the foundation of Graph Neural Networks (GNNs) and why they are well-suited to the problem of atmospheric modeling and weather forecasting.

Graph Neural Networks (GNNs) are a class of deep learning models specifically designed to operate on graph-structured data. In the context of weather forecasting, where atmospheric variables exhibit complex spatial relationships, GNNs offer a framework for capturing these dependencies (Wu et al., 2020). At their core, GNNs function by iteratively updating the representations (or embeddings) of

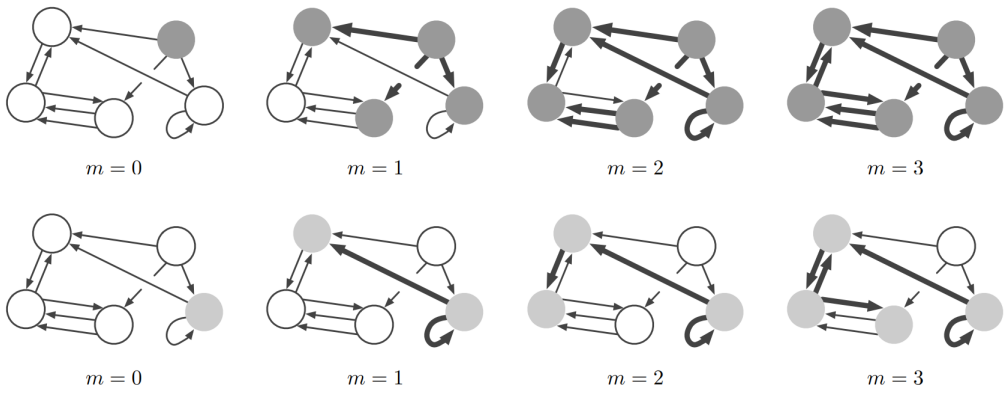


Figure 2.1: Two diffusion paths through a graph using message passing. Each row shows how information starting in the top right node (top row) and bottom right node (bottom row) diffuses through the graph with each subsequent message passing step (m=0,1,2,3). After each round of message passing the color of a note or edge is updated if it has seen new information. Note that in the bottom row the top right node never sees new information, as all edges are directed away from that note: the color remains white. (P. W. Battaglia et al., 2018)

nodes and edges in a graph based on the representations of their neighboring nodes and edges. This process, often referred to as message passing, allows information to propagate through the graph, enabling the model to learn about the relationships between different nodes and their features (Bronstein et al., 2021). Figure 2.1 shows how information propegates through two different graphs over a series of message passing steps.

The concept of using GNNs for relational reasoning draws heavily from the foundational work on Interaction Networks (P. Battaglia et al., 2016), which introduced a framework for reasoning about complex systems by modeling objects and their interactions separately. Interaction Networks explicitly separate the reasoning about relations (how objects interact) from the reasoning about objects themselves, allowing for a more structured and scalable approach to learning complex dynamics. This idea was further generalized to Graph Networks (P. W. Battaglia et al., 2018), providing a flexible blueprint for building GNN architectures capable of learning and reasoning about entities and their relationships.

In the case of GraphCast, the nodes in the graph represent spatial locations in the latent space (the multimesh: more on this in Figure 2.1.2), and the edges encode the connectivity between these locations. The features associated with each node encapsulate the relevant weather variables at that specific location. During the message-passing process, each node aggregates information from its neighbors, allowing the model to learn how weather patterns at one location influence those at neighboring locations.

The power of GNNs in weather forecasting is twofold:

1. Their ability to learn from the relational structure of the data. In weather forecasting, this translates to the model's capacity to capture how atmospheric variables interact across different spatial scales, from local interactions to global phenomena. By learning these relationships, GNNs can make more informed predictions about the evolution of weather patterns over time.
2. GNNs can efficiently handle high-dimensional input and output data, reducing the complexity of the model while maintaining high resolution and accuracy.

1)

The following section will delve deeper into the specific architecture of GraphCast and its set of different GNN's used at various steps of the forecasting process.

2.1.2. Model Architecture

GraphCast employs an encoder-processor-decoder framework, which allow is the model to first encode input data to a latent space, do processing in this latent space, and finally decode it back to gridded

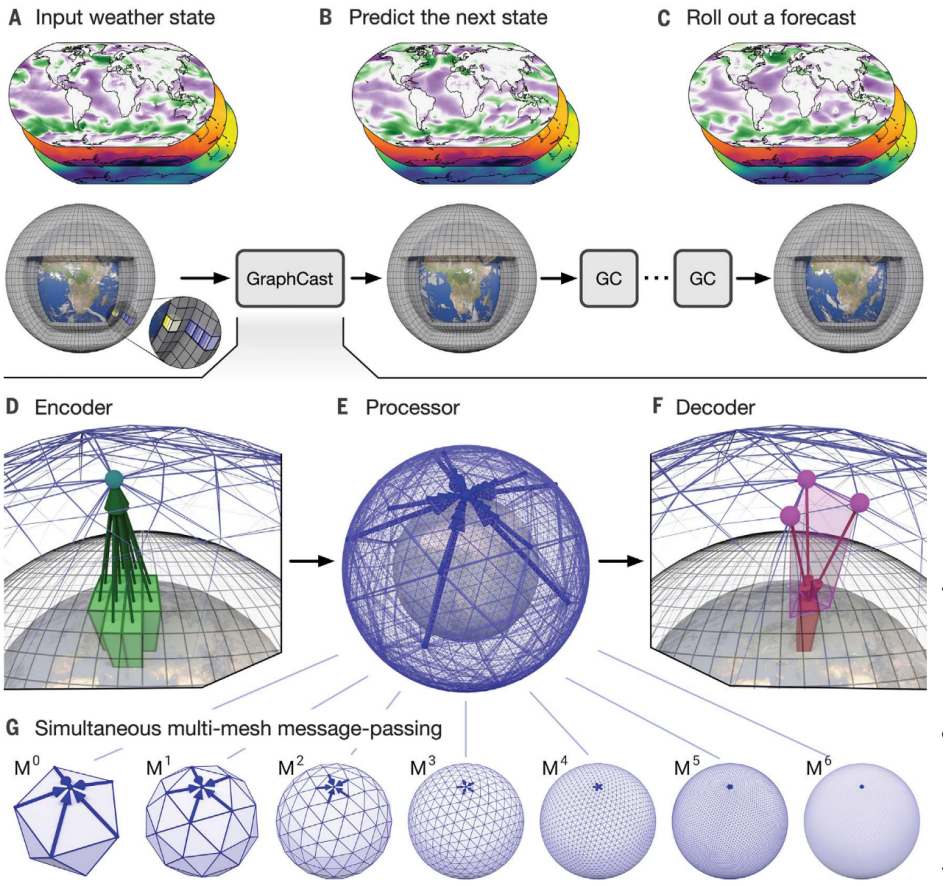


Figure 2.2: The process through which GraphCast generates forecasts. First mapping the current weather state to a latent representation in form of the multimesh (D), then performing 16 message-passing steps with unshared neural network weights (E). Once all nodes and edges of the graph have been updated, the nodes are mapped to the future weather state as a grid representation (F). This can be autoregressively repeated to create each subsequent forecast timestep. (G) describes how the multimesh latent space is the union of 7 different mesh structures, each of these structures allows for a different connection length and their combination therefore allows for both long and short range dependancies to be learned. (Lam et al., 2023a)

output as a future weather state. The architecture incorporates a multi-scale mesh for spatial representation in the latent space, and a GNN facilitates message-passing across these mesh scales. The model comprises a total of 36.7 Million parameters. For reference, Meta AI's open source Llama 3's largest model consists of 405 Billion parameters (Team & Meta, 2024).

Encoder

The encoder serves as the initial translator, mapping weather variables from a conventional latitude-longitude grid to a latent space embodied by a refined icosahedral multimesh (see (G) in Figure 2.2). The input weather state encompasses surface variables (such as 2 m temperature, 10 wind components, and mean sea level pressure) and atmospheric variables (like temperature, wind components, geopotential, and humidity) across 37 pressure levels. This translates to a substantial 227 variables associated with each grid point.

GraphCast employs a bipartitie GNN layer to embed each grid point into a feature representation on the multimesh. This process effectively captures the local dynamics at each grid point while also setting the stage for the integration of global interactions within the subsequent processor stage.

Processor

The processor constitutes the computational heart of GraphCast's GNN. It employs 16 layers of unshared GNNs to execute learned message-passing operations on the multimesh. This design facilitates the efficient propagation of information across both short- and long-range spatial dependencies. The message-passing mechanism allows the model to capture the intricate interplay of local weather

phenomena, such as wind patterns and temperature gradients, with larger-scale atmospheric features like pressure systems and tropical cyclones.

The design of GraphCast's processor is deeply influenced by the Interaction Network framework, which emphasizes reasoning about complex systems through the interactions of entities, represented as nodes, and their relationships, represented as edges (P. Battaglia et al., 2016). The GNN layers in the processor effectively learn to model these interactions, allowing the network to capture the complex, dynamic dependencies of atmospheric systems.

In essence, each GNN layer in the processor refines its understanding of the weather state by exchanging information between nodes connected by edges in the multimesh. The 'unshared' nature of these GNN layers implies that each layer possesses its own set of learnable parameters, enabling the model to capture progressively more complex and nuanced patterns in the weather data as it traverses deeper into the network.

A key feature of the processor GNN is that while each layer is different, there is only a single GNN for all nodes and edges. This implies that the same process is applied across the spatial dimension, which makes sense from a physical intuition point-of-view.

Decoder

Following the message-passing phase in the processor, the decoder assumes the role of translating the refined latent representations on the multimesh back to the familiar latitude-longitude grid. A final GNN layer generates the weather forecasts, framed as residual updates to the most recent input state. This approach is inspired by the separation of object and relational reasoning found in Interaction Networks (P. Battaglia et al., 2016), where the relational dynamics are first updated, followed by the object state. Normalization is applied to these updates to ensure consistency across different time steps. The output encompasses both surface and atmospheric variables, providing a comprehensive prediction of the future weather state.

Multimesh Representation

The multimesh stands as a pivotal innovation within GraphCast, enabling efficient weather prediction on a global scale. It is constructed through the iterative refinement of a regular icosahedron mesh. Each refinement step introduces additional nodes and edges, culminating in a mesh with 40,962 nodes and 327,660 edges at its highest resolution. This hierarchical mesh structure empowers GraphCast to generate fine-grained local predictions while simultaneously accounting for coarse, global-scale interactions, all within a computationally manageable number of message-passing steps Lam et al., 2023b.

2.1.3. Data Sources and Inputs

GraphCast is trained on reanalysis data sourced from the ERA5 archive, which offers a wealth of historical weather data spanning from 1979 to 2017. This data is available at an hourly resolution on a 0.25° latitude-longitude grid. The dataset encompasses surface variables such as 2-meter temperature, mean sea level pressure, and 10-meter wind components, alongside atmospheric variables across 37 pressure levels. More on this in subsection 3.2.3.

2.1.4. Training Methodology

GraphCast's training revolves around minimizing the mean squared error (MSE) between its forecasts and the corresponding ERA5 reanalysis data. The model is optimized using gradient descent applied over 12 autoregressive steps, which equates to 3-day forecasts. This training regime compels the model to learn how to predict multiple time steps into the future, thereby enhancing its accuracy for both short- and medium-range forecasts.

The training process is executed on 32 Cloud TPU v4 devices, with the data partitioned into training (1979-2015) and validation (2016-2017) sets. A curriculum learning strategy is adopted, progressively increasing the number of autoregressive steps during training. This approach facilitates the model's ability to generate accurate long-term forecasts.

A second model was created for operational settings, referred to as GraphCast Operational. This model exhibits key distinctions from the primary model:

- Reduced Vertical Resolution: It operates at 13 pressure levels, compared to the 37 levels of the full-sized model.
- Fine-Tuning with Recent HRES Analysis Data: It is fine-tuned using HRES data from 2018-2021. Incorporating more recent data has been shown to enhance skill compared to training on less recent data.
- Initialization with HRES Data: This model can be initialized using HRES data, making it suitable for operational use. HRES data is disseminated in real-time, whereas ERA5 is released with a delay.

2.1.5. Forecasting Methodology

GraphCast generates weather forecasts spanning 10 days, with predictions made at 6-hour intervals. It adheres to an autoregressive methodology, wherein the model leverages the two most recent weather states (current and previous time steps) to predict the subsequent state. This process is iterated to produce a sequence of weather states that constitute the forecast for the specified horizon.

The forecasting process unfolds through a series of well-defined stages. The encoder initially processes the input weather data into a multimesh representation. Subsequently, the processor engages in message-passing operations to effectively propagate weather information across the mesh. Finally, the decoder translates the processed data back to the original grid, yielding the weather forecasts.

GraphCast's computational efficiency is noteworthy. It generates predictions in under 60 seconds on a TPU device, a stark contrast to traditional numerical models like HRES, which can take up to an hour on extensive computing clusters. This efficiency opens doors to the creation of very large ensemble forecasts with perturbed initial conditions, enhancing the capability to predict low-probability, high-impact events—a critical aspect of extreme weather forecasting. In scenarios where a storm has a less than 5% probability of making landfall, it might be overlooked in an ensemble forecast with fewer than 20 members. GraphCast's low inference cost allows for the generation of ensembles with hundreds of members, significantly improving the ability to capture such rare events.

2.1.6. Model Performance and Evaluation

GraphCast has been evaluated against HRES, the leading operational deterministic forecast system. Lam et al., 2023a show that GraphCast consistently demonstrates better performance than the current best NWP model, outperforming HRES on 90% of 1,380 verification targets. It particularly excels in predicting mid-tropospheric Z500 and surface-level variables. Furthermore, GraphCast exhibits improved skill over HRES in forecasting severe events like tropical cyclones and atmospheric rivers, yielding more accurate predictions for cyclone tracks, integrated vapor transport, and extreme temperature events. Figure 2.3 shows the RMSE skill for various variables as a function of forecast lead time as found by Lam et al., 2023b. Note that this is for the full resolution model against ERA5, the implication of which is twofold:

- This cannot be repeated in operational setting due to a lack of real-time ERA5 data needed for initialization, and;
- Performance is assessed against ERA5 reanalysis data and not against observations, more on the discrepancy between these in section 3.3.

2.1.7. Limitations and Future Work

While GraphCast represents a substantial leap forward in medium-range forecasting, it is not without limitations. The model's training objective, rooted in minimizing MSE, can lead to blurrier forecasts at extended lead times. This blurring arises from the model's tendency to average over potential future weather states, which can be detrimental in scenarios where precise predictions of extreme weather events are crucial. Additionally, GraphCast, in its original formulation, is deterministic, lacking the inherent ability to quantify uncertainty in its predictions.

Addressing these limitations has been a focus of subsequent research. Notably, Google DeepMind has introduced GenCast, Price et al. (2023), a diffusion-based model designed to generate ensemble forecasts. GenCast has demonstrated remarkable capabilities, significantly outperforming the ECMWF's Ensemble Prediction System (ENS) in terms of probabilistic forecasting skill.

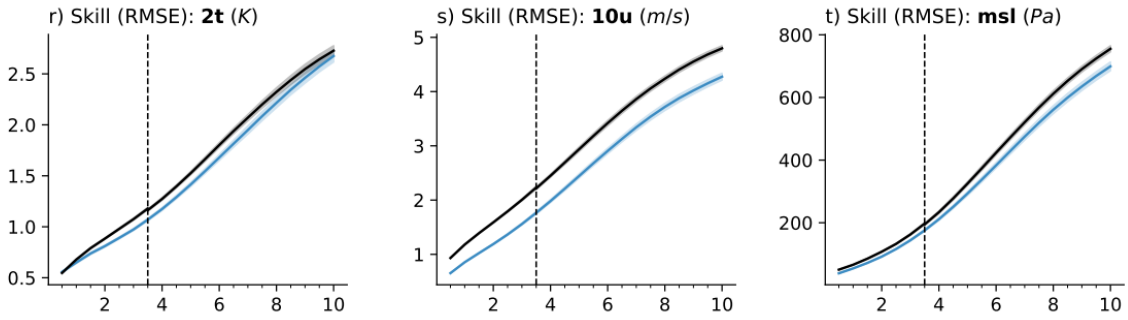


Figure 2.3: GraphCast's (blue lines) absolute RMSE skill versus HRES (black lines) in 2018 (lower is better). Each subplot represents a single surface level variable, as indicated in the subplot titles: 2 meter temperature, 10m u-component of wind and mean sea level pressure. The x-axis represents lead time, at 12-hour steps over 10 days. The y-axis represents RMSE. Figure as in Lam et al., 2023b.

Furthermore, there exists potential for enhancing the spatial resolution of GraphCast's forecasts. The current model is trained on ERA5 data at a 0.25° resolution. However, future iterations could leverage finer-resolution datasets to improve the model's ability to capture and predict smaller-scale weather phenomena. This, however, remains a challenge for current data-driven models due to their reliance on ERA5 as the primary data source.

Additionally, the performance of GraphCast can be further augmented by retraining it on more recent data. This would enable the model to adapt to evolving weather patterns and the impacts of climate change.Lam et al. (2023a) have already showcased the skill improvements achievable through retraining with more recent data, highlighting a key advantage of GraphCast: its capacity for continuous learning and adaptation.

However this doesn't address one of the core problems of global models: their unified global nature inherently means it will be better in some areas than in others. Their reliance on global data leads to a certain inflexibility in spatial fine tuning.

2.1.8. Conclusion

GraphCast embodies a significant milestone in the realm of machine learning-based weather forecasting. Its efficient utilization of GNNs and the innovative multimesh representation empowers it to deliver rapid and accurate 10-day forecasts on a global scale. By surpassing the performance of traditional models like HRES, GraphCast underscores the potential of machine learning to not only complement but also enhance existing numerical weather prediction systems.

In essence, GraphCast exemplifies the power of harnessing machine learning to tackle the complexities of weather forecasting. Its ability to learn from vast datasets and capture intricate spatio-temporal patterns positions it as a valuable tool for improving the accuracy and accessibility of weather predictions, ultimately benefiting a wide array of human activities and decision-making processes.

The following section will explore a second AI-based weather model, highlighting the different approaches taken to learn from historical data.

2.2. Pangu-Weather

Pangu-Weather is a model for medium-range global weather forecasting that utilizes a three-dimensional deep neural network architecture tailored to atmospheric dynamics. Developed by (Bi et al., 2023), this model introduces Earth-specific priors and a hierarchical temporal aggregation strategy. In this section, we examine the architecture, data sources, and features that underpin Pangu-Weather's approach to forecasting meteorological phenomena.

2.2.1. Transformer Architecture in Weather Forecasting

Transformers are a class of deep learning models that have achieved significant success in sequence modeling tasks due to their ability to capture long-range dependencies through self-attention mech-

anisms (Vaswani et al., 2017). In the context of weather forecasting, transformers offer a powerful framework for modeling the complex spatial-temporal relationships inherent in atmospheric data.

The core component of the transformer is the **self-attention** mechanism, which allows the model to weigh the influence of different parts of the input data when making predictions. This is particularly useful for weather forecasting, where the atmospheric state at a given location and time is influenced by conditions at distant locations and previous times.

In Pangu-Weather, the transformer architecture is adapted to handle three-dimensional atmospheric data, encompassing latitude, longitude, and altitude (or pressure levels). By utilizing self-attention across these dimensions, the model can effectively capture the interactions between different regions of the atmosphere, both horizontally and vertically.

Additionally, transformers incorporate **positional encoding** to retain information about the spatial and temporal positions of data points. Pangu-Weather extends this concept by integrating Earth-specific positional biases, accounting for geographical variations in weather patterns. This enhances the model's ability to generalize across diverse climatic regions.

Overall, the transformer architecture provides Pangu-Weather with the capability to model the complex dependencies in atmospheric data, enabling more accurate and efficient weather forecasts.

2.2.2. 3D Earth-Specific Transformers

Building upon the transformer architecture described in Section 2.2.1, Pangu-Weather extends the model to suit the complexities of atmospheric data. Traditional transformers are designed for one-dimensional sequences, but atmospheric data is inherently three-dimensional.

Pangu-Weather addresses this by integrating height information as an additional dimension, conceptualizing the atmosphere as a three-dimensional structure. The core of Pangu-Weather is its **3D Earth-specific Transformer (3DEST)** architecture, which incorporates Earth-specific positional biases. These biases account for geographical variations in weather patterns, enhancing the model's ability to generalize across diverse climatic regions.

By employing a 3D architecture, Pangu-Weather effectively models dependencies between atmospheric variables both horizontally and vertically, capturing the complex spatial relationships inherent in meteorological data (Bi et al., 2023). This spatial dependence is notably different from the way GraphCast uses the same GNN across space.

2.2.3. Model Architecture

Pangu-Weather employs an encoder-decoder framework within the 3DEST architecture. The model is trained on ERA5 reanalysis data (see subsection 3.2.3), with the input consisting of 13 pressure levels for upper-air variables and several surface variables, represented in a 3D data cube. The encoder processes these inputs through a series of transformer blocks, each equipped with Earth-specific priors that incorporate information about the absolute geographical position of each grid cell (Bi et al., 2023).

The encoder gradually reduces the spatial resolution of the input data while increasing the depth of the feature space, capturing progressively higher-level representations of the atmospheric state. The decoder then reverses this process, mapping the high-dimensional feature space back to the original resolution, yielding forecasts for both upper-air and surface variables using a series of transformer blocks.

2.2.4. Hierarchical Temporal Aggregation

A distinctive aspect of Pangu-Weather is its hierarchical temporal aggregation strategy. All weather models accumulate errors as they iteratively predict future states over long lead times. Pangu-Weather attempts to address this issue by training multiple models with different lead times (1 hour, 3 hours, 6 hours, and 24 hours). During inference, the model employs a greedy algorithm that selects the model with the longest feasible lead time at each step, minimizing the number of iterations and reducing cumulative forecast errors (Bi et al., 2023).

This approach both accelerates the forecasting process and aims to enhance accuracy for medium-range forecasts spanning up to 7 days. The use of hierarchical temporal aggregation allows Pangu-

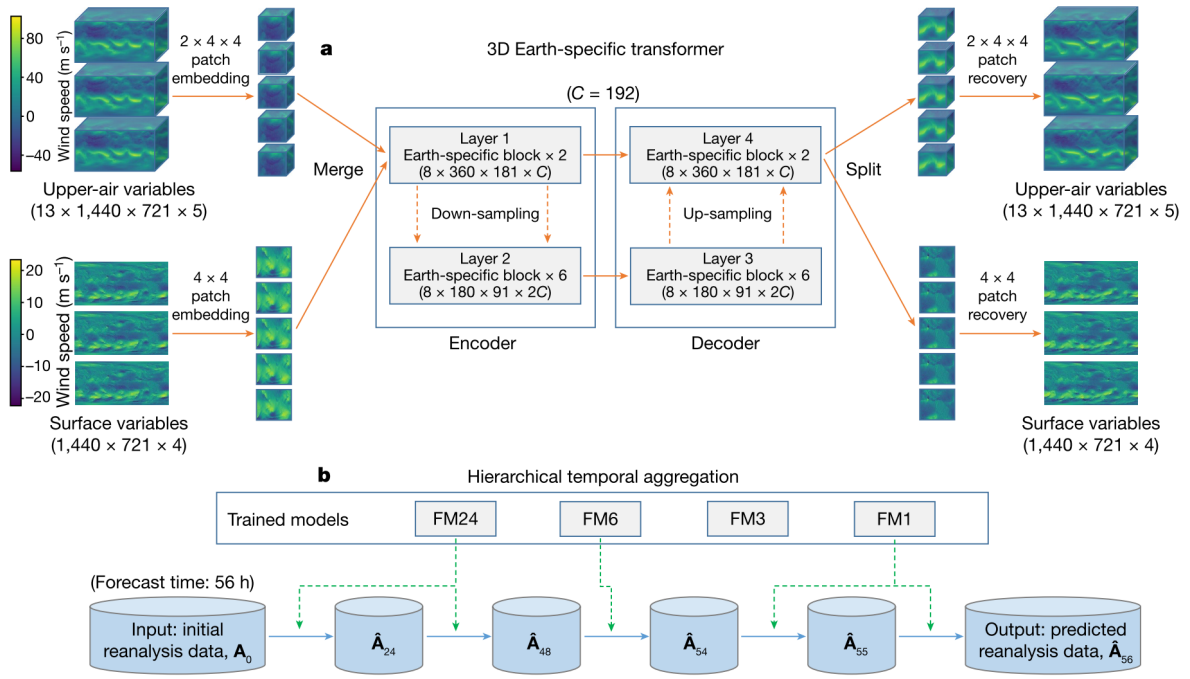


Figure 2.4: The architecture of the Pangu-Weather model. The 3D data cube is processed through the encoder and decoder stages, incorporating Earth-specific priors and hierarchical temporal aggregation. Each block represents a series of transformer layers that model the interactions between atmospheric variables at different pressure levels. Figure as in Bi et al., 2023.

Weather to generate forecasts over longer time horizons with potentially fewer errors compared to conventional models.

2.2.5. Data Sources and Inputs

Similar to GraphCast, Pangu-Weather is trained on the ERA5 reanalysis dataset. The training set consists of hourly data for a 39-year period, encompassing 13 upper-air variables (such as geopotential height, temperature, and wind components) and 4 surface variables (including 2-meter temperature and mean sea level pressure) (Bi et al., 2023). The model's spatial resolution is $0.25° \times 0.25°$, matching the ERA5 dataset.

The input data is processed into a 3D grid with dimensions corresponding to latitude, longitude, and pressure levels. This format enables the model to capture interactions across different layers of the atmosphere.

2.2.6. Training Methodology

Pangu-Weather's training process is computationally intensive, utilizing a cluster of 192 NVIDIA Tesla-V100 GPUs over 16 days for each model (Bi et al., 2023). The training procedure is designed to minimize the Mean Absolute Error (MAE) between the predicted and true atmospheric states. This is different from GraphCast's approach, which optimizes for Mean Squared Error (MSE) instead.

- **MAE**:
  - Less sensitive to outliers; penalizes errors linearly.
  - Provides consistent performance across typical conditions.
  - May underemphasize extreme weather events due to equal weighting of all errors.
  - Leads the model to predict the **median** of the target distribution.
- **MSE**:
  - More sensitive to outliers; penalizes larger errors more heavily.

- Emphasizes reducing large errors, potentially improving extreme event predictions.
- May overfit to outliers, affecting performance on common conditions.
- Leads the model to predict the **mean** of the target distribution.

The model's performance is validated on a separate dataset from 2019 and tested on data from 2018, providing evaluation across different temporal segments.

2.2.7. Forecasting Methodology

Pangu-Weather generates forecasts by iteratively applying the trained models over the desired forecast horizon. The hierarchical temporal aggregation strategy is employed during this process, selecting the most appropriate model for each forecast step. This methodology reduces the number of required iterations, enabling the model to produce medium-range forecasts more efficiently than NWP systems (Bi et al., 2023).

2.2.8. Model Performance and Evaluation

Pangu-Weather has been evaluated against the ECMWF's (HRES) and other AI-based models like FourCastNet. The model reportedly outperforms these benchmarks in terms of both Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC) for various tested variables, including geopotential height, temperature, and wind components (Bi et al., 2023).

For example, Pangu-Weather achieves an RMSE of 296.7 for a 5-day Z500 forecast, compared to 333.7 for HRES and 462.5 for FourCastNet. Moreover, Pangu-Weather's inference time is 1.4 seconds on a single GPU, making it significantly faster than HRES, which requires orders of magnitude longer on a supercomputer (Bi et al., 2023).

2.2.9. Limitations and Future Work

Pangu-Weather does have limitations, similar to GraphCast. The model has been trained only on reanalysis data, and its performance on observational data requires further investigation. Additionally, some important weather variables, such as precipitation, were not included in the current study, which may limit the model's applicability for certain extreme weather events (Bi et al., 2023).

It does suffer from the same data-driven model drawbacks highlighted earlier, with limited ability for scaling, low interpretability, and large upfront costs during training. Besides, various assessments (Lam et al., 2023a; Liu et al., 2024; Olivetti & Messori, 2024) have shown GraphCast to outperform Pangu-Weather at a significantly lower parameter count, 36.7 million vs. 256 million, respectively. This discounts the merit of the architectural innovations (Bi et al., 2023) showcase.

Future work could involve extending the model to incorporate more atmospheric variables and integrating observational data for training. There is also potential for enhancing the model's spatial resolution and exploring ensemble forecasting methods to better capture uncertainty in predictions.

2.2.10. Conclusion

Pangu-Weather represents a different approach in the use of AI for weather forecasting, combining deep learning techniques with a transformer architecture that integrates Earth-specific priors. Its 3D architecture and hierarchical temporal aggregation strategy enable it to deliver reasonably good medium-range forecasts with speed and efficiency. As AI continues to evolve, models like Pangu-Weather may play an increasingly important role in enhancing our ability to predict and respond to complex weather phenomena. In this report, it will serve a key function as a second data-driven model in the comparison, allowing for more general and robust conclusions regarding this new approach to weather modeling.

2.3. HRES

The High Resolution (HRES) deterministic forecasting model, developed by the ECMWF, represents the current cutting-edge in numerical weather prediction on a global scale (ECMWF, 2020b). This chapter provides a detailed examination of the architecture of HRES, focusing on its spectral transform method, hybrid vertical coordinates, time integration, physical parameterizations, and data assimilation system.

At the core, any NWP model attempts to solve the Navier-Stokes equations in order to predict the future atmospheric states. Solving complex partial differential equations numerically over the globe is an immense computational challenge due to the nonlinear nature of atmospheric dynamics and the vast range of spatial and temporal scales involved. The atmosphere exhibits processes ranging from small-scale turbulence and cloud microphysics to large-scale planetary waves, all of which interact with each other.

Numerical weather prediction models discretize the continuous equations of motion using numerical methods, dividing the atmosphere into a finite number of grid points or spectral components. The choice of discretization methods, grid resolution, and numerical schemes affects the model's ability to accurately simulate atmospheric processes and its computational efficiency. High-resolution models can capture finer details but require significantly more computational resources.

Another fundamental challenge in NWP is the accurate representation of sub-grid-scale processes, such as convection, radiation, and boundary layer turbulence, which cannot be resolved explicitly due to computational limitations. These processes are represented through physical parameterization schemes, which approximate their effects on the larger-scale flow. The development and tuning of these parameterizations are critical for model performance.

Furthermore, the sensitivity of the atmosphere to initial conditions necessitates precise and accurate initialization of the model state. Data assimilation techniques are employed to integrate observational data from various sources into the model, producing the best possible estimate of the current atmospheric state. However, observations are unevenly distributed spatially and temporally, and contain errors, which adds complexity to the assimilation process.

HRES model addresses these challenges through advanced numerical methods, sophisticated physical parameterizations, and a robust data assimilation system. In the following sections, we examine the architecture of HRES in detail, highlighting how it implements these components to achieve high accuracy in global weather forecasting.

2.3.1. Model Architecture

The Integrated Forecasting System (HRES) developed by ECMWF is a highly sophisticated numerical weather prediction (NWP) model designed to simulate the evolution of the Earth's atmosphere (ECMWF, 2020a). The architecture of HRES combines a spectral transform method for horizontal discretization, a hybrid vertical coordinate system, and a semi-Lagrangian semi-implicit (SLSI) time integration scheme, enabling efficient and accurate forecasting of large-scale atmospheric dynamics (ECMWF, 2020c).

The spectral transform method uses spherical harmonics to represent meteorological fields horizontally, offering computational efficiency by reducing the dimensionality of the problem (ECMWF, 2020a). This approach is well-suited to capturing large-scale atmospheric waves and dynamics. Non-linear terms, such as advection, are computed in grid-point space, allowing the model to handle both large-scale processes and smaller-scale interactions (ECMWF, 2020a).

Vertically, HRES uses a hybrid coordinate system, combining terrain-following coordinates near the surface with pressure-based levels aloft (ECMWF, 2020a). This enables accurate representation of near-surface processes and smooth atmospheric profiles at higher altitudes, crucial for regions with complex topography.

The semi-Lagrangian semi-implicit (SLSI) time integration scheme allows the model to take longer time steps without compromising numerical stability (ECMWF, 2020c). By handling fast-moving gravity waves implicitly and tracing air parcel trajectories backward in time, HRES efficiently handles fast-moving features like advection while maintaining computational efficiency.

2.3.2. Physical Parameterization Schemes

In the Integrated Forecasting System (HRES), sub-grid-scale processes that cannot be explicitly resolved by the model are represented through physical parameterization schemes (ECMWF, 2020d). These include parameterizations for radiation, convection, turbulence, clouds, and drag.

The radiation scheme simulates the transfer of solar and terrestrial radiation through the atmosphere, interacting with atmospheric constituents such as clouds, aerosols, and greenhouse gases (ECMWF,

2020d). These radiative processes are crucial for driving the diurnal cycle and global energy balance.

Convection, responsible for vertical transport of heat, moisture, and momentum, is parameterized to represent both shallow and deep convection (ECMWF, 2020d). The scheme models processes like entrainment, detrainment, and convective momentum transport, which affect the larger-scale momentum field and weather phenomena.

Turbulent transport, particularly in the atmospheric boundary layer, models the mixing of heat, moisture, and momentum due to small-scale turbulence (ECMWF, 2020d). HRES accounts for exchanges with different surface types such as land, oceans, and sea ice, which are critical for simulating near-surface weather phenomena.

Clouds and precipitation are modeled using prognostic equations for cloud water, ice, rain, and snow (ECMWF, 2020d). These schemes handle processes like condensation, evaporation, and freezing, providing accurate predictions of large-scale and convective precipitation.

Orographic and non-orographic drag are parameterized to represent momentum loss due to unresolved terrain features and gravity waves, influencing large-scale circulation patterns and atmospheric dynamics (ECMWF, 2020d).

2.3.3. Data Assimilation System

HRES employs a sophisticated data assimilation system known as 4D-Var (Four-Dimensional Variational Data Assimilation) to merge observational data with model forecasts, producing an optimal estimate of the current atmospheric state for forecast initialization (ECMWF, 2013).

The 4D-Var system minimizes discrepancies between the model's background state (short-term forecast) and available observations by solving a cost function over a specified time window (ECMWF, 2013). This approach integrates spatial and temporal information, allowing for the optimal assimilation of observations. A schematic of this process is shown in Figure 2.5, where the background forecast (blue dashed line) is adjusted through the assimilation process to yield an analysis (solid red line) that lies between the background forecast and observational data. This analysis plays an important role in this research, refer to subsection 3.2.4 for further details.

The data assimilation system ingests a broad range of observational data, including satellite radiances, radiosonde measurements, surface stations, aircraft observations, and ground-based remote sensing instruments (ECMWF, 2020b). Before assimilation, these data undergo quality control and bias correction, with satellite radiances providing essential global coverage. The adjustments to the model's initial conditions enhance forecast accuracy. However, as illustrated by the distribution of observational data in Figure 2.5, assimilation quality may vary across different locations due to differences in observational coverage.

2.3.4. Forecast Generation Process

The forecast generation process in HRES begins with the preparation of initial conditions through data assimilation. The 4D-Var system combines observational data with the model's background state to create an analysis, which serves as the best estimate of the current atmospheric state.

Once the initial conditions are prepared, HRES integrates them forward in time using its dynamical core. The raw model output is post-processed to generate useful forecast products, converting data from spectral space back to a latitude-longitude grid.

Post-processing generates gridded fields of essential meteorological variables such as temperature, wind speed, humidity, and precipitation. Derived variables like Convective Available Potential Energy (CAPE) are also calculated to provide insights into specific weather phenomena. Bias correction methods and downscaling techniques are applied to refine the forecast output, improving accuracy at finer spatial scales.

2.3.5. Model Performance and Evaluation

The performance of HRES is evaluated through metrics such as Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC), which assess forecast accuracy relative to real-world observations.

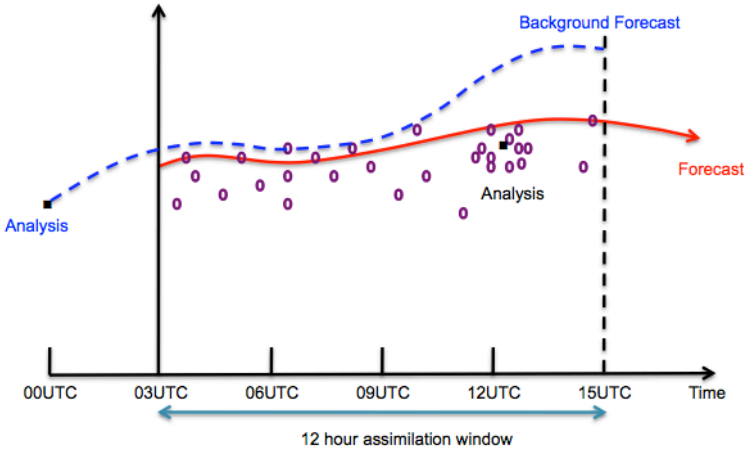


Figure 2.5: The ECMWF 4-dimensional data assimilation system calculates an adjustment to the background forecast (blue dashed line) to produce an analysis (solid red line) that balances both the background and the observations, effectively acting as a weighted average of the two (Owens, 2022).

RMSE provides an estimate of forecast error, sensitive to large discrepancies between forecasted values and observations. ACC compares the spatial pattern of forecast anomalies with observations, particularly for mid-range to long-term forecasts. HRES typically performs well in medium-range forecasts (up to 10 days), with high ACC scores and competitive RMSE values compared to other global models.

2.3.6. Limitations and Future Work

Despite its strong performance, HRES faces limitations related to resolution and computational cost. Higher resolution models capture fine-scale weather features more accurately but require significantly more computational resources.

Parameterization schemes for sub-grid-scale processes, such as convection and turbulence, introduce uncertainty, particularly in predicting heavy precipitation or tropical cyclone intensity (ECMWF, 2020d). Further developments in ensemble approaches, uncertainty quantification, and data assimilation could enhance HRES's performance.

Future work on HRES may focus on improving parameterization schemes, increasing forecast resolution, and advancing coupled modeling systems to integrate ocean, land surface, and cryosphere models, enhancing both weather prediction and climate simulations (ECMWF, 2020e, 2023b).

The ECMWF's research efforts are also increasingly directed towards data driven models with the first promising results shown with their AIFS model (Lang et al., 2024).

2.4. Aspire: LES and Mesoscale Models

The Aspire framework is a state-of-the-art modeling platform designed to simulate both fine-scale turbulent structures of the atmosphere and larger mesoscale atmospheric dynamics. Aspire was developed as an extension of the GRASP framework, which in turn originated from the Dutch Atmospheric Large Eddy Simulation (DALES) model (Siebesma, Wyszogrodzki, et al., 2010). DALES itself was built to study the dynamics of the atmospheric boundary layer (ABL), and Aspire further refines this approach to be useful in both research and practical applications, particularly for renewable energy. By integrating Large Eddy Simulation (LES) with a mesoscale model, Aspire can represent complex atmospheric processes across a wide range of scales, making it highly suitable for wind energy applications (Whiffle, 2024).

In the following sections we will highlight how the Aspire model works and describe and the key as-

sumptions underlying it. To this end we start by describing the Atmospheric Boundary Layer (ABL) and the role of turbulence.

2.4.1. Atmospheric Boundary Layer (ABL)

The atmospheric boundary layer (ABL) is the lowest part of the atmosphere, directly interacting with the Earth's surface (Stull, 1988). Turbulence plays a dominant role in the ABL, where energy is transferred through eddies of various sizes. The energy cascade, first introduced by Richardson (Richardson, 1922), describes how energy is injected into the atmosphere at large scales and transferred to smaller scales until it dissipates at the viscous subrange (Kolmogorov, 1941). This turbulent behavior must be accurately captured to understand surface-layer interactions and to model phenomena like wind behavior near wind farms.

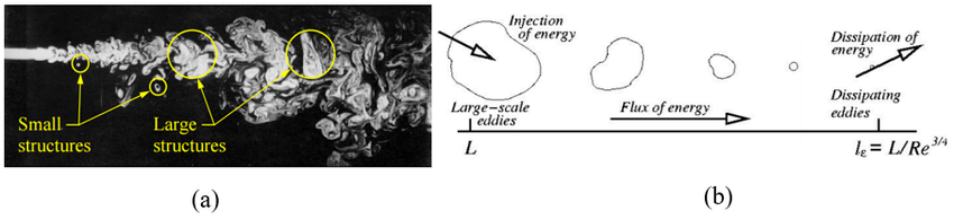


Figure 2.6: (a) Turbulent Flow Structures (b) The Energy Cascade according to Richardson (1922)

As shown in Figure 2.6, turbulent flow structures vary from large eddies, where energy is injected, to small eddies where energy is dissipated. The correct representation of this energy cascade is critical for high-fidelity atmospheric modeling.

2.4.2. Large Eddy Simulation (LES)

LES is a core component of Aspire and resolves the large eddies in the atmosphere while modeling the effects of smaller eddies. The main assumption behind LES is that the large, energy-containing eddies are directly simulated, while the smaller, sub-grid-scale (SGS) eddies are parameterized. This constitutes a stark difference with NWP models such as HRES, which parameterizes *all* turbulence. This approach allows LES to capture the most important turbulent structures without the prohibitive computational cost of resolving all scales.

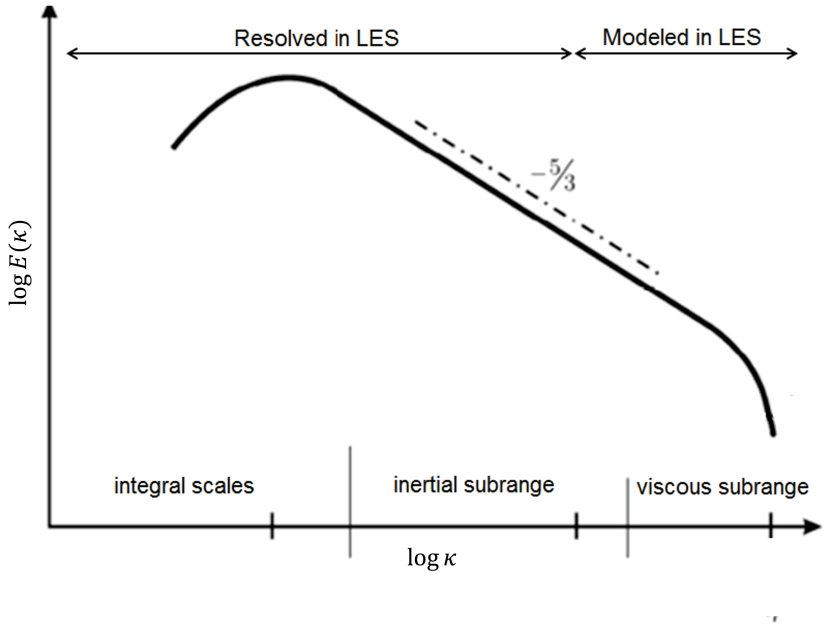


Figure 2.7: Energy Spectrum showing resolved and modeled eddies in LES

Figure 2.7 illustrates how Aspire-LES resolves the large eddies, capturing the inertial subrange, while smaller, unresolved eddies are modeled using the SGS approach. This balance allows Aspire to focus computational resources on the large, energy-containing structures, which are crucial for accurate simulation of the ABL.

Assumptions in Aspire-LES

Aspire's LES model is built on several key assumptions that balance physical realism with computational efficiency:

- **Anelastic Approximation**: Aspire employs the anelastic approximation to simplify the governing equations by filtering out sound waves, which are not of primary interest in atmospheric flows but can impose restrictive time-step limits if included. By assuming a base-state reference density that varies only with height, denoted $\rho_0(z)$, and neglecting variations in density associated with small-scale pressure fluctuations, the model avoids the need to solve for acoustic modes, allowing for larger time steps and more efficient simulations. This assumption is justified for atmospheric flows where vertical density stratification is significant, but local density variations due to small pressure perturbations are negligible.
- **Filtered Navier-Stokes Equations**: Aspire filters the governing Navier-Stokes equations spatially, separating the large, energy-containing eddies from the smaller, dissipative eddies. This filtering process removes small-scale turbulence that cannot be resolved on the computational grid.
- **Subgrid-Scale (SGS) Modeling**: Aspire can use various SGS models to parameterize the unresolved scales of turbulence, including the Smagorinsky model, Rozema, Vreman model, and Verstappen subgrid model. Each model makes distinct assumptions about turbulence and is suited for different flow conditions.
- **Boundary Conditions**: Aspire employs different boundary condition strategies, such as periodic boundary conditions for spatially homogeneous flows and non-periodic boundary conditions for complex environments. Surface boundary conditions are based on Monin-Obukhov similarity theory (MOST), which relates surface fluxes to local shear stress and buoyancy flux.

These assumptions allow Aspire to balance computational efficiency with accurate turbulence representation in atmospheric boundary layer modeling.

Handling of Subgrid-Scale (SGS) Turbulence

Aspire parameterizes SGS turbulence using various closure models, each with distinct assumptions about turbulence:

Rozema Model The Rozema model, part of the minimum-dissipation class of subgrid-scale (SGS) models, aims to achieve optimal energy dissipation in large-eddy simulation (LES) by minimizing the eddy viscosity required to dissipate energy from unresolved scales. Unlike traditional eddy-viscosity models, which often lead to excessive dissipation, the Rozema model uses the invariants of the rate-of-strain tensor to ensure dissipation only when necessary. This approach provides accurate energy dissipation in turbulent flows, avoids over-damping in laminar flows, and adapts well to anisotropic grids without requiring a filter-width approximation. The model has shown effectiveness in simulating complex flows, including turbulent mixing layers and channel flows, enhancing LES performance across varying grid types. (Rozema et al., 2015)

2.4.3. Model Architecture and GPU Parallelization

LES is computationally expensive because of the high grid resolution and fine temporal scales required to resolve large eddies. Aspire addresses these computational challenges by using GPU acceleration to solve the filtered Navier-Stokes equations efficiently (Schalkwijk et al., 2015). The model leverages fast, GPU-optimized solvers, including a pressure solver to handle the anelastic condition, which allows for real-time simulation and visualization of atmospheric fields.

Numerical Stability and Time-Stepping Aspire adheres to the Courant-Friedrichs-Lewy (CFL) condition for numerical stability. In simple terms, the CFL condition states that the time step in a simulation

must be short enough that information (like a wave or particle movement) doesn't travel more than one grid cell in a single time step. This means that the model sees changes happening gradually from one grid cell to the next, rather than jumping over several cells at once.

If this condition is not met, the simulation can become unstable, resulting in errors, unrealistic results, or even complete breakdown of the model. This is particularly important in high-resolution simulations where small time steps are required to capture the fast dynamics of small eddies. Additionally, numerical dissipation and stability are maintained through careful selection of numerical schemes and time-stepping methods, which are optimized for GPU architectures.

By addressing numerical stability and leveraging GPU parallelization, Aspire can handle large-scale simulations efficiently, making it suitable for practical applications like wind farm modeling, where detailed simulations are needed over long time periods.

2.4.4. Boundary Conditions and Data Coupling

Accurate boundary conditions are essential for LES models like Aspire to simulate the correct atmospheric state. Aspire employs a range of boundary condition strategies, each based on assumptions about the behavior of the atmosphere at the edges of the computational domain:

Periodic Boundary Conditions In scenarios where the atmospheric flow is considered spatially homogeneous, such as over flat terrain or open ocean, Aspire assumes periodic boundary conditions in the horizontal directions. This means that the flow leaving one side of the domain re-enters from the opposite side, effectively creating an infinite domain. This assumption is valid for large-scale turbulence in uniform environments where flow structures on either side of the domain are statistically similar.

Non-periodic Boundary Conditions In more complex environments involving terrain, obstacles, or varying atmospheric conditions, Aspire supports non-periodic boundary conditions. These include inflow and outflow conditions, where the model assumes that turbulence is generated upstream and propagates downstream. This approach requires assumptions about the upstream flow being known or prescribed, often based on observational data or outputs from larger-scale models. The model may also use radiation boundary conditions at the top of the domain to prevent artificial reflection of gravity waves, assuming that such waves can propagate out of the domain without reflection.

Surface Boundary Conditions At the lower boundary, Aspire uses Monin-Obukhov similarity theory (MOST) to describe surface fluxes of momentum, heat, and moisture. MOST assumes that turbulent fluxes near the surface are related to the local shear stress and buoyancy flux, which depend on surface roughness and stability parameters. The model incorporates surface energy balance calculations, which are crucial for capturing the effects of land-surface interactions, such as heat fluxes over land or sea-surface fluxes over oceans.

Data Coupling with Mesoscale and Global Models In Aspire, the LES domain is typically nested within a mesoscale model, which acts as an intermediary between the LES and global weather data/models such as ERA5 or IFS HRES. These global models provide the ”real” weather state outside of the LES domain, ensuring that the LES receives appropriate atmospheric information at its boundaries. The mesoscale model bridges the gap between the global models and the fine-scale LES, ensuring that the LES receives realistic atmospheric forcing. Two-way coupling ensures that the mesoscale model informs the LES while the LES provides feedback to the mesoscale model for improved accuracy near the surface.

By integrating various boundary condition strategies and data coupling methods, Aspire can simulate atmospheric flows accurately across different scales and environments.

2.4.5. Aspire Mesoscale Model

The mesoscale component of Aspire is designed to simulate larger-scale atmospheric phenomena that extend beyond the capabilities of Large Eddy Simulation (LES). Operating at a coarser resolution over larger geographical areas and longer time periods, it focuses on parameterizing turbulence effects rather than resolving individual turbulent eddies. By using the anelastic approximation—similar

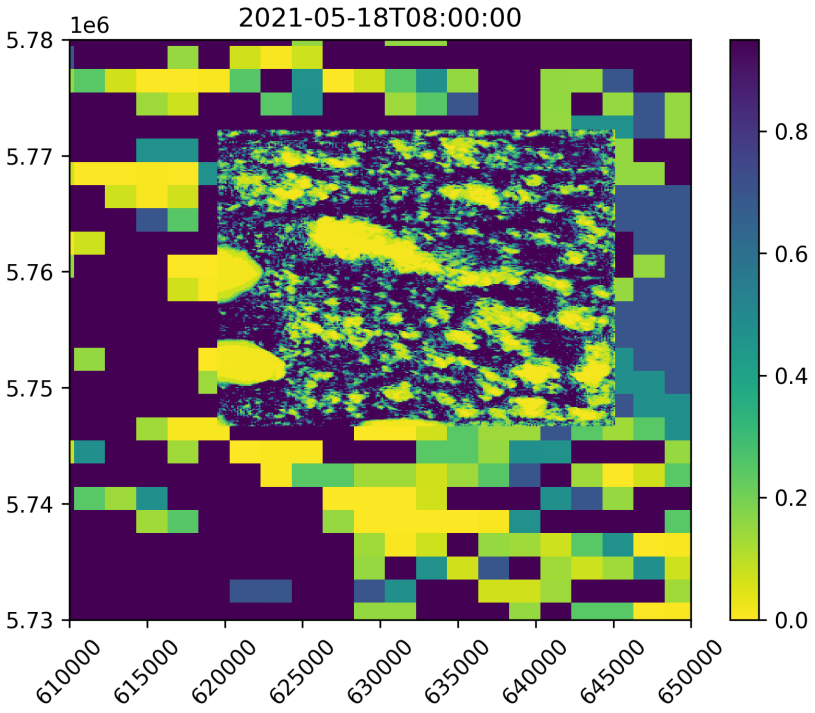


Figure 2.8: Illustration of how a mesoscale simulation bridges the gap between boundary conditions and a high resolution LES simulation.

to LES—the mesoscale model filters out sound waves to simplify the governing equations. However, it differs by not explicitly resolving turbulence; instead, it employs closure schemes to efficiently simulate large-scale atmospheric processes.

Due to its coarse grid resolution, the mesoscale model entirely parameterizes turbulence, distinguishing between horizontal and vertical mixing to account for the anisotropic nature of atmospheric turbulence at larger scales. Horizontal mixing is parameterized using schemes akin to the Rozema model applied at the mesoscale, accounting for the spreading and mixing of atmospheric properties like momentum and heat over vast horizontal distances. Vertical mixing utilizes mixing-length approaches dependent on atmospheric stability, with stability corrections based on the Richardson number to consider the balance between buoyancy and shear.

By parameterizing turbulence, the mesoscale model avoids the computational expense of resolving small-scale motions, enabling it to simulate atmospheric processes over larger domains and longer timescales. This efficiency makes it ideal for simulating regional weather patterns such as fronts, cyclones, and jet streams. It provides the broader atmospheric context necessary for nested LES simulations by supplying realistic boundary conditions and large-scale forcing.

The mesoscale model supplies LES with essential information about the larger-scale atmospheric environment, including background wind profiles, temperature, pressure gradients, and other synoptic-scale features. The coupling between the mesoscale model and LES ensures consistency across scales. The mesoscale model provides time-varying boundary conditions for the LES domain, ensuring that small-scale simulations are influenced by realistic atmospheric conditions. Conversely, LES can feed back information to the mesoscale model, particularly near the surface, improving the overall accuracy of the simulation by accounting for fine-scale processes that affect larger scales.

2.4.6. Applications to Renewable Energy

Aspire's most significant application is in wind energy, where it models the complex interactions between the atmosphere and wind turbines. Using LES, Aspire can resolve the fine-scale turbulence that affects wind turbine performance, allowing for more accurate wind farm simulations. A key feature of Aspire's

LES is its integration of the actuator disk model, which represents wind turbines as porous disks that extract momentum from the flow (Sørensen, 2022).

2.4.7. Limitations and Future Developments

While Aspire is a powerful tool, it has limitations. One of the primary challenges is the high computational c ost associated with LES, even with GPU acceleration. Moreover, the accuracy of the subgrid-scale models used in LES remains an area of active research, particularly in highly complex flows, such as those in wind farms. Future developments in Aspire are likely to focus on improving SGS models, incorporating better boundary condition treatments, and improve wind farm modelling capabilities.

2.5. Summary

This chapter reviewed the four weather forecasting models used in this study: GraphCast, Pangu-Weather, HRES, and Aspire, each employing different methodologies for atmospheric prediction.

GraphCast utilizes Graph Neural Networks with an encoder-processor-decoder architecture and a multimesh representation to capture local and global atmospheric interactions efficiently. It delivers accurate medium-range forecasts with high computational efficiency by learning from extensive datasets.

Pangu-Weather introduces a 3D Earth-specific transformer architecture that models both vertical and horizontal dependencies of atmospheric variables. By incorporating Earth-specific biases and a hierarchical temporal aggregation strategy, it achieves high accuracy in medium-range forecasts with remarkable speed, outperforming traditional operational models.

HRES, developed by ECMWF, represents a traditional numerical weather prediction approach. It employs a spectral transform method, hybrid vertical coordinates, and comprehensive physical parameterizations to simulate atmospheric processes. Despite being computationally intensive, HRES provides reliable forecasts through detailed modeling and data assimilation.

Aspire combines Large Eddy Simulation with mesoscale modeling to simulate atmospheric processes across various scales. By resolving large turbulent eddies and parameterizing smaller scales, it offers high-fidelity simulations crucial for applications like wind energy. GPU acceleration enhances its computational efficiency.

In essence, these models highlight the diversity in weather forecasting methodologies. Machine learning models like GraphCast and Pangu-Weather offer efficient and accurate alternatives to traditional numerical models. HRES continues to be a benchmark for operational forecasting due to its thorough physical modeling, while Aspire provides detailed simulations for specific applications. Understanding each model's strengths and limitations is key to advancing weather prediction technologies.

3

Research Design, Datasets and Methodology

This chapter will cover the research design and experimental setup, as well as the evaluation methods used. First, a brief overview will be provided, followed by a detailed description of the datasets used in this study. We then describe the specific settings of the weather models used in our analysis. The chapter concludes with the evaluation and statistical methods to be used in the comparison of the models.

3.1. Research Design

The main goal of this research is to evaluate AI based weather forecasting models in operational setting against observations, and judge their performance against the current state of the art in numerical weather forecasting. The Netherlands was chosen because this area is rapidly building out wind power infrastructure and dense observation data are available for both onshore and offshore regions in the form of Royal Netherlands Meteorological Institute (KNMI) Surface Synoptical Observation (SYNOP) data. The AI models chosen were GraphCast by Google DeepMind Lam et al., 2023a and Pangu-Weather by Bi et al., 2023. We will compare these against the state-of-the-art in numerical weather prediction: the High Resolution (HRES) model by the European Center for Medium-Range Weather Forecasting (ECMWF). Additionally, the Aspire model developed by Whiffle was used, both in mesoscale setting for the comparison across all weather station and in turbulence resolving LES setting for two specific station locations.

In order to perform a fair comparison, an entire year of forecasts were generated for each of the models. The year chosen was 2022, for the following reasons:

- It is not in the training datasets for Graphcast and Pangu-Weather, a core requirement for evaluating machine learning models. This ensures the models can show whether they have learned the underlying patterns in the data and do not overfit.
- HRES forecast data was readily available in the right resolution from Weatherbench2 (Rasp et al., 2024).
- SYNOP data from the KNMI was available for 47 stations in the European Netherlands, allowing for assessment in the geographical region of interest.

In the field of atmospheric science, weather forecasting is divided into short-range, medium-range, and long-range categories, each characterized by distinct timescales and methodologies. Short-range forecasts, typically spanning up to 72 hours, offer high precision due to the reliance on high-resolution models and real-time observational data, making them essential for immediate operational decision-making. Medium-range forecasts, covering 3 to 7 days, focus on broader atmospheric patterns, but their accuracy diminishes as uncertainties compound over time. Long-range forecasts, extending beyond a

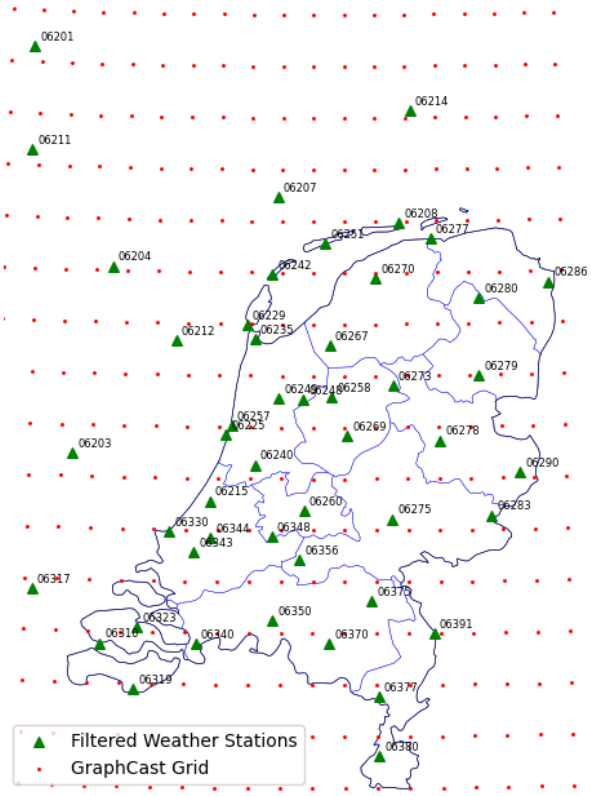


Figure 3.1: Map of KNMI SYNOP stations in the Netherlands shown in green. The red dots represent the 0.25° grid point spacing of GraphCast, Pangu-Weather and the downsampled HRES data.

week to up to 30 or more days, provide probabilistic insights into general trends and patterns, relying on lower resolution models and climate drivers such as the El Niño-Southern Oscillation.

The choice of a 96-hour (4-day) forecast window for this study represents a strategic balance between computational efficiency and the ability to capture the breakdown of forecast skill in various weather models. Short-range models, particularly Large Eddy Simulation (LES) models, are computationally expensive but offer high precision within the first few days. By selecting a 4-day lead time, this study allows for a robust assessment of both short-term accuracy and the point at which model performance begins to degrade. On the other hand, a 96-hour window does allow the medium range models to show their skill on this longer horizon.

The data were analyzed both in aggregate and along various spatial and temporal dimensions, in order to clearly highlight the different strengths and weaknesses of each model and the possible underlying causes for those differences.

3.2. Datasets

Observation data selection was based on location, quality, and availability. Once the Netherlands was chosen as the region of interest, two main options emerged: a very high-quality dataset extending up to 200 m in the vertical domain from the Cabauw meteorological mast, and a large set of surface-level observations from the KNMI's network of SYNOP stations. Given the global nature of the main models considered, the decision was made to use the SYNOP observations. This choice allows the global models, such as GraphCast and HRES, to demonstrate their generalization skills while enabling us to highlight specific differences in skill across different stations. Furthermore, the relatively dense distribution of stations (47 stations across the European Netherlands, both onshore and offshore) made it feasible to perform a mesoscale run with the Aspire model.

3.2.1. SYNOP Observation Data

The SYNOP dataset contains a total of 56 stations. Out of these, 3 stations lacked location data and were therefore discarded. From the remaining 53 stations, two were filtered out to limit the latitude extent of the area, aiming to decrease the computational cost of the mesoscale Aspire run. Another 3 stations were excluded due to low data quality, as they contained a disproportionate number of NaN-valued data points. This process left us with 50 stations: 47 in the European Netherlands and 3 in the Caribbean. For the purposes of this analysis, we focus on the stations in the European Netherlands. A map showing the locations of these weather stations is presented in Figure 3.1.

3.2.2. SYNOP Processing

After filtering the dataset to include only the stations of interest, the following processing steps were applied. The original data contains a wide range of variables, measured at 10-minute intervals. First, the dataset was reduced to the variables of interest, and the 10 m u and v components of wind were calculated by decomposing the wind speed and direction. Conveniently, the dataset already included the variable R6H, which represents the accumulated rainfall over the last 6 hours.

Next, the data was processed in the time domain. The dataset was downsampled to 6-hour intervals without averaging. This approach ensured that the observation data retained the true measurements at the specific moments of interest (UTC00, UTC06, UTC12, and UTC18). While averaging could create a fairer comparison by smoothing out short-term fluctuations, we opted to keep the observations as close to the ground truth as possible, leaving any smoothing operations to the models.

After all adjustments, the following variables remained:

- 1.5 m ambient temperature (° C), 10-minute average;
- Air pressure at sea level (hPa), 1-minute average;
- 10 m wind speed (m/s), 10-minute average;
- 10 m wind direction (°), 10-minute average;
- 10 m u-component of wind (m/s), 10-minute average;
- 10 m v-component of wind (m/s), 10-minute average;
- Rainfall in the last 6 hours (mm), accumulated.

It is important to note that the temperature measurements are taken at 1.5 m above ground level, whereas all other models produce temperature at 2 m height. No adjustments were made to the SYNOP data to correct for this difference, in order to maintain the integrity of the original observations. Any errors resulting from this height discrepancy will be considered during the analysis.

Additionally, the actual wind speed and direction sensors are usually not located exactly at 10 m above ground level. The KNMI interpolates these measurements to 10 m, and we used these interpolated values to ensure consistency across stations and to minimize the introduction of errors. The same approach applies to the conversion of air pressure measurements to sea level pressure.

Lastly, the SYNOP dataset available for this study started on 12-01-2022 instead of 01-01-2022. Consequently, all forecasts were generated for the period from 12-01-2022 until 31-12-2022.

3.2.3. ERA5 Reanalysis Dataset

Reanalysis datasets are fundamental in meteorology and climate science, providing comprehensive and consistent records of the Earth's atmosphere by assimilating historical observational data into numerical weather prediction (NWP) models. The **ERA5 dataset** is the fifth-generation global reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) as part of the Copernicus Climate Change Service (C3S) (Hersbach et al., 2020). It offers hourly estimates of a wide range of atmospheric, land, and oceanic variables from 1950 onwards, with a native horizontal resolution of approximately 31 km (0.25°).

ERA5 serves as the primary training dataset for the AI-based weather forecasting models GraphCast and Pangu-Weather. By providing a detailed and physically consistent representation of the historical

atmospheric state, ERA5 enables these models to learn complex patterns and relationships within the climate system.

However, it is important to acknowledge that ERA5 is not a perfect ground truth. While it assimilates a vast array of observations, the reanalysis process involves model assumptions and simplifications that can introduce errors and biases. These limitations may affect the performance of AI models trained on ERA5 data, particularly when applied to local scales or specific regions.

3.2.4. ECMWF Analysis Dataset

An additional dataset used in this study is the **ECMWF Analysis dataset**. Similar to reanalysis datasets, the Analysis dataset combines available observations with previous forecasts through a data assimilation process to produce the best estimate of the atmospheric state at a specific time, typically referred to as $t = 0$. This is achieved using the High-Resolution Forecast (HRES) model, which, as discussed in subsection 2.3.3, utilizes the 4D-Var data assimilation system to optimally incorporate current observational data with the model's background state. This process ensures that the dataset remains physically consistent while assimilating as much real-time observational data as possible.

The Analysis dataset differs from ERA5 in several key aspects:

- **Temporal Production:** The Analysis dataset is generated operationally in real-time, whereas ERA5 is a historical reanalysis product.
- **Data Availability:** Because the Analysis is produced in real-time, not all observational data may be available at the time of assimilation, potentially leading to a greater reliance on the model background and introducing different errors compared to ERA5.
- **Resolution and Variables:** The Analysis dataset used in this study is derived from the HRES forecast at 0 lead time, obtained from WeatherBench 2 (Rasp et al., 2024). It has a horizontal resolution of 0.25° (approximately 31 km) and includes variables consistent with those used by GraphCast (see Table 3.2 and Table 3.3), except for precipitation data, which is not available.

In the context of this study, the Analysis dataset serves three main purposes:

1. **Model Finetuning:** Following training on ERA5, GraphCast is finetuned on Analysis data from 2016-2021.
2. **Model Initialization:** It is used to initialize the GraphCast and Pangu-Weather models in operational settings (see subsection 3.4.1 and subsection 3.4.2).
3. **Comparison Benchmark:** Including the Analysis dataset in our comparisons allows us to evaluate the difference between this 'best estimate' and the actual SYNOP observations (see section 3.3). It also enables us to assess whether models are better at forecasting the Analysis dataset versus actual observations.

Understanding the differences between ERA5 and the Analysis dataset is crucial, as they both play significant but distinct roles in weather forecasting and in this study. By incorporating data assimilation as described in subsection 2.3.3, the Analysis dataset provides a high-quality, real-time estimate of atmospheric conditions that is vital for model finetuning, initialization, and benchmarking.

3.2.5. Comparison of Datasets

To enhance clarity, Table 3.1 summarizes the key differences between the SYNOP observations, ERA5 reanalysis, and the ECMWF Analysis dataset.

Table 3.1: Comparison of datasets used in this study.

Dataset	Temporal Resolution	Spatial Resolution	Role in This Study
SYNOP	10 minutes	Station locations	Ground truth for model evaluation
ERA5	Hourly	0.25° (∼30 km)	Training data for AI models
Analysis	6-hourly	0.25° (∼30 km)	AI fine-tuning and model initialization

3.2.6. Understanding the Limitations

Both the ERA5 and Analysis datasets are essential for weather forecasting but are not perfect representations of the ground truth. Their limitations stem from:

- **Model Dependence:** Both datasets rely on NWP models, which introduce errors due to approximations in model physics and numerical methods. Additionally their reliance on NWP models comes with the significant computational cost associated with running such models to create these datasets.
- **Data Assimilation Timing:** The Analysis dataset may have larger errors than ERA5 because it is produced in real-time and may not include all observational data that become available later.
- **Resolution Differences:** Differences in spatial and temporal resolutions can affect the accuracy and comparability of the datasets. While both datasets are used at a 0.25° grid in this study, their native resolutions and the methods used to produce them differ.

In the following section, we will perform an error and bias analysis of both the ERA5 and Analysis datasets against the SYNOP observations. This analysis will highlight the baseline errors inherent in these datasets and provide context for interpreting the performance of the forecasting models.

3.3. Baseline Error Assessment

In this section, we assess the baseline errors inherent in the ERA5 and ECMWF Analysis datasets by comparing them against SYNOP observations. Although both ERA5 and Analysis datasets are considered as representations of the atmospheric state and are used extensively for model training and initialization, they are not perfect ground truths. Understanding the magnitude and characteristics of their errors is essential for interpreting the performance of forecasting models that rely on them.

We evaluate the Mean Absolute Error (MAE) and Bias as functions of lead time for key meteorological variables: 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total 6-hour precipitation. The MAE provides a measure of the average magnitude of errors without considering their direction, while Bias indicates the average tendency of the dataset to overestimate or underestimate the observations. More on these methods and how they are calculated in subsection 3.6.2.

Figures 3.2 and 3.3 present the MAE and Bias, respectively, as functions of lead time for the ERA5 and Analysis datasets compared to SYNOP observations. Each line represents the error metric for a specific dataset, averaged across all station locations. Keep in mind that the notion of lead time does not apply to these datasets as all points are $t = 0$ best estimates following data assimilation. Consider the temporal aspect of the figures to indicate variation in performance at different moments throughout the day.

3.3.1. Analysis of Mean Absolute Error

The MAE plots in Figure 3.2 reveal several important observations across the key meteorological variables. For *2 m temperature*, both ERA5 and Analysis datasets exhibit a consistent average MAE of approximately 0.8° C relative to the SYNOP observations, with ERA5 marginally outperforming the Analysis dataset across all lead times. A diurnal pattern is evident in the errors, with higher MAE observed at lead times corresponding to 0, 24, 48, 72, and 96 hours, and lower MAE at 12, 36, 60, and 84 hours. This suggests larger errors during certain times of the day, potentially related to diurnal temperature variations.

For *mean sea level pressure*, the average MAE is around 0.3 hPa for both datasets, with the Analysis dataset slightly outperforming ERA5. Similar to temperature, a diurnal pattern is present in the errors.

Regarding *10 m wind speed*, both datasets show an average MAE of approximately 1.1 m/s when compared to SYNOP observations, indicating very similar performance and suggesting that neither dataset provides a clear advantage in terms of wind speed accuracy.

For *10 m wind direction*, the average MAE is around 35 degrees for both datasets, a relatively large error indicating significant challenges in accurately capturing wind direction at the 10 m level.

Finally, in the case of *total 6-hour precipitation*, only ERA5 provides data, with an average MAE of approximately 0.45 mm. The absence of Analysis data for precipitation precludes a direct comparison.

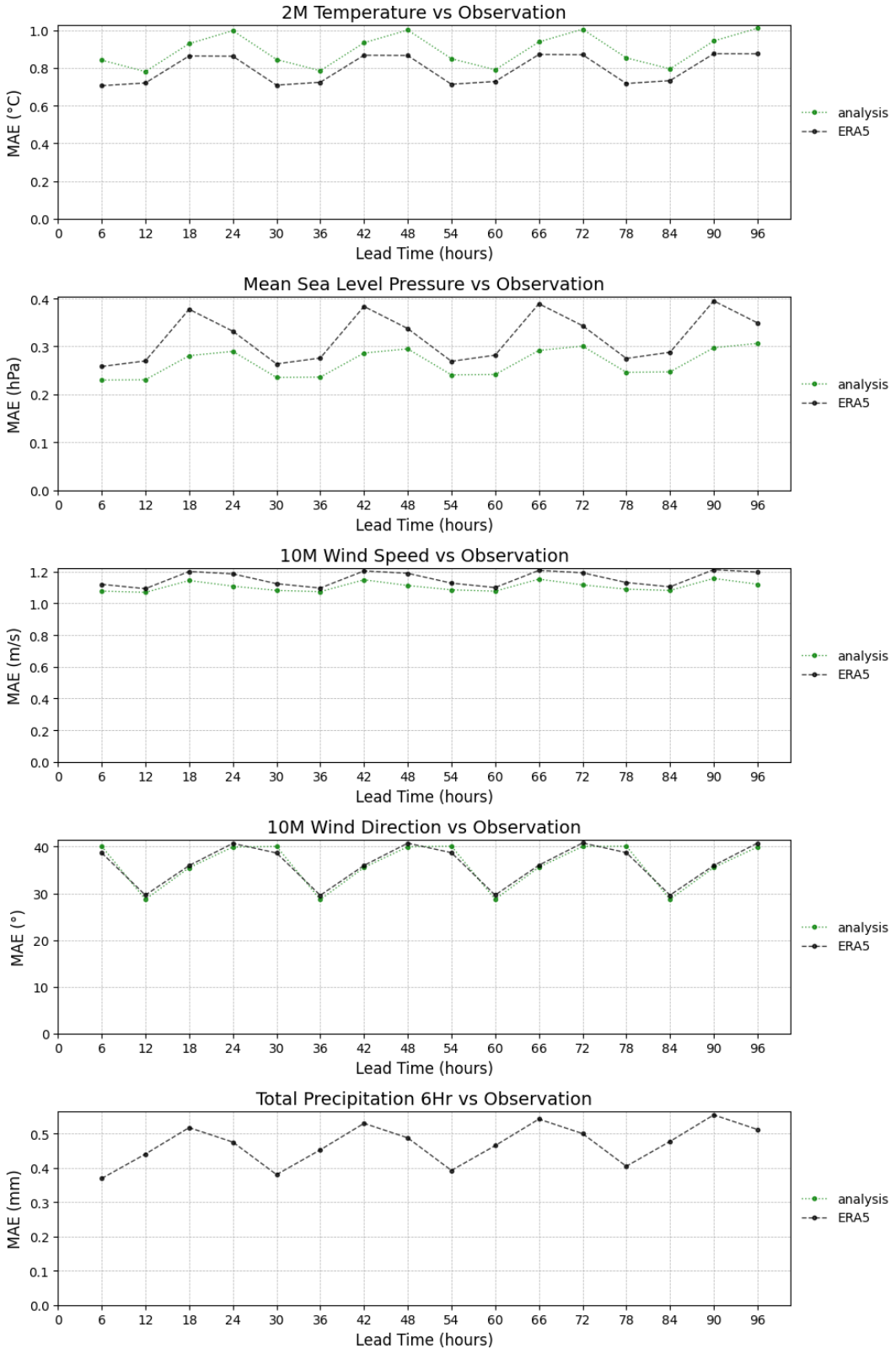


Figure 3.2: Mean Absolute Error (MAE) as a function of lead time for ERA5 and Analysis datasets against SYNOP observations. Variables include 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total 6-hour precipitation. Each line represents the MAE for a specific dataset, averaged across all station locations.

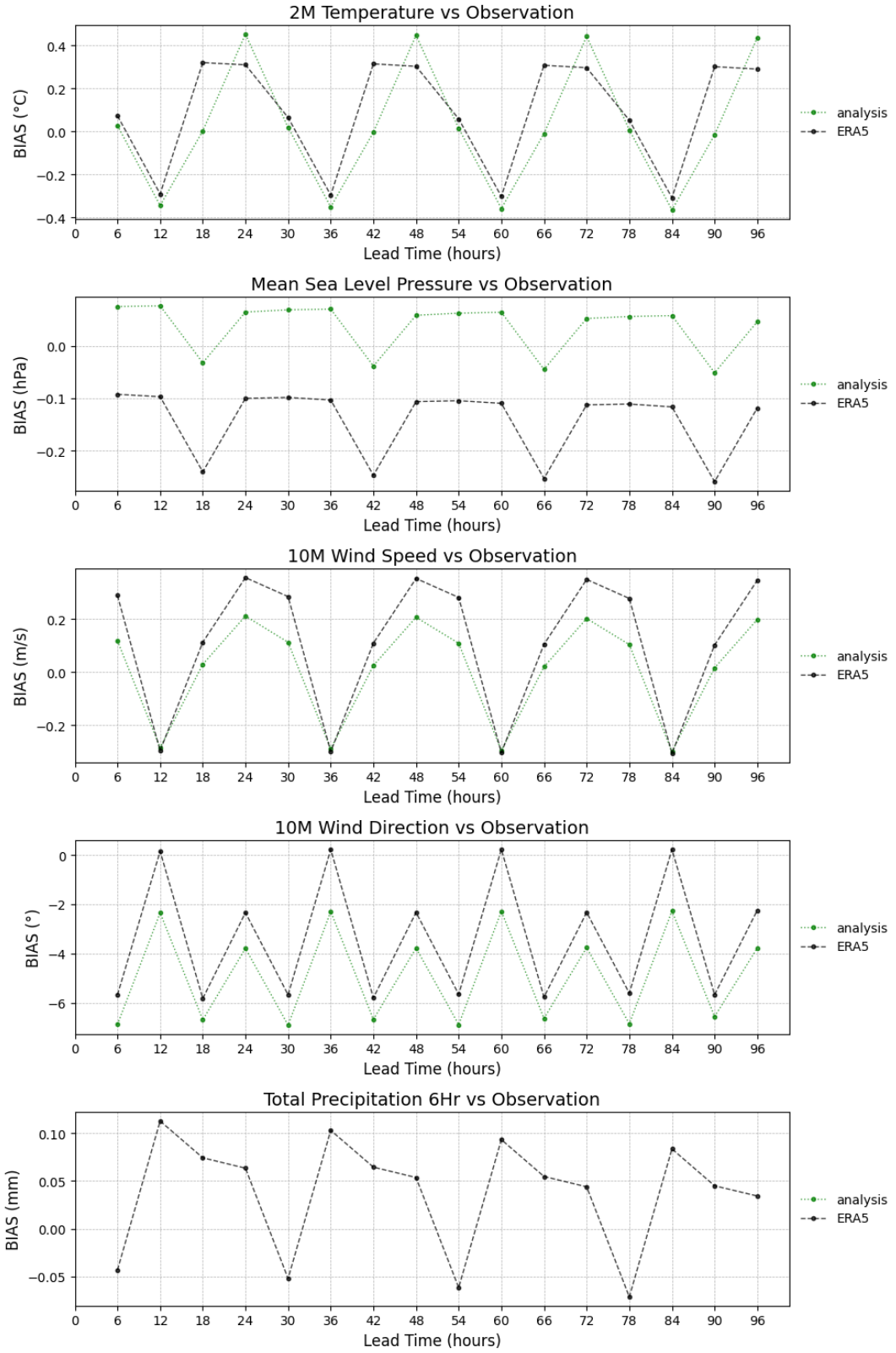


Figure 3.3: Bias as a function of lead time for ERA5 and Analysis datasets against SYNOP observations. Variables include 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total 6-hour precipitation. Each line represents the Bias for a specific dataset, averaged across all station locations.

3.3.2. Analysis of Bias

Figure 3.3 illustrates the Bias as a function of lead time for the datasets across the key variables. For *2 m temperature*, both datasets exhibit biases ranging between $-0.4\,°$C at lead times corresponding to 12-hour multiples and $+0.4\,°$C at 24-hour multiples. This oscillation suggests systematic overestimation and underestimation patterns aligned with the diurnal cycle.

In the case of *mean sea level pressure*, the Analysis dataset shows biases fluctuating between -0.5 hPa and $+0.5$ hPa, while ERA5 has a slightly narrower bias range between -0.25 hPa and -0.1 hPa. The consistent negative bias in ERA5 indicates a tendency to underestimate pressure compared to observations.

For *10 m wind speed*, both datasets have biases oscillating between -0.2 m/s and $+0.2$ m/s. The small magnitude of the bias suggests that, despite significant average wind speed errors (as indicated by the MAE), the errors are not systematically overestimating or underestimating wind speeds.

Regarding *10 m wind direction*, the biases range between 0 and -6 degrees, indicating a slight tendency to underestimate the wind direction angle. However, given the large MAE, this bias is relatively small.

For *total 6-hour precipitation*, ERA5 shows a bias ranging between -0.05 mm and $+0.1$ mm. The small magnitude of the bias compared to the MAE suggests that, although there are significant errors in precipitation amounts, they are not consistently overestimated or underestimated.

3.3.3. Spatial Patterns

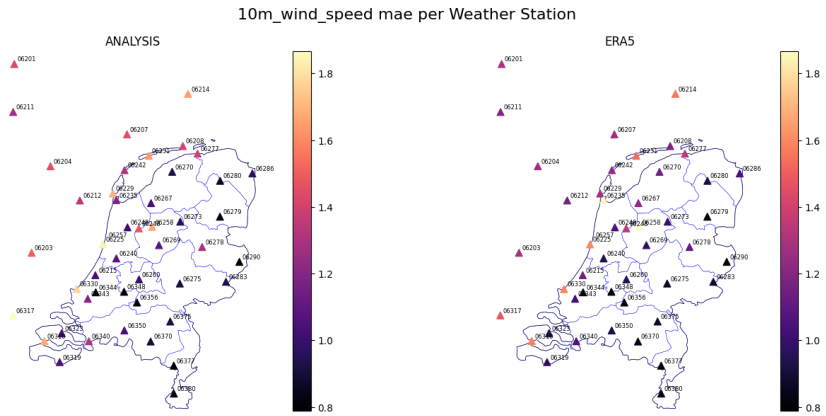


Figure 3.4: 10 m wind speed Mean Absolute Error (MAE) per weather station for the two datasets: Analysis (left), ERA5 (right). MAE values are averaged over all lead times and for the entire year of 2022, with darker colors representing lower error and lighter colors indicating higher error.

The spatial distribution of Mean Absolute Error (MAE) and Bias for 10 m wind speed, as illustrated in Figures 3.4 and 3.5, reveals several key patterns common to both the Analysis and ERA5 datasets. Higher errors are predominantly observed offshore, with onshore locations generally displaying lower MAE values. The transition zone between land and sea presents the highest errors, suggesting that complex boundary conditions and local interactions at this interface introduce additional challenges for accurate wind speed predictions.

Bias patterns further highlight distinct tendencies between offshore and onshore areas. Offshore, a consistent negative bias is apparent, indicating that both datasets tend to underestimate wind speeds in these regions, with a few exceptions where positive bias occurs. Conversely, onshore regions typically exhibit a positive bias, signifying a tendency to overestimate wind speeds, though some stations show exceptions to this trend. These observations underscore the influence of geographical and boundary characteristics on the performance of wind speed predictions.

3.3.4. Critical Assessment and Conclusions

The baseline error assessment highlights that both ERA5 and the Analysis datasets, while valuable for training and initializing models, exhibit significant errors when compared to actual observations. These

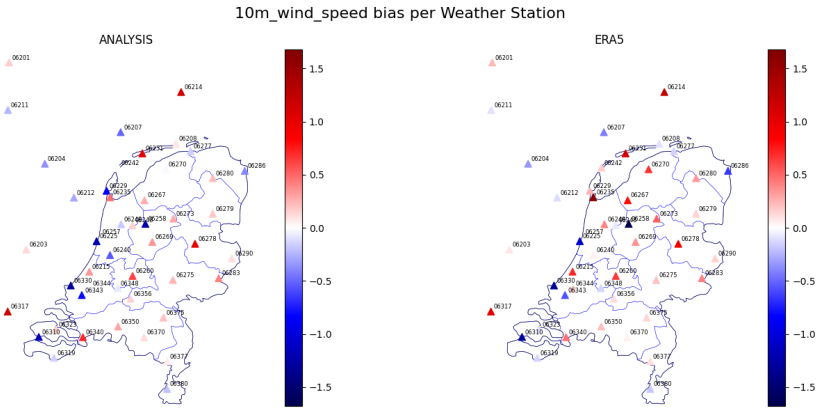


Figure 3.5: 10 m wind speed Bias per weather station for the two datasets: Analysis (left), ERA5 (right). Bias is averaged over all lead times and for the entire year of 2022. Cooler colors represent larger negative bias, while warmer colors represent higher positive bias.

errors are variable-dependent and exhibit temporal patterns, particularly diurnal cycles.

For *temperature and pressure*, the datasets perform relatively well, with MAE values within acceptable ranges for many applications. However, the diurnal patterns observed in both MAE and Bias indicate time-dependent errors, possibly due to model limitations in capturing diurnal cycles accurately.

In contrast, the errors in *wind speed and direction* are notably larger. An average MAE of 1.1 m/s for wind speed and 35 degrees for wind direction is significant, especially for applications like wind energy forecasting where precise wind information is critical. The relatively small biases suggest that these errors are more random rather than systematic. These significant baseline errors in wind speed and direction underscore the challenges in accurately forecasting wind resources. Since wind power generation is highly sensitive to wind speed—power output is proportional to the cube of wind speed—even small errors can lead to substantial discrepancies in energy production estimates.

These findings illustrate the limitations of datasets often considered as 'ground truth', which still contain inherent errors and biases. This is particularly important when these datasets are used to train AI models or initialize forecasts, as the errors can propagate through the modeling process. For applications sensitive to wind conditions, such as renewable energy forecasting, reliance solely on these datasets without accounting for their inherent errors may lead to suboptimal decisions.

Reducing baseline errors requires enhancements in both observational data quality and modeling techniques. Incorporating more high-resolution observational data and refining data assimilation methods may help in reducing these errors. Additionally, this emphasizes the need to assess model performance against observation data and not just against (re)analysis, which is currently the norm in AI-based weather forecasting.

Therefore, it is crucial to consider these baseline errors when evaluating the performance of forecasting models and to explore methods to mitigate their impact, such as incorporating local observational data or employing bias correction techniques.

In the subsequent chapters, the performance of various forecasting models will be assessed in light of these baseline errors, providing a more informed context for interpreting their relative accuracies.

3.4. Model Setup and Configuration

The following section will describe how each of the models was set up to generate forecasts. What data were used as initial/boundary conditions and what the model output and resolution were.

3.4.1. GraphCast

To generate forecasts with GraphCast, we use the operational model called GraphCast Operational (from here on referred to as GraphCast). This model has a horizontal resolution of 0.25°, resulting in

1440x721 grid points, meaning approximately one grid point every 31 kilometers. See Figure 3.1 for an overview of the grid spacing over the Netherlands. Vertically, the model is split between a set of surface variables, defined at specific heights above the surface, and atmospheric variables, distributed across 13 pressure levels from 50 hPa to 1000 hPa, as shown in Figure 3.6. The details of the surface and atmospheric variables in the model are provided in Table 3.2 and Table 3.3.

Table 3.2: Surface variables in the GraphCast model, defined at specific heights.

Surface Variable	Units
2m temperature	°C
10m u-component of wind	m/s
10m v-component of wind	m/s
Mean sea level pressure	hPa
Total precipitation (last 6 hours)	mm

Table 3.3: Atmospheric variables in the GraphCast model, available at multiple pressure levels.

Atmospheric Variable	Units
Temperature	°C
u-component of wind	m/s
v-component of wind	m/s
Geopotential	m
Specific humidity	g/g

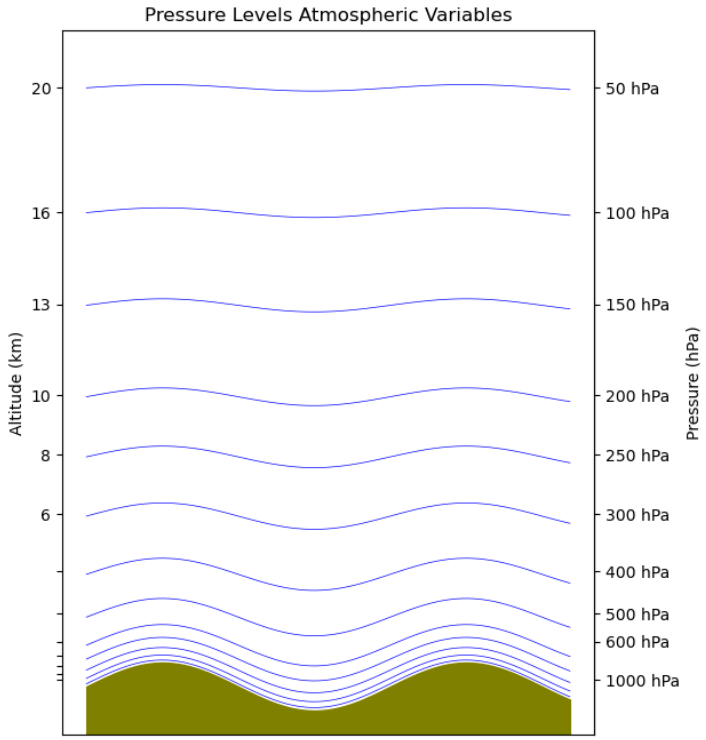


Figure 3.6: Pressure levels for atmospheric variables in the GraphCast model, ranging from 50 hPa to 1000 hPa.

Forecast Generation

The process of generating forecasts for 2022 was as follows:

1. Retrieve model weights from Google Deepmind's Google Cloud Services bucket [1];
2. Retrieve t=-6hr and t=0 HRES analysis data (see subsection 3.2.4) and initialize the model;
3. autoregressively generate forecasts up to 96 hours (16 steps) ahead;
4. Write forecasts to disk based on the region of interest.

A note to be made is that while GraphCast is a global model, with it's input, intermediate states and output all spanning the entire atmosphere, only a subset of the models outputs were kept in this study.

[1]The model weights are made available for use under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). You may obtain a copy of the License at: https://creativecommons.org/licenses/by-nc-sa/4.0/.

A single timestep at full resolution equates to $\sim$300MB, and conversely a 96 hour set of forecasts is $\sim$5GB. Since we are only interested in the a very small geographic subset it made sense to slice the data before writing to storage.

Computation Cost
Generating forecasts with GraphCast was done on 4 Nidia A100 40GB GPUs. Every forecast could be run on a single GPU, the forecasts where therefore generated in parallel. Generation of a single 96 hour, 6 hour interval, global forecast at full 1440x721x13 resolution, took on average 392 seconds. So for the date range considered in this study (12-01-2022 until 31-12-2022) the total was: 38.54 hours. That did however include loading the analysis data used for the initialization from Google Cloud.

The cost for a single 6.5 min run using a cloud compute provider using an A100: 3.35 euros/hour $\times$ 6.5 minutes/forecast $=$ € 0.36 per forecast, at Google Cloud Platform (GetDeploying, 2024).

However, factoring in the training cost of the model is relevant since most of the compute is invested before even a single forecast is generated. As mentioned in subsection 2.1.4, GraphCast is trained for 4 weeks on 32 cloud TPUv4 devices. The cost for training is therefore estimated as such:

€3.22 /TPUv4/hour $\times$ 32TPUv4's $\times$ 672hours minutes/forecast $=$ € $69,243$, at Google Cloud Platform (Google Cloud Platform, 2024).

Assuming either an updated architecture or new ERA dataset (ERA6 is set to be released in 2027 ECMWF, 2024) would warant retraining of the model, we thus say that the practical lifespan of GraphCast is 4 years. Although marginal costs of running an additional forecast are small we rely on initialization with analysis data, this data is released 4 times per day. If we thus assume that Graphcast is run 4 times daily, and spread the training cost over the assumed 4 year lifespan we arrive at the following estimation of total cost, including training, per forecast:

$$\frac{€69,243}{365 \times 4 \times 4} + €0.36 = €12.22 \text{ per forecast} \tag{3.1}$$

The costs of the various models are summarized in Table 3.4.

Post-processing
Besides limiting the extent of the forecasts a few other post-processing steps were performed before analysing the results.

1. Take only the surface variables: 2m temperature, 10m u/v components of wind, mean sea level pressure and total precipitation in the last 6 hours;
2. Add derived variables: 10m wind speed and direction;
3. perform unit conversions: temperature Kelvin to Celsius and Pascal to Hectopascal;
4. Perform bilinear interpolation to station locations.

Additionally, after inspecting preliminary results, it was found that the GraphCast forecasted data had a phase shift of 6 hours. This was accounted for by shifting the forecasted dataset 6 hours to the right, thereby loosing the first forecasted lead time step.

Lastly, in all further visualizations GraphCast will be indication with the shorthand: gc.

3.4.2. Pangu-Weather

The forecasts for the Pangu-Weather model were not created for this study, instead they were retrieved from Weatherbench2 (Rasp et al., 2024). However we will still highlight the process through which a forecast could be generated in the following section.

Similar to GraphCast the Pangu operational model uses the same shape and variables as ERA5. Refer to Table 3.2 and Table 3.3 for an overview of the variables and the pressure levels present in Pangu-Weather. The resolution is again: 6 hour intervals (although 1 hour and 24 hour models are also available), at full 1440x721x13 resolution.

The Pangu-Weather model was not run for this research. Instead a dataset of forecasts for 2022 was accessed from Rasp et al., 2024. The dataset accessed was that of the Pangu-Weather model in operational setting, i.e. initialized using IFS HRES initial conditions. These forecasts where generated following the code provided on the Github repository accompanying Bi et al., 2023. The 6-hour model was used such that the data is of comparable shape to the other models used.

Computation Cost

Generating forecasts with Pangu-Weather can be done on a single NVIDIA V100 GPU. Generation of a single 7-day global forecast takes only 1.4 seconds (Bi et al., 2023). The cost for a single 1.4-second run using a cloud compute provider using a V100 GPU is calculated as:

$$€2.33 \text{ EUR/hour} \times \frac{1.4 \text{ seconds/forecast}}{3600 \text{ seconds/hour}} = €0.00090 \text{ EUR per forecast}, \tag{3.2}$$

at Google Cloud Platform (Google Cloud Platform, 2024).

However, factoring in the training cost of the model is relevant since most of the compute is invested before even a single forecast is generated. As mentioned in (Bi et al., 2023), Pangu-Weather consists of four deep networks with lead times of 1 h, 3 h, 6 h, and 24 h, respectively. Each of these networks was trained for 100 epochs, taking approximately 16 days on a cluster of 192 NVIDIA Tesla V100 GPUs.

The cost for training is therefore estimated as:

$$\begin{aligned} \text{Total training cost} &= €2.33 \text{ EUR/GPU-hour} \times 192 \text{ GPUs} \times 384 \text{ hours} \times 4 \text{ networks} && (3.3)\\ &= €2.33 \times 192 \times 384 \times 4 && (3.4)\\ &= €687,491.33, && (3.5) \end{aligned}$$

at Google Cloud Platform (Google Cloud Platform, 2024), where 16 days $\times$ 24 hours/day $=$ 384 hours per network.

Assuming that either an updated architecture or new data would warrant retraining of the model, we consider the practical lifespan of Pangu-Weather to be 4 years. Although the marginal cost of running an additional forecast is small, operational forecasts are typically generated 4 times per day. If we assume that Pangu-Weather is run 4 times daily, and spread the training cost over the assumed 4-year lifespan, we arrive at the following estimation of total cost, including training, per forecast:

$$\frac{€687,491.33}{365 \times 4 \times 4} + €0.00090 = €117.76 \text{ per forecast}. \tag{3.6}$$

This calculation demonstrates that while the inference cost per forecast is minimal, the substantial training cost significantly increases the overall cost per forecast when amortized over the model's lifespan.

3.4.3. HRES

For the HRES forecasts a historic dataset of forecasts was used. Instead of retrieving the full dataset from the ECMWF directly a similar method was used to retrieving the analysis dataset. Historic forecast data generated using the ECMWF's IFS HRES are publicly available on Weatherbench2, this meant the data already in the same resolution as GraphCast.

Processing the HRES data followed the same pattern as processing the GraphCast data: see Equation 3.4.1 for full details. Note that the highest resolution HRES is available in is 0.1° or 9km horizontal resolution. Downsampling this to 0.25° and then interpolating to the station locations means HRES is not shown at its full potential.

Cost Estimation for HRES

Estimating the cost of running high-resolution weather forecasting models, such as the ECMWF's High-Resolution Forecast (HRES), involves several factors. The primary computational resource for these forecasts is the Atos BullSequana supercomputer, located in ECMWF's data center in Bologna, Italy. The total cost of the supercomputer contract is approximately €80 million over a four-year period ECMWF, 2021. This facility supports both operational forecasts and research activities.

To estimate the cost of a single forecast run, we begin by allocating the total computational resources between operational forecasts and research. Based on typical usage patterns, approximately 70% of the supercomputer's resources are dedicated to operational forecasts, with the remaining 30% used for research and other purposes ECMWF, 2023a.

Each day, the HRES and Ensemble (ENS) forecasts are run four times, producing a total of 5,840 forecast runs over the four-year period ECMWF, n.d. Assuming that 70% of the total cost is allocated to operational forecasting and 20% of that is HRES (which requires less compute than the Ensemble), the cost attributed to forecasting over the four years is

$$\text{Cost for Forecasting} = 0.70 \times 0.20 \times €80 \text{ million} = €11.2 \text{ million}$$

The cost per forecast run is then calculated as:

$$\text{Cost per Forecast Run} = \frac{€11.2 \text{ million}}{5840 \text{ runs}} \approx €1918$$

Considering this would be for 10 day forecasts and we only use 4 days, we arrive at a very rough estimate of **€ 762 per forecast run**. This estimation provides a rough idea of the cost per single forecast run. The actual cost may vary based on specific operational and research usage, as well as additional factors such as maintenance and upgrades. The costs of the various models are summarized in Table 3.4.

3.4.4. Aspire

In order to generate forecasts for an entire year for the full domain of the station locations, a trade-off was made between resolution and computational cost. The Aspire model has the capability to run in mesoscale mode, generally used to bridge the gap between ECMWF grid-scale boundary conditions and the high-resolution LES nests for the area of interest.

Mesoscale Simulation

The mesoscale weather model is configured to simulate atmospheric conditions over the Netherlands, covering the 47 SYNOP stations, centered at $(5.03°\text{E}, 52.61°\text{N})$. The model domain extends horizontally with a grid of 256×256 grid points, and vertically with 64 levels. The horizontal grid spacing is set to 2000 meters in both the east-west and north-south directions, while the vertical grid spacing is set to 40 meters.

The vertical grid structure begins with a uniform resolution of 40 meters up to 300 meters above the terrain, after which a growth rate is applied to increase the grid spacing with altitude. The domain height is capped at 8000 meters, ensuring that the model captures the lower to mid-troposphere dynamics effectively.

These settings provide a high-resolution simulation environment that balances computational efficiency and the ability to resolve mesoscale phenomena such as convective systems and boundary layer processes. The chosen grid resolution allows for detailed representation of meteorological variables and interaction with topography, which is essential for accurate weather forecasting over complex terrain and coastal areas.

Lastly in all further visualizations the Aspire Meso run will be indicated with the shorthand: meso.

LES Simulation

A second run of the Aspire model was performed to include LES nests at station 06203, P11-B (offshore), and station 06348, Cabauw (onshore). Refer to Figure 3.1 for the exact locations.

The LES setup in Aspire focuses on resolving smaller-scale turbulence and boundary-layer processes at a higher resolution than the mesoscale run. The configuration includes the following settings:

- **Location and Grid Structure:** The center coordinates for the LES domain are $(3.34°\text{E}, 52.36°\text{N})$ and $(4.93°\text{E}, 51.97°\text{N})$, for station 06203 and 06348 respectively. Both with a grid size of $128 \times 128 \times 64$.
- **Grid Spacing:** Horizontal grid spacing is set to 100 meters, while the vertical grid spacing is 25 meters up to a domain height of 2500 meters.
- **Growth Rate and Vertical Stretching:** Grid spacing grows above 300 meters to capture vertical variation effectively, with a growth rate applied to adaptively increase the vertical spacing with altitude.
- **Boundary Conditions:** Large-scale boundary conditions are provided by the mesoscale run, using HRES data, allowing realistic atmospheric inputs into the LES domains.

Diagnostics and Data Collection Diagnostics are set to record key variables with high temporal resolution, using the STATTFMETMAST module:

- **Variables:** Temperature, wind components (u, v), moisture (M), surface pressure, and precipitation rate.
- **Averaging Interval (dtav):** 1 second.
- **Write Interval (dtwrite):** 600 seconds.

This LES setup provides a high-resolution environment suitable for capturing boundary-layer dynamics and small-scale turbulent structures. The mesoscale output supplies boundary conditions, ensuring that the LES nests operate within a realistic and dynamically consistent framework.

Lastly in all further visualizations the Aspire LES run will be indicated with the shorthand: LES.

Computation Cost

The computation cost for generating forecasts with both the Aspire Meso and Aspire LES models was assessed using two GPUs: an Nvidia P100 and a V100, each with 16GB RAM. Similar to GraphCast, both models could be run on a single GPU and in parallel. However, the average run time differed for each model and GPU.

For the **Aspire Meso model**, a single 96-hour mesoscale run with a 10-minute interval, covering the extent of the Netherlands, took on average:

- 2300 seconds on a V100
- 3850 seconds on a P100

With all forecasts generated for this study, the total computation time for the Aspire Meso model was approximately 226 hours (assuming all runs were on a V100).

The cost for a single 38-minute run using a V100 on a cloud compute provider is calculated as:

$$2.48 \text{ euros/hour} \times \frac{38 \text{ minutes}}{60 \text{ minutes/hour}} = € \, 1.57 \text{ per forecast}$$

This estimate is specific to an area covering the Netherlands, unlike global forecasts for other models.

For the **Aspire LES model**, the simulation contained two LES nests, and since these are nested within a mesoscale run, we can attribute the computational cost based on the actual computational effort recorded. According to the timer statistics from the simulation, the mesoscale simulation accounted for 54.7% of the total computational effort, while the two LES nests accounted for 21.1% and 24.2%, respectively. The average total run time for this setup was:

- 4762 seconds on a V100
- 7447 seconds on a P100

Breaking down the total run time on a V100:

- Mesoscale simulation (54.7%): 0.547×4762 seconds $= 2605$ seconds
- LES Nest 1 (21.1%): 0.211×4762 seconds $= 1006$ seconds
- LES Nest 2 (24.2%): 0.242×4762 seconds $= 1151$ seconds

The total computation time for all Aspire LES forecasts in this study was approximately 396 hours (assuming all runs were on a V100). This total includes the computational time for the mesoscale simulation and both LES nests.

The cost for a single Aspire LES run using a V100 on a cloud compute provider is:

$$2.48 \text{ euros/hour} \times \frac{4762 \text{ seconds}}{3600 \text{ seconds/hour}} = € \, 3.28 \text{ per forecast}$$

Allocating the costs based on computational effort:

- Mesoscale simulation cost (54.7%): $0.547 \times € \, 3.28 = € \, 1.79$
- LES Nest 1 cost (21.1%): $0.211 \times € \, 3.28 = € \, 0.69$
- LES Nest 2 cost (24.2%): $0.242 \times € \, 3.28 = € \, 0.79$

As with the Aspire Meso model, this estimate is specific to a limited regional forecast rather than a global forecast, aligning both Aspire models' cost considerations with the designated study area.

The costs of the various models are summarized in Table 3.4.

Post-processing

Similar to Equation 3.4.1 the output data from the Aspire model, both in meso and in LES setting required post-processing before analysis could be performed. The steps performed were:

1. Aggregate rainfall using average precipitation rate;
2. Downsample from 10-minute frequency to data every 6 hours;
3. Concatenate 0-24, 24-48, 48-72 and 72-96 hour forecasts;
4. Convert surface pressure to mean sea level pressure, see below for a more detailed explanation.

The mean sea level pressure (MSLP) is calculated using the barometric formula, which relates the atmospheric pressure at a given altitude to the corresponding pressure at sea level. The barometric formula is given by:

$$P_0 = P \cdot \exp\left(\frac{g \cdot h}{R \cdot T}\right)$$

where:

- P_0 is the sea level pressure (Pa),
- P is the observed pressure at height h (Pa),
- $g = 9.80665\,\text{m/s}^2$ is the acceleration due to gravity,
- h is the height above sea level (m),
- $R = 287.05\,\text{J/(kg·K)}$ is the specific gas constant for dry air,
- T is the temperature in Kelvin (K).

This process allows for an accurate estimation of the mean sea level pressure based on the observed pressure at the station's altitude, adjusted for temperature variations.

Table 3.4: Summary of Computational Costs and Characteristics of Different Models

				Cost			
Model	**Domain**	**Resolution (H × V)**	**Grid Points**	**Time**	**Single Run**	**Incl. Training**	**Per M Grid Points**
GraphCast	Global	0.25° (31 km) × 13 levels	13,478,400	6.5 min	€0.36	€12.22	€0.91
Pangu-Weather	Global	0.25° (31 km) × 13 levels	13,478,400	1.4 sec	€0.00090	€117.76	€8.74
HRES	Global	0.10° (9 km) × 137 levels	84,240,000	Several hours	N/A	€762	€9.05
Aspire Meso	Regional (NL)	2 km × 64 levels	4,194,304	38 min	€1.58	€1.58	€0.38
Aspire LES	Local (Stations)	100 m × 64 levels	1,048,576	80 min	€3.28	€3.28	€3.13

3.4.5. Computational Cost Summary

In this section, we summarize the computational costs associated with each of the models used in this study. The models differ not only in their underlying architecture and forecasting capabilities but also in their operational domain (global, regional, or local) and their setup configurations. Table 3.4 provides a detailed comparison of the models, including their domains, resolutions, grid points, computation times, and associated costs.

As shown in Table 3.4, the computational costs vary significantly between models due to differences in domain coverage, resolution, and computational requirements. Global models like GraphCast, Pangu-Weather, and HRES cover the entire globe and have a larger number of grid points compared to regional or local models like Aspire Meso and Aspire LES.

The **GraphCast** model operates on a global scale with a horizontal resolution of 0.25° (approximately 31 km) and 13 vertical levels, resulting in 13,478,400 grid points. The computation time for a single forecast is approximately 6.5 minutes, with an inference cost of €0.36 per run. When including the amortized training cost over the model's operational lifespan (assumed to be 4 years), the total cost per run is €12.22. This results in a cost of €0.91 per million grid points.

The **Pangu-Weather** model also operates globally at the same resolution as GraphCast, with 13,478,400 grid points. However, it is significantly faster, with a computation time of only 1.4 seconds per forecast and an inference cost of €0.00090. Due to the substantial training costs amortized over the operational lifespan, the total cost per run is €117.76, leading to a cost of €8.74 per million grid points.

The **HRES** model from ECMWF operates globally at a higher horizontal resolution of 0.10° (approximately 9 km) with 137 vertical levels, resulting in 84,240,000 grid points. The computation time is several hours per forecast, and the total cost per run is estimated at €762. This translates to a cost of €9.05 per million grid points.

The **Aspire Meso** model is a regional model covering the Netherlands, with a horizontal resolution of 2 km and 64 vertical levels, totaling 4,194,304 grid points. Each forecast takes approximately 38 minutes to compute, with a cost of €1.58 per run. Since there are no significant training costs, the total cost per run remains €1.58, resulting in a cost of €0.38 per million grid points.

The **Aspire LES** model operates at a local scale, focusing on specific stations with a high horizontal resolution of 100 meters and 64 vertical levels, amounting to 1,048,576 grid points. Each forecast requires around 80 minutes of computation time, costing €3.28 per run. Based on the updated computational effort breakdown, we attribute €1.79 of the cost to the mesoscale simulation and €1.49 to the two LES nests. Without additional training costs, the total cost per run is €3.28, leading to a cost of €3.13 per million grid points.

Including the vertical resolution in the grid point calculations provides a more accurate representation of the computational effort required for each model, as atmospheric models solve equations in three dimensions. The cost per million grid points normalizes the costs, allowing for a fair comparison between models of different scales and resolutions.

It is important to note that models like GraphCast and Pangu-Weather have significant training costs, which are amortized over their operational lifespans. In contrast, the Aspire models have no training costs, and their costs are entirely operational (excluding R&D costs). Additionally, the high cost per million grid points for models like Pangu-Weather and HRES reflects the substantial computational resources required for global forecasting at higher resolutions.

3.5. Ensemble Forecasting

Ensemble forecasting is a methodological approach in meteorology where multiple forecasts are generated and combined to produce a single, more reliable prediction. This technique addresses the inherent uncertainties in weather prediction models by considering a range of possible future states, thereby capturing the uncertainty and reducing the errors associated with individual model forecasts.

Single-model forecasts are limited by their inherent biases, parameterization schemes, and sensitivities to initial conditions. By combining forecasts from different models, ensembles mitigate these limitations through the averaging of individual model errors. The ensemble mean often provides a more accurate forecast than any single model due to the cancellation of random errors and the reinforcement of common signals.

Ensemble forecasting is widely adopted in operational meteorology. Agencies such as ECMWF produce ensemble forecasts to enhance predictive skill and quantify forecast uncertainty. In operational scenarios, ensembles inform decision-making processes by providing probabilistic forecasts, which are essential for risk assessment in sectors like energy, aviation, agriculture, and disaster management.

In operational practice, ensemble forecasts often employ *weighted averaging*, where individual model forecasts are assigned weights based on their historical performance, error characteristics, and biases. This approach acknowledges that not all models contribute equally to forecast accuracy. By assigning higher weights to more reliable models, the overall forecast can be improved. However, determining the appropriate weights requires extensive statistical analysis and can be complex. (De Menezes et al., 2000; Elliott, 2011; Nielsen et al., 2007)

The theoretical foundation of ensemble forecasting lies in the understanding that atmospheric systems are sensitive to initial conditions—a concept known as the butterfly effect. Ensembles address this by incorporating variations in initial conditions and model formulations, leading to a spread of possible outcomes that better represent the range of potential future states. This approach enhances the reliability and robustness of weather forecasts.

3.5.1. Implementation in This Study

In this research, we employed simple equal-weighted averaging to create ensemble forecasts from the individual models: GraphCast, HRES, Aspire Meso, and Pangu-Weather. The ensembles were constructed by taking the arithmetic mean of the forecasted values from two or more models for each meteorological variable at each forecast lead time. This method is straightforward to implement and has been shown to improve forecast accuracy.

The ensemble combinations considered in this study are as follows:

- **GraphCast + HRES**
- **GraphCast + Aspire Meso**
- **GraphCast + Pangu-Weather**
- **HRES + Aspire Meso**
- **HRES + Pangu-Weather**
- **GraphCast + HRES + Aspire Meso + Pangu-Weather**

The ensembles were created using Python and numerical libraries such as NumPy and Xarray for efficient data manipulation and computation. The forecast data from each model were first interpolated onto a common spatial grid corresponding to the locations of the 47 KNMI SYNOP stations. For each time step and variable, the ensemble forecast was calculated as:

$$F_{\text{ensemble}} = \frac{1}{N} \sum_{i=1}^{N} F_i, \tag{3.7}$$

where F_i is the forecast from model i, and N is the number of models in the ensemble.

Regarding computational cost, running multiple models independently can be resource-intensive. However, since the individual models are already operational and their forecasts are readily available, the additional computational expense of creating ensembles is minimal. The ensemble process involves simple arithmetic operations, which are computationally negligible compared to the cost of running full numerical weather prediction or AI-based models.

3.5.2. Summary

Ensemble forecasting enhances weather prediction by combining multiple model outputs to reduce errors and quantify uncertainties. It is a well-established practice in operational meteorology, providing more reliable forecasts than single-model predictions. In this study, we implemented ensemble forecasting through equal-weighted averaging of the GraphCast, HRES, Aspire Meso, and Pangu-Weather models. This approach leverages the strengths of each model while mitigating their individual weaknesses. While equal weighting is a limitation compared to more sophisticated weighted methods used operationally, it provides a clear and straightforward assessment of the benefits of ensemble forecasting. The ensembles are expected to improve forecast accuracy and provide a more robust evaluation of weather variables. The implementation is computationally efficient, given that the models are already operational, and is expected to add significant value to the forecasting process. In the next chapter (section 4.7) we will assess the performance of the ensembles as well as whether these ensembles improve performance with respect to their individual members.

3.6. Evaluation of Results

This section outlines the methods used to evaluate model performance, detailing the metrics and dimensions of analysis applied in comparing forecasts to observations. By defining key error metrics and explaining how results are aggregated across different temporal and spatial dimensions, this section establishes a structured approach for assessing forecast accuracy. These methods provide the foundation for interpreting model behavior and identifying patterns in the actual evaluation, discussed in later Chapters.

3.6.1. Dimensions of Analysis: Aggregating and Averaging

In this research, three key dimensions of analysis are considered, two of which are time-based, and one spatial. The dimensions are:

1. Initialization time,
2. Lead time,
3. Station location.

Each of these dimensions offers unique insights into model performance, but averaging across any of these dimensions can obscure important variability. It is essential to consider the trade-offs between reducing variance for clarity versus retaining detailed information that might reveal specific strengths or weaknesses in the models. In the following subsections, we will discuss each dimension and how aggregating across it affects the interpretation of the error metrics introduced earlier.

Initialization Time

Initialization time refers to the moment when a forecast was generated. In this research, forecasts are initialized daily for the year 2022 (as mentioned earlier: starting 12-01-2022). Seasonal variations in atmospheric conditions may cause model performance to differ depending on the time of year. For example, some models may perform better during the winter months than during summer, or vice versa.

Averaging across initialization times can smooth out seasonal variability, providing a general sense of model performance over the year. However, this approach risks losing insights into season-specific performance, which may be crucial for applications such as energy production forecasting or agricultural planning. Additionally, keeping the initialization time dimension separate allows for the identification of seasonal patterns or biases that may be masked by aggregation.

Averaging across initialization times results in reduced variance and provides an overall assessment of annual model performance. However, it may hide important seasonal trends. For example, models could exhibit biases related to specific weather patterns that dominate during certain times of the year

(e.g., summer heat waves or winter storms). Therefore, in later sections, we present results both with and without this averaging.

Lead Time

Lead time refers to the time interval between the forecast initialization and the time for which the forecast is made. In this research, lead times range from 6 to 96 hours, providing forecasts at 6-hour intervals.

Analyzing model performance across different lead times is essential for understanding how quickly a model's accuracy degrades over time. In general, we expect the accuracy of weather forecasts to decrease as lead time increases. This degradation reflects the limits of the model's predictive capability and the increasing uncertainty in atmospheric conditions as time progresses, as well as the aggregation of errors resulting from the initial conditions not perfectly representing the real state of the atmosphere.

When aggregating results over multiple lead times, it is important to note that this can hide valuable information about the temporal evolution of model accuracy. A model that performs well at short lead times (e.g., 6-24 hours) may perform poorly at longer lead times (e.g., 72-96 hours), or vice versa. Therefore, aggregating across lead times should be done cautiously and with specific goals in mind. A second consideration is that performance at different lead time might show specific diurnal patterns that would get lost averaged over. For temperature and rainfall there are known diurnal patterns that some models could pick up on while others could fail at identifying these.

In the subsequent chapters, we will present performance across different lead times individually and discuss the consequences of averaging over this dimension.

Station Location

The third dimension of analysis is station location. Each forecast is compared to observed values recorded at multiple weather stations distributed across different geographic regions. The stations in this study cover a range of environments, from onshore to offshore and urban to rural settings. Model performance may vary significantly depending on the location, as different models might handle local weather phenomena like sea breezes, elevation effects, or urban heat islands better or worse than others.

Averaging across station locations allows for a more generalized assessment of model performance over a region. However, this approach risks obscuring localized patterns, such as a model performing well in offshore areas but poorly in inland regions. Therefore, it is important to assess model performance at individual stations before aggregating results. In this study, we will compare model results at individual stations as well as in aggregate to balance the need for detail and generalization.

Combining Dimensions

In some cases, it may be necessary to average across multiple dimensions for ease of interpretation, such as averaging across both initialization times and station location. However, combining dimensions increases the risk of losing detailed insights, as aggregation removes important variability that might reveal model-specific strengths or weaknesses.

In this research, we will present results both aggregated and disaggregated across dimensions. This approach ensures that we retain critical insights while providing a clear overall picture of model performance. Careful consideration will be given to the choice of dimensions for aggregation in order to highlight the most relevant aspects of the models' strengths and limitations.

An important 'sanity check' will be in the ability to visually inspect the results in the contiguous timeseries plots. These will constitute the most disaggregated form of data: with a single variable, at a single location, for a continuous time range. More on this in section 4.2.

3.6.2. Methods

The evaluation of model performance is based on several error metrics, chosen to give a comprehensive understanding of forecast accuracy across multiple variables. The following metrics were used: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Bias, and Standard Deviation. Each of these metrics offers unique insights into model behavior, from penalizing large errors to identifying systematic over- or under-forecasting.

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is used to measure the difference between predicted and observed values. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

where y_i is the observed value, $\hat{y_i}$ is the forecasted value, and n is the number of observations. RMSE gives more weight to larger errors, making it particularly useful for identifying forecasts with significant deviations from observations. This metric is critical for assessing model accuracy, especially for variables like temperature and wind speed, where large deviations may have operational consequences.

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) provides a straightforward interpretation of the average error in forecasts. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

MAE does not penalize large errors as strongly as RMSE, offering a more balanced view of forecast performance. It complements RMSE by providing an intuitive understanding of how much a typical forecast deviates from observations. This metric is useful for operational decision-making, where knowing the average error is often more informative than focusing on extreme outliers.

Bias

Bias is a measure of the systematic error in the forecasts, indicating whether a model tends to consistently over-forecast or under-forecast a particular variable. It is calculated as:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)$$

A positive bias indicates over-forecasting, while a negative bias suggests under-forecasting. Bias is essential for understanding consistent directional errors in the model, which can help in fine-tuning or adjusting the model for specific applications, such as wind power forecasting where systematic errors can affect energy yield predictions.

Standard Deviation

The standard deviation of forecast errors assesses how well a model captures the natural variability of atmospheric conditions. It is calculated as:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the observed values. A lower standard deviation indicates that the model's forecasts closely match the variability in the observed data, while a higher standard deviation suggests that the model is less able to capture fluctuations in the atmosphere. This metric is especially important for understanding forecast uncertainty, which can be critical for decision-making in fields like aviation and renewable energy.

Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) measures the strength and direction of the linear relationship between forecasted and observed values. It is calculated as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y_i} - \bar{\hat{y}})^2}}$$

where y_i and $\hat{y_i}$ are the observed and forecasted values, and $\bar{y}$ and $\bar{\hat{y}}$ are the means of the observed and forecasted values, respectively. The coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

The Pearson Correlation Coefficient is useful for understanding how well the model captures the overall trend between observed and forecasted data. A higher correlation suggests that the model successfully predicts the general direction of change in the variable of interest, even if the magnitudes differ. This metric is particularly important when assessing variables like wind direction and precipitation patterns, where the ability to forecast directional changes is crucial.

4

Analysis of Forecast Performance

In this chapter, we evaluate the performance of various weather models in predicting key atmospheric variables in the Netherlands by comparing their forecasts with observational data from 47 KNMI SYNOP stations. Our aim is to assess the accuracy, reliability, and practical applicability of each model, focusing on variables such as 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total accumulated precipitation over 6 hours.

We begin with visualizations of model forecasts over time, focusing on significant weather events like Storms Dudley and Eunice. These visualizations offer an initial insight into how well models capture the spatial and temporal characteristics of extreme weather phenomena.

Next, we perform time series analyses at specific locations to observe model performance over time, especially during severe weather events. This detailed view highlights the strengths and weaknesses of the models in predicting the intensity and timing of these events.

Lead time analysis follows, examining how forecast errors evolve over increasing lead times. This analysis reveals patterns such as diurnal cycles and baseline errors inherent in the models, offering insight into their predictive capabilities over different horizons.

We also conduct spatial analyses to reveal geographical patterns in forecast errors, exploring how accuracy varies between onshore and offshore locations and discussing the impact of localized phenomena, such as wind farms, on model performance.

Scatter density plots are utilized to compare forecasts against observations, identifying biases and variances in the models' predictions, particularly at higher wind speeds. These plots help in diagnosing systematic errors and evaluating the models' ability to capture extreme values.

To explore relationships between models, we use correlation matrices to assess inter-model similarity and agreement, highlighting the influence of shared data sources and methodologies on model outputs.

We then assess the performance of ensemble models, demonstrating how combining forecasts from different models can enhance accuracy and reliability. We discuss the benefits and limitations of equally weighted ensembles and consider the potential for optimized weighting schemes to further improve forecast skill.

Finally, we analyze the performance of the Aspire LES model compared to other models using Taylor diagrams, providing insights into the LES model's ability to capture fine-scale atmospheric processes relative to traditional numerical weather prediction and AI-based models.

Overall, this comprehensive analysis sheds light on the strengths and limitations of each model.

4.1. Visualization of Model Forecasts

To gain an initial understanding of how the GraphCast model forecasts weather variables over time, we present a series of forecast maps focusing on significant weather events. Figure 4.1 shows the 10 m

wind speed forecasts up to 96 hours ahead for February 16–19, 2022, produced by the GraphCast model. This period encompasses the storm events Dudley and Eunice, which had a substantial impact on the Netherlands.

In mid-February 2022, the Netherlands experienced an extraordinary sequence of three storms—Dudley, Eunice, and Franklin—over six consecutive days. This series of storms set a record for the country, with storm-force winds recorded each day. Storm Eunice, the most severe of the three, struck on February 18 and was classified as a heavy storm with wind force peaks reaching 11 on the Beaufort scale. Wind gusts exceeded 160 km/h near the Wadden Sea, and the storm caused significant disruptions, including halting public life in the afternoon and resulting in five fatalities.

Figure 4.1 displays the spatial distribution of wind speed across the Netherlands during the forecast period, enabling a visual assessment of the model's ability to capture the development and progression of these storm events. Each panel represents the forecast at 6-hour intervals, extending up to 96 hours ahead from the initialization time on February 16, 2022.

The maps illustrate a clear increase in wind speeds associated with the approaching storms. The GraphCast model effectively forecasts the general patterns of wind speed, showing elevated wind speeds along the coast and over the North Sea, consistent with the typical characteristics of such storm events. The temporal progression captured in the forecasts aligns with the timing of Storm Dudley and Storm Eunice, indicating the model's capability to predict significant weather phenomena.

However, some differences may be observed in the finer spatial details and the precise timing of peak winds. The model may underpredict or overpredict localized wind speed maxima, and there may be slight discrepancies in the onset and dissipation times of the storms. Additionally it can be seen that the model output, while accurate, is relatively smooth and increasingly blurry with longer lead times. A rule of thumb in weather forecasting is that the *Effective Resolution* of a model, i.e. the size of weather patterns it can capture effectively, is $6\times$ its actual resolution. In the case of GraphCast ($\sim$ 31 km resolution) this implies that the effective resolution is $\sim$186 km, which means it would struggle with smaller scale events. This can be seen in the the smoothness near the land-sea boundary during the storm events. This highlight the importance of detailed verification and model evaluation, which are further explored in subsequent analyses.

A more detailed examination of the model's performance during this period is provided in Figure 4.2, which presents time series of mean sea level pressure and wind speed. This analysis allows for a quantitative assessment of the model's accuracy in forecasting the intensity and timing of the storm events.

These visualizations provide valuable insights into the model's strengths and limitations in forecasting severe weather events. By analyzing the forecast maps over time, we can assess how well the model predicts significant changes in weather variables and identify areas where model improvements may be necessary. The next section will take a deeper look at how GraphCast and other models forecast the extreme wind events in Februari 2022.

4.2. Time Series Analysis of Forecasts

To further evaluate the performance of the models during significant weather events, we analyze contiguous time series plots of key meteorological variables. Figure 4.2 presents the 10 m wind speed and mean sea level pressure (MSLP) from February 14 to February 28, 2022, at offshore station 06203 (P11-B). This period includes the record-breaking six consecutive days of storm-force winds caused by the triplet of storms: Dudley, Eunice, and Franklin.

The plots clearly illustrate the three storm events, as well as the rapid drop in mean sea level pressure that led to extremely high wind speeds. The highest recorded 10 m wind speed at the offshore station is 25 m/s, occurring at the peak of Storm Eunice. It is worth noting that the observations have been subsampled every 6 hours, so the actual maximum wind speeds may be higher. Each storm is labeled for clarity.

The models exhibit various strengths and weaknesses during these storms. Aspire Meso shows a consistent underestimation of mean sea level pressure, though it accurately predicts the direction of

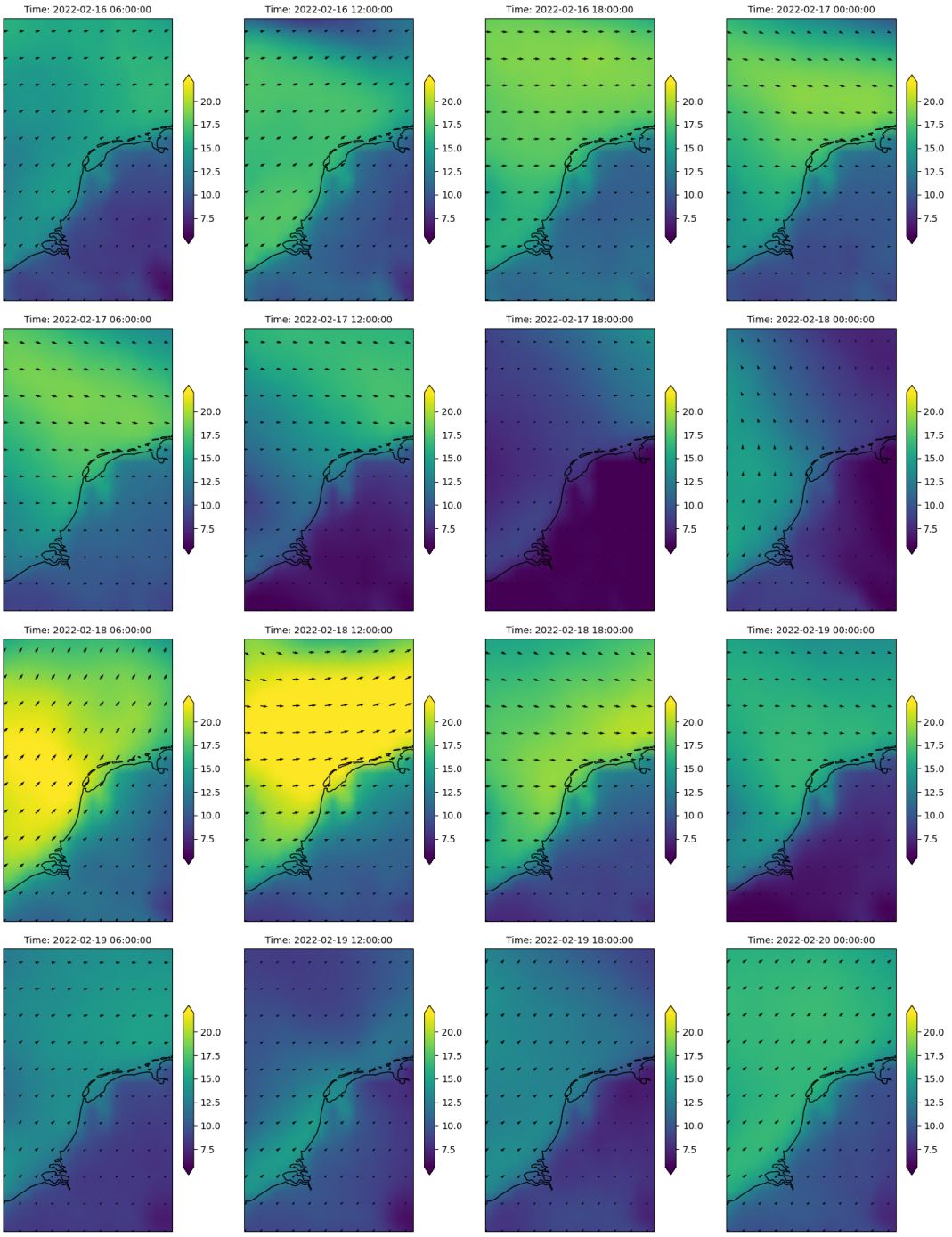
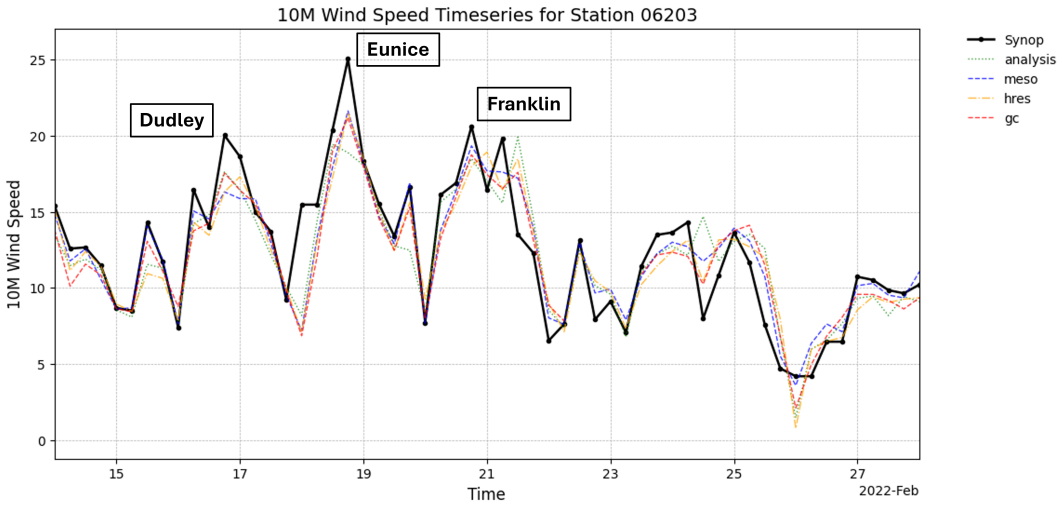


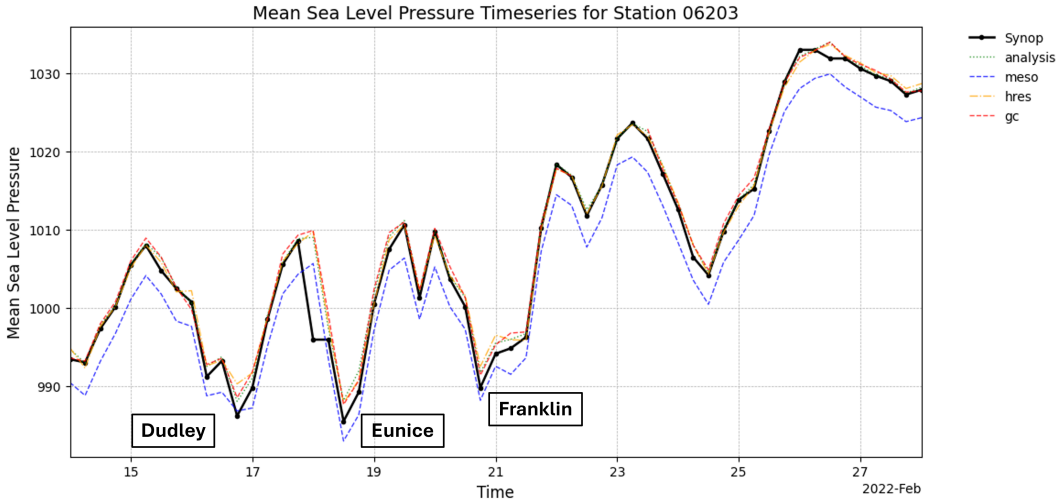
Figure 4.1: Forecasts up to 96 hours ahead for February 16–19, 2022, produced by the GraphCast model. The maps display the 10 m wind speed over the Netherlands, highlighting the progression of Storms Dudley and Eunice. Each panel represents a forecast at 6-hour intervals from the initialization time on February 16, 2022.

changes. All models manage to capture the occurrence of Storms Dudley and Franklin, successfully forecasting the drop in MSLP and the subsequent increase in wind speed. However, peak wind speeds are generally underpredicted across models.

During Storm Eunice, there is a notable forecast timing error, with all models missing the critical drop in MSLP by approximately 6 hours. Observations indicate a significant drop in MSLP at 2022-02-17T18, leading to a rapid increase in wind speed. The models, however, forecast this drop at 2022-02-18T00, resulting in a delayed wind speed increase and a lower forecasted peak wind speed. Specifically,



(a) 10 m Wind Speed at Station 06203



(b) Mean Sea Level Pressure at Station 06203

Figure 4.2: Time series plots of 10 m wind speed and mean sea level pressure from February 14 to February 28, 2022, at offshore station 06203 (P11-B). Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue. The record-breaking six consecutive days of storm wind speeds, caused by the triplet of storms Dudley, Eunice, and Franklin, is highlighted.

GraphCast forecasts the peak wind speed to reach only 22 m/s, as shown in Figure 4.1.

For Storm Franklin, two distinct wind speed peaks are observed. The models predict the timing of the first peak well but tend to underpredict the wind speed. For the second peak, both the timing and wind speed are slightly off, with the models showing a delay.

A direct link between MSLP and wind speed forecast errors is evident from comparing the two sets of plots. If a model misestimates MSLP—whether in timing or in the rate of change—this is reflected in the timing and magnitude of its wind speed predictions. In the case of these storms, a less steep or delayed forecasted drop in MSLP results in lower maximum forecasted wind speeds compared to observed values.

These observations highlight the inherent challenges in accurately forecasting both the timing and intensity of severe storm events. While the models demonstrate skill in capturing general trends and patterns, they face limitations in achieving precise timing and magnitude for extreme weather. This analysis underscores the importance of continuous model refinement and the need for high-resolution

data to enhance forecast accuracy for critical weather events.

4.3. Lead Time Analysis

Lead time plots illustrate how forecast errors evolve with increasing lead time, offering insights into the models' performance over different forecast horizons. Figure 4.3 presents the Root Mean Squared Error (RMSE) as a function of lead time for various weather variables and models. It is, however, important to keep in mind the baseline error all models exhibit: note the y-axes do not start at 0. Refer to section 3.3 for a more detailed exploration of this baseline error, as this is already evident in the ground-truth datasets (ERA5 and Analysis.

Analysis of the lead time plots reveals several key insights into the models' performance over different forecast horizons. First, the comparison among the models—including Aspire Meso, HRES, GraphCast, Pangu-Weather, and Ensemble—shows variations in performance across different variables and lead times (Figure 4.3). Generally, GraphCast (red) demonstrates superior performance, consistently exhibiting lower RMSE values across most variables and lead times. HRES (orange) performs moderately well, often outperforming Pangu-Weather (purple) and Aspire Meso (blue) but not matching the accuracy of GraphCast. The Ensemble model (green) shows superior performance to any individual model, at all lead times, suggesting that ensemble forecasting effectively captures uncertainty and improves prediction accuracy.

Notably, diurnal patterns are visible in the bias of temperature, wind, and precipitation forecasts, reflecting the influence of daily cycles on these variables (Figure 4.4). Specifically, for 2 m temperature, Aspire Meso displays an inverted bias compared to the other models. While GraphCast, Pangu-Weather, HRES, and the Ensemble tend to underestimate daytime temperatures and overestimate nighttime temperatures, Aspire Meso does the opposite—overestimating daytime temperatures and underestimating nighttime temperatures. This inversion suggests that Aspire Meso may handle diurnal temperature variations differently, potentially due to differences in its physical parameterizations or the representation of boundary layer processes.

In the case of wind variables, the inclusion of 10 m wind direction (°) alongside wind speed (m/s) provides a more comprehensive understanding of wind forecasts. The models show varying degrees of bias in wind direction forecasts, with ensemble methods potentially offering advantages due to their ability to represent a range of possible outcomes.

Additionally, the analysis of correlation coefficients (Figure 4.5) shows that variables like 2 m temperature and mean sea level pressure are relatively easier to forecast, as evidenced by higher correlation coefficients across all models. This indicates that these variables are more predictable and less sensitive to model errors. In contrast, variables like 10 m wind speed, wind direction, and total 6-hour precipitation exhibit lower correlation coefficients and higher forecast errors. These variables are inherently more volatile and influenced by localized phenomena, making accurate forecasting more challenging.

The inclusion of total 6-hour precipitation (mm) highlights the models' capabilities in predicting precipitation events, which are critical for hydrological applications and disaster management. The models generally struggle with precipitation forecasts, as indicated by higher RMSE values and lower correlation coefficients. Ensemble forecasting shows some promise in this area, but in this instance the all model ensemble appears to suffer from the inclusion of the Aspire Meso model. Ensembling is explored further in section 4.7.

In summary, the analysis provides a comprehensive evaluation of forecasting performance. GraphCast exhibits superior accuracy, particularly for variables that are traditionally challenging to predict. The Ensemble model's performance underscores the value of probabilistic forecasting in capturing uncertainties inherent in weather prediction.

4.4. Spatial Analysis of Forecast Errors

To understand how forecast errors vary geographically, we present map plots of error metrics for each station. The metrics are averaged over the entire 96-hour lead time window. Figures 4.6a and 4.6b show the RMSE of 2 m temperature and 10 m wind speed, respectively, for four models: GraphCast, Pangu-Weather, HRES, and Aspire Meso. The RMSE is averaged over all lead times and for the entire

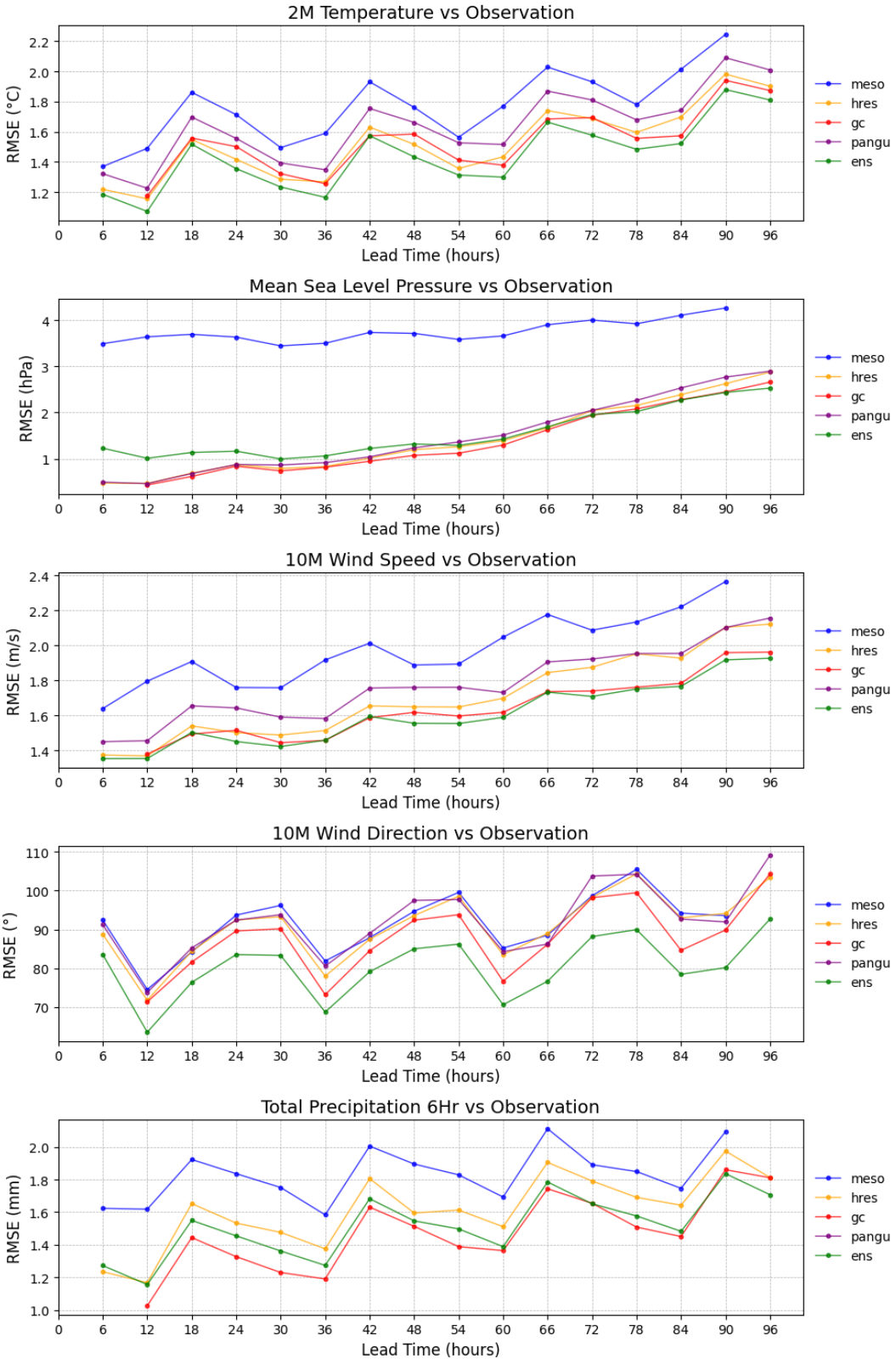


Figure 4.3: Root Mean Squared Error (RMSE) as a function of lead time for various weather variables and models. The variables include 2 m temperature (°C), mean sea level pressure (hPa), 10 m wind speed (m/s), 10 m wind direction (°), and total 6-hour precipitation (mm). Each line represents the RMSE for a specific model—Aspire Meso (blue), HRES (orange), GraphCast (red), Pangu-Weather (purple), and Ensemble (green)—averaged across all station locations.

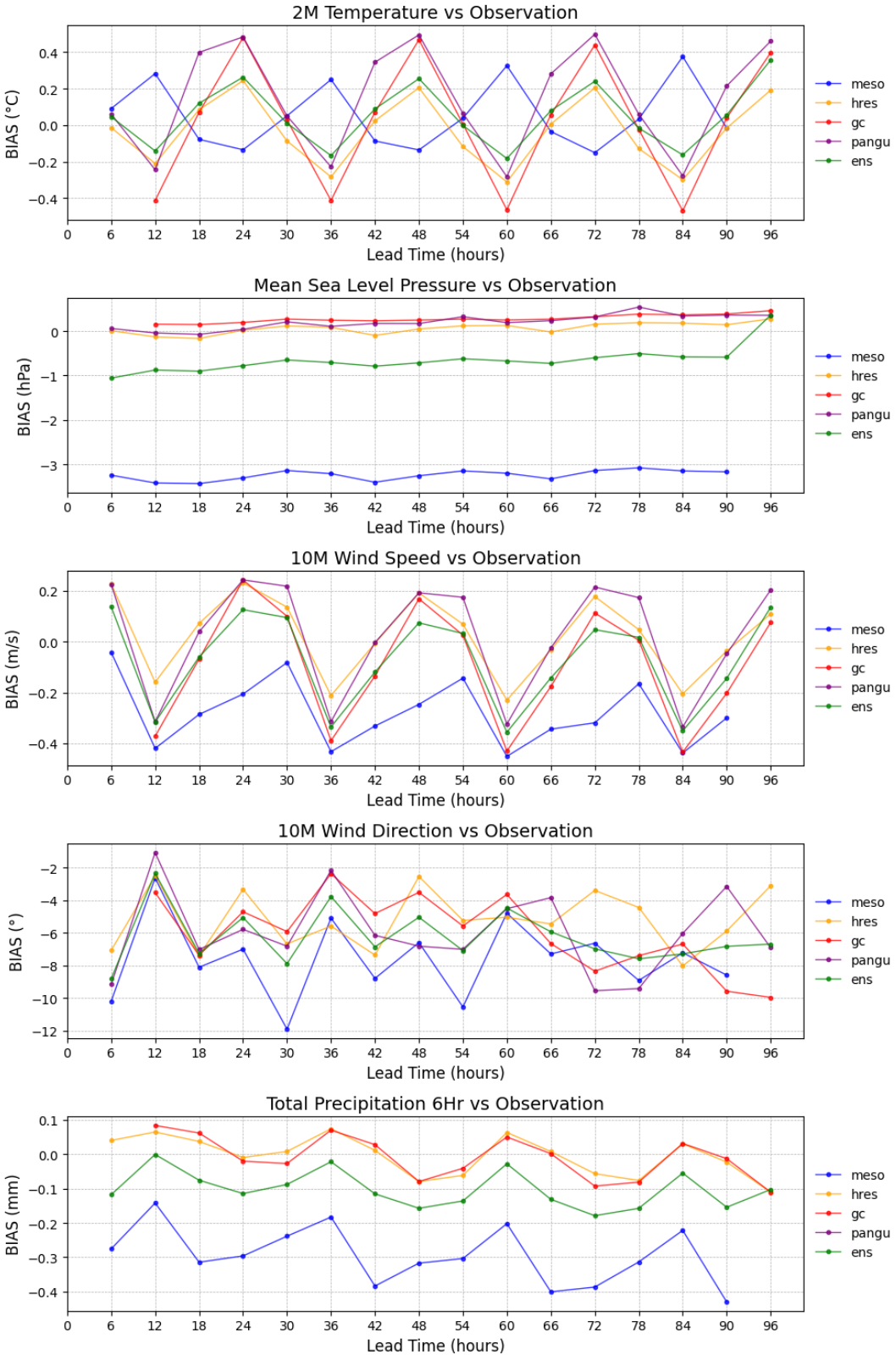


Figure 4.4: Bias as a function of lead time for various weather variables and models. The variables include 2 m temperature (°C), mean sea level pressure (hPa), 10 m wind speed (m/s), 10 m wind direction (°), and total 6-hour precipitation (mm). Each line represents the bias for a specific model—Aspire Meso (blue), HRES (orange), GraphCast (red), Pangu-Weather (purple), and Ensemble (green)—averaged across all station locations.

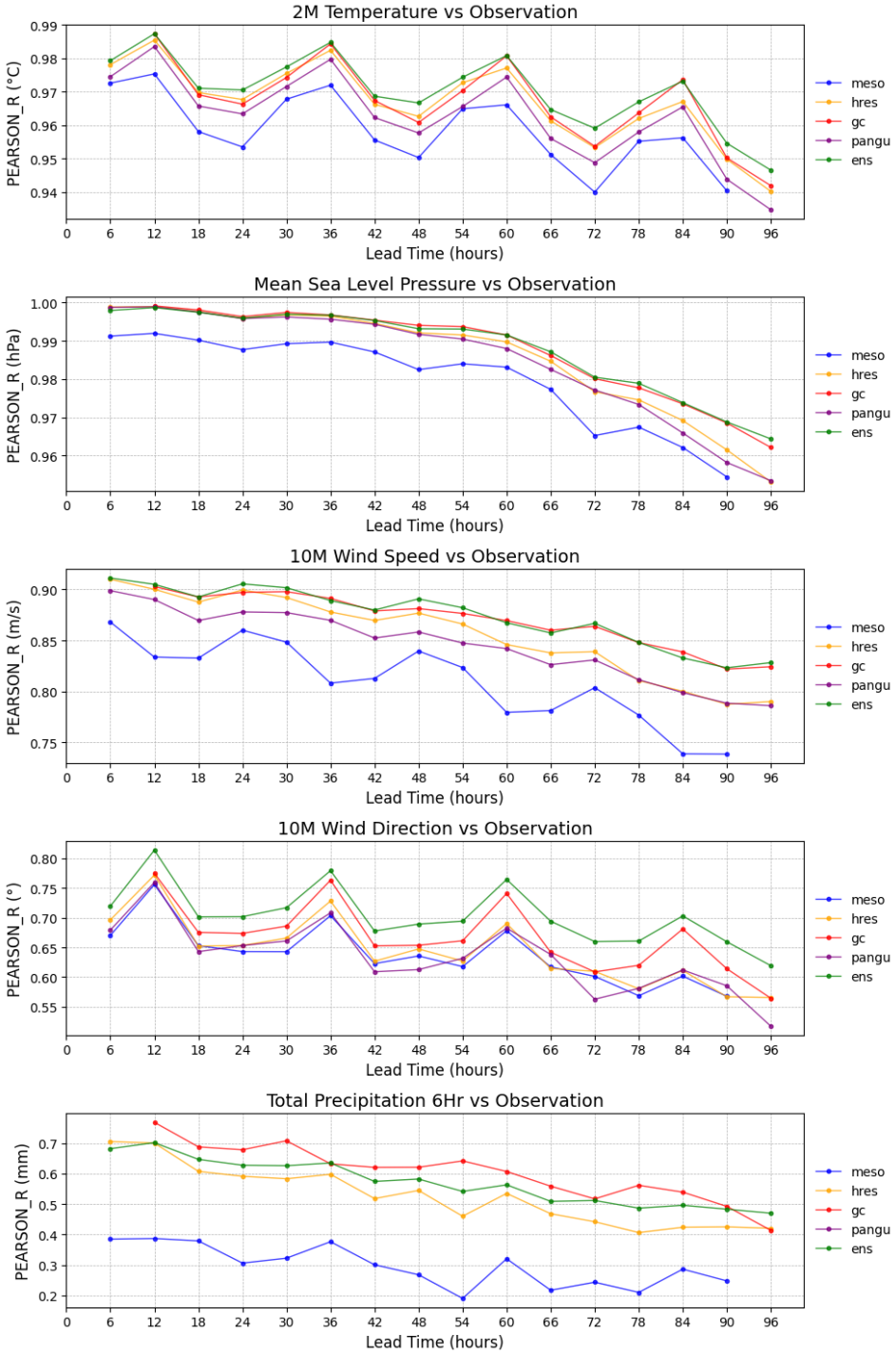


Figure 4.5: Pearson correlation coefficient as a function of lead time for various weather variables and models. The variables include 2 m temperature (°C), mean sea level pressure (hPa), 10 m wind speed (m/s), 10 m wind direction (°), and total 6-hour precipitation (mm). Each line represents the correlation coefficient for a specific model—Aspire Meso (blue), HRES (orange), GraphCast (red), Pangu-Weather (purple), and Ensemble (green)—averaged across all station locations.

year of 2022.

The spatial analysis of forecast errors reveals several noteworthy patterns, which for GraphCast and Pangu-Weather closely mirror the error and bias distributions observed in the ERA5 and Analysis datasets discussed in Section 3.3.3 (Figures 3.4 and 3.5). Notably, there is a distinct onshore-offshore divide in errors for different variables. For temperature, higher errors are observed onshore in all models

(Figure 4.6a), suggesting that temperature forecasts are less accurate over land areas. This is possibly due to the increased complexity of land surface processes and higher spatial variability compared to the relatively homogeneous sea surface.

In contrast, for wind speed, higher errors are observed offshore for GraphCast, Pangu-Weather, and HRES (Figure 4.6b), consistent with the patterns seen in the ERA5 and Analysis datasets (Figure 3.4). This indicates that wind speed forecasts are less accurate over the sea, which could be due to challenges in modeling marine atmospheric dynamics or the influence of data sparsity in these regions.

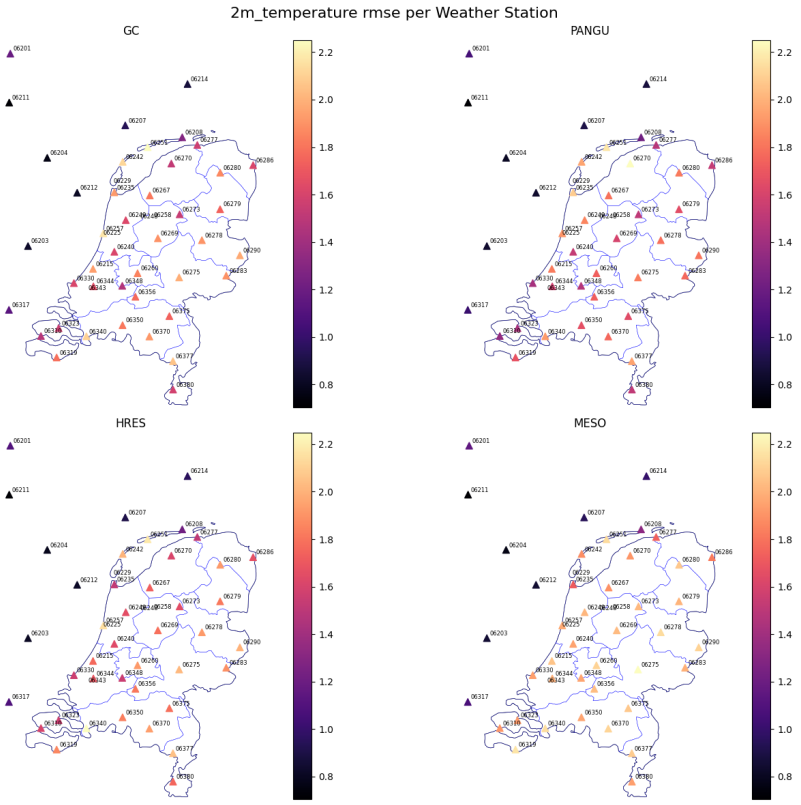
Moreover, the AI models, GraphCast and Pangu-Weather, which are trained using ERA5 and Analysis data, have learned and replicated these error patterns. Their spatial distributions of errors and biases closely resemble those of their training data, indicating that the models have inherited the strengths and weaknesses present in the datasets. For instance, the higher errors and negative biases offshore in wind speed forecasts (Figure 4.7) reflect similar tendencies in the ERA5 and Analysis datasets (Figure 3.5). This suggests that the AI models have effectively learned from the data but also that they propagate existing biases and errors inherent in the training datasets. This hints at a barrier in skill imposed by this training data, we will discuss this further in chapter 5.

Anomalies are also evident at specific stations. For example, station 06225 in IJmuiden consistently shows higher errors in all models, with GraphCast exhibiting the least error among them. This suggests that local factors, such as complex terrain or coastal effects, may be affecting forecast accuracy at this location. Examination of the bias maps (Figure 4.7) reveals that all models significantly underestimate the wind speed at this station, further highlighting the need for localized model improvements.

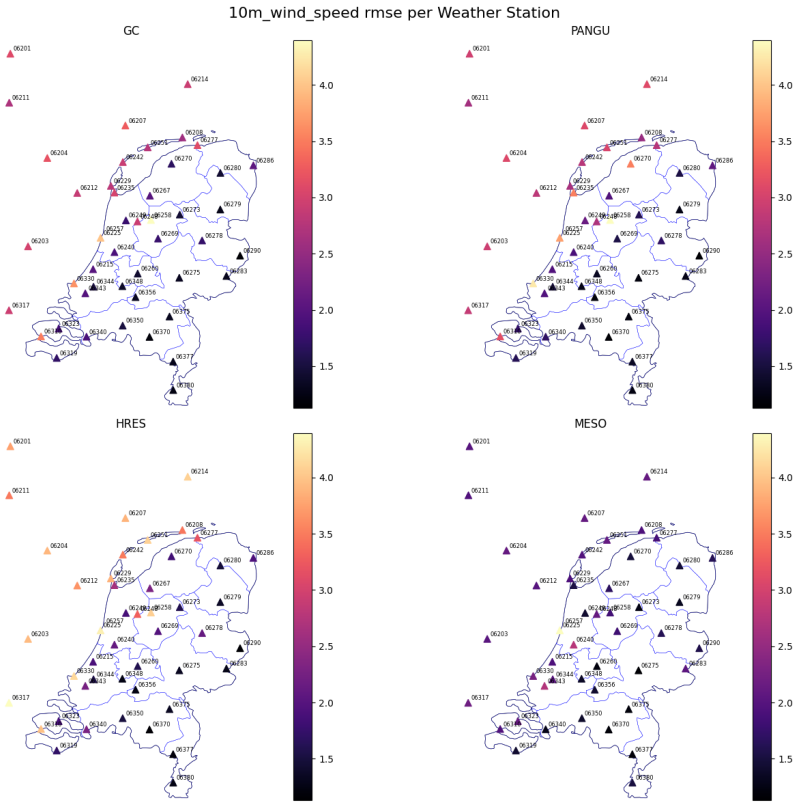
Furthermore, the bias patterns in wind speed forecasts indicate predominantly negative bias offshore and positive bias onshore, again mirroring the biases observed in the ERA5 and Analysis datasets (Figure 3.5). This means that models tend to underpredict wind speeds over the sea and overpredict them over land, which could be due to misrepresentation of surface roughness or other boundary layer parameters in the models. The AI models' replication of these bias patterns underscores the influence of training data on model performance and highlights areas where the underlying datasets may need refinement.

The impact of wind farms on forecast accuracy is also significant. At locations of offshore wind farms, such as Gemini (station 06214) and Borssele (station 06317), large positive biases are observed in all models. This is likely because the presence of wind turbines, which are not explicitly modeled, affects the wind speed by extracting energy from the flow and inducing turbulence. This effect is clearly seen in time series plots (Figure 4.8), where all models consistently over-forecast the 10 m wind speed at Borssele. This highlights how generalized models miss the localized effects of wind farms, indicating a need for models to account for such infrastructure to improve forecast accuracy in these areas.

In summary, the spatial patterns of forecast errors in the AI models and traditional NWP models reflect the biases and errors present in the ERA5 and Analysis datasets from which they were trained or developed. Recognizing these inherited patterns is crucial for guiding future model development and data assimilation strategies to enhance forecast accuracy across different geographical regions.



(a) 2 m temperature RMSE



(b) 10 m wind speed RMSE

Figure 4.6: Root Mean Square Error (RMSE) per weather station for four models: GraphCast (top left), Pangu-Weather (top right), HRES (bottom left), and Aspire Meso (bottom right). RMSE values are averaged over all lead times for the entire year of 2022, with darker colors representing lower error and lighter colors indicating higher error.

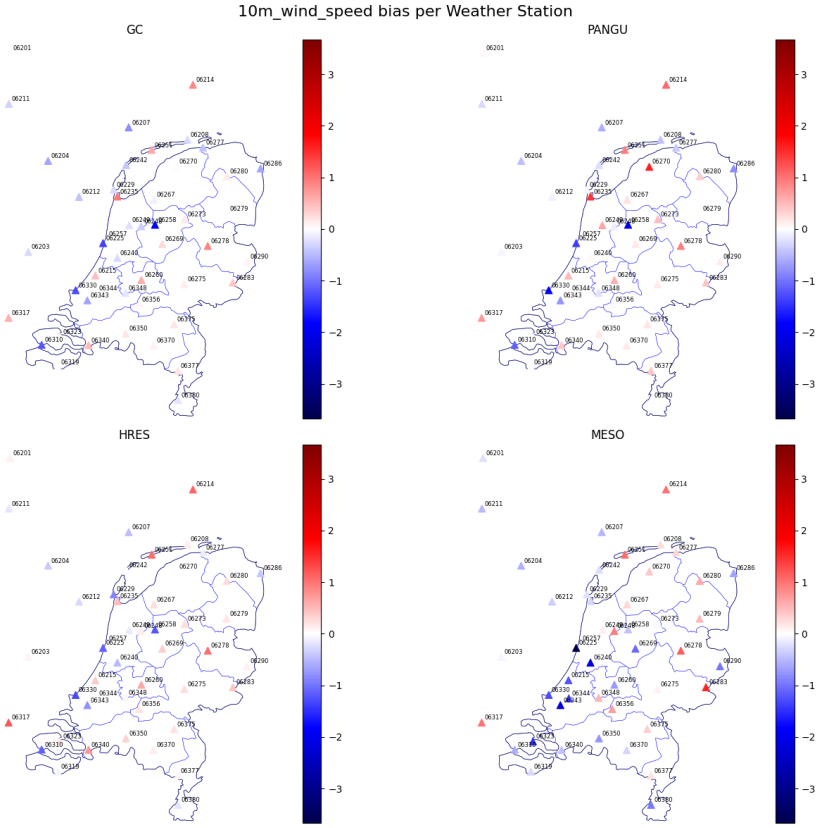


Figure 4.7: 10 m wind speed bias per weather station for four models: GraphCast (top left), Pangu-Weather (top right), HRES (bottom left), and Aspire Meso (bottom right). Bias is averaged over all lead times and for the entire year of 2022. Cooler colors represent larger negative bias, while warmer colors represent higher positive bias.

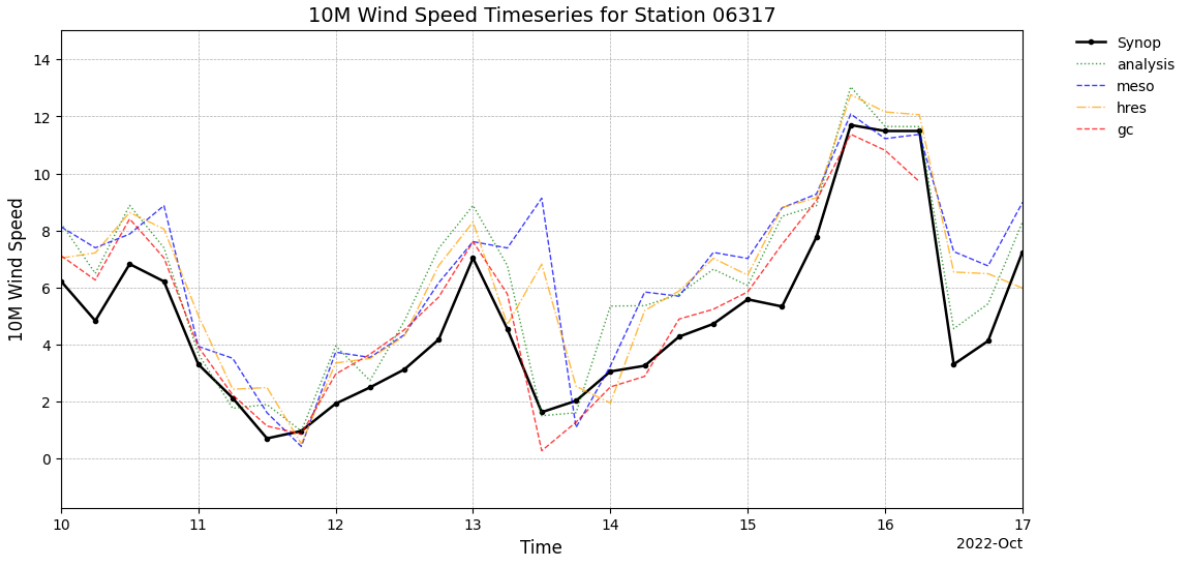


Figure 4.8: Time series plots for 10 m wind speed from October 10 to October 17, 2022, at station 06317 at the Borssele offshore wind farm. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue. All models tend to forecast the wind speed to be higher than the observations, indicating the effect a wind farm has on the wind speed and demonstrating how generalized models miss these effects.

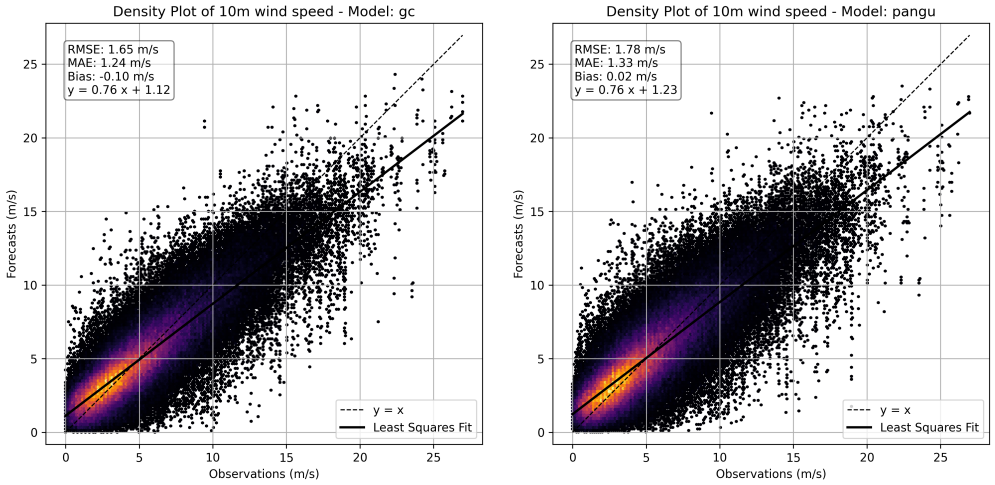
4.5. Scatter plots

In this section, we present the scatter density plots of 10m wind speed forecasts against observed values for four different weather models: GraphCast, Pangu-Weather, HRES, and Aspire Meso. Figure 4.9 displays the scatter plots, where each model's forecast is plotted against observed values with color intensity indicating the density of data points. The dashed line represents the ideal 1:1 relationship, and the solid line is the linear least-squares fit.

From these plots, we observe that all models have a slope less than 1, which indicates an increasing negative bias at higher wind speeds and a slight positive bias at lower wind speeds. This suggests that, while the models generally perform well at lower wind speeds, they tend to underpredict as wind speed increases, diverging more significantly from observations in the upper ranges.

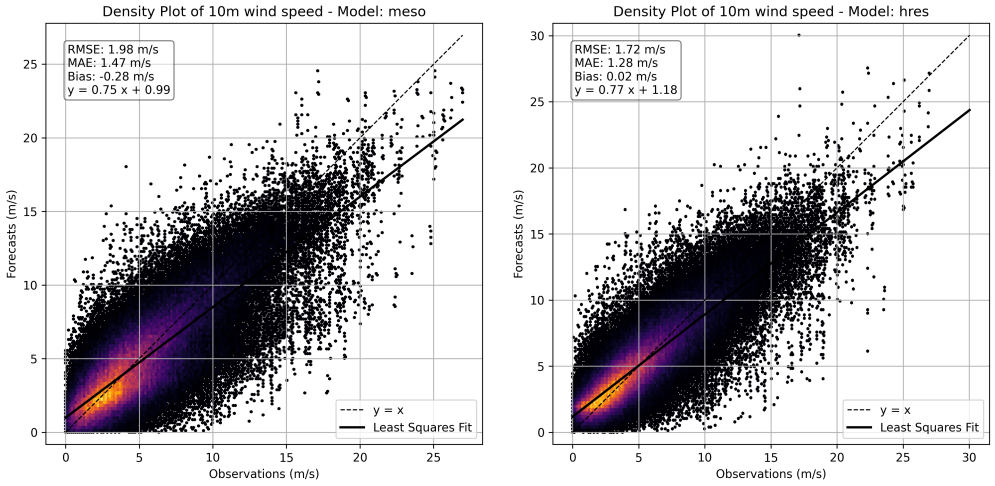
Additionally, we find that GraphCast and HRES demonstrate tighter clustering around the y=x line, which is reflected in their lower RMSE and MAE scores compared to Pangu-Weather and Aspire Meso models. This tighter grouping indicates a closer alignment with observed values, highlighting these two models as more accurate in their forecasts. Between GraphCast and HRES, GraphCast shows a slight performance edge, providing the most reliable predictions among the models analyzed.

Aspire Meso shows marginally better performance in predicting high wind events than the other models. When looking at these high wind events (around the 25 m/s observed wind speed line) we find that meso shows higher forecasted values. while not perfectly alligned with y=x, there is noteably better performance during these storm events.



(a) GraphCast

(b) Pangu-Weather



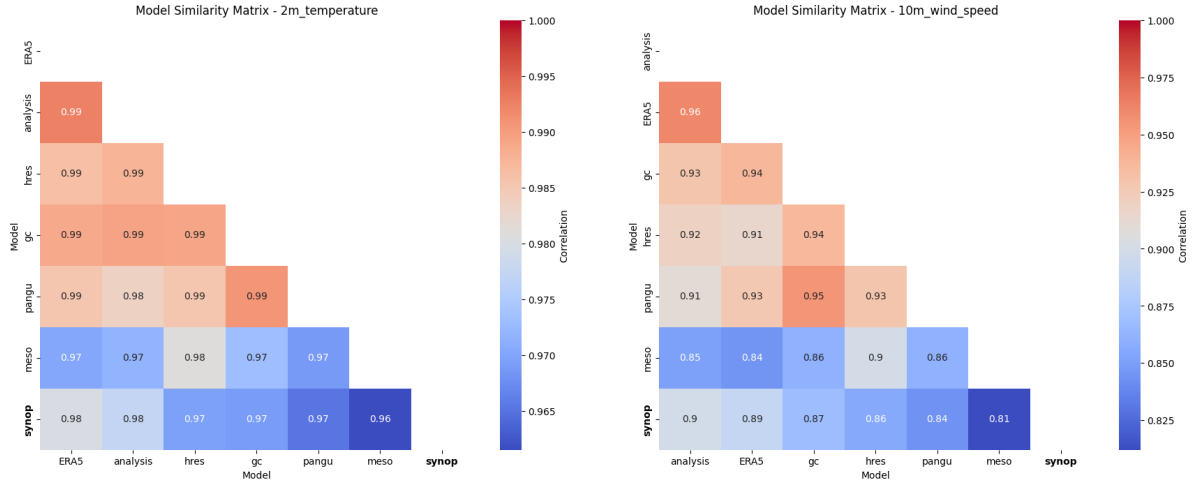
(c) Aspire Meso

(d) HRES

Figure 4.9: Density plots of 10m wind speed forecasts against observations for different weather models. Each plot shows the forecast data compared to observed values, with color intensity representing data density. The dashed line indicates the ideal y=x relationship, while the solid line represents the least-squares fit for each model. Error metrics (RMSE, MAE, Bias) are included to summarize model performance.

4.6. Inter-model Similarity

Model similarity matrices provide insights into the agreement between different models. They are constructed using the Pearson correlation coefficient between the outputs of each pair of models for a given variable. Figure 4.10 shows the model similarity matrices for 2 m temperature and 10 m wind speed.



(a) 2 m Temperature

(b) 10 m Wind Speed

Figure 4.10: Model similarity matrices showing the Pearson correlation coefficients between different models and observations for (a) 2 m temperature and (b) 10 m wind speed. Higher values indicate stronger agreement between models and observations.

The model similarity matrices (Figure 4.10) provide insights into how closely the models agree with each other and with the observations. The analysis reveals a very strong agreement among Pangu-Weather, GraphCast, HRES, and the ECMWF Analysis. Specifically, for 10 m wind speed, the Pearson correlation coefficient between Pangu-Weather and GraphCast is 0.955, which is higher than the correlation between GraphCast and ERA5 (0.938) or GraphCast and the Analysis (0.929). This suggests that Pangu-Weather and GraphCast are capturing very similar patterns in the data. This further strengthens the conclusions drawn in the Spatial Analysis, section 4.4.

Aspire Meso remains distinct from the other models but still shows significant correlation with them, notably with HRES (0.899). This could be because Aspire Meso's boundaries are set by HRES at initialization time, causing its outputs to be influenced by HRES data. This interdependence highlights that despite differences in model configurations and resolutions, there is a degree of similarity in their outputs due to shared data inputs or initial conditions.

When comparing the models to observations (denoted as "synop"), we see that the correlations are generally lower than the inter-model correlations. For example, the correlation between Pangu-Weather and observations is 0.844, while between GraphCast and observations it is 0.868. This suggests that while the models agree strongly with each other, they are less aligned with the actual observations. This could be due to various factors such as model biases, differences in spatial and temporal resolutions, limitations in the observational data or overfitting to training data that already contains a baseline error (section 3.3).

These findings highlight that while the models are highly consistent with each other, discrepancies exist when compared to observations. The highest correlation with observations is seen with the Analysis (0.898), followed by ERA5 (0.890), and GraphCast (0.868). Aspire Meso shows the lowest correlation with observations (0.812), indicating that it deviates more from the observed data compared to the other models.

The high inter-model correlations suggest that the models are capturing similar patterns in the data, which is expected given their similar training data and initialization procedures. However, this also raises concerns about the lack of diversity in model approaches. If models are too similar, they may

collectively miss certain patterns or errors that could be captured by more diverse modeling strategies. This underscores the importance of developing models with different methodologies or incorporating independent data sources to enhance the robustness of weather forecasts.

In conclusion, we observe that AI-based models like Pangu-Weather and GraphCast are closely aligned with traditional numerical weather prediction models like HRES and reanalysis datasets like ERA5 and the ECMWF Analysis. The high degree of similarity among these models indicates that they are effectively capturing the large-scale patterns in the Data. However, the lower correlations with observations highlight the need for continuous validation against real-world data to improve model accuracy as well emphesize the need for diverse training datasets.

4.7. Assessment of Ensemble Model Performance

In this section, we assess the performance of ensemble models compared to individual weather prediction models using the Root Mean Square Error (RMSE) metric for the 10-meter wind speed variable. The models considered are GraphCast (*gc*), HRES (*hres*), Pangu-Weather (*pangu*), and Aspire Meso (*meso*). We create ensembles by combining pairs of these models with equal weighting (following section 3.5) and evaluate the RMSE of these ensembles.

It is important to note that the ensembles in this study are equally weighted, meaning each model contributes equally to the ensemble prediction. This approach differs from operational settings, where ensemble members are often weighted based on their historical performance or optimized using machine learning techniques and historical data. Such methods can potentially improve the ensemble's predictive accuracy by assigning more weight to better-performing models.

Figure 4.11 presents a 3D bar chart that visualizes the RMSE performance of the individual models and their ensembles for the 10-meter wind speed. The base models are represented along the Y-axis, while the ensemble models are along the X-axis. The height of each bar corresponds to the RMSE value, with lower values indicating better performance. The color gradient of the bars represents the percentage improvement or degradation relative to the base model, providing a visual cue of the ensemble's effectiveness for that model in particular.

From Figure 4.11, we observe that equally weighted ensemble models generally exhibit improved performance over individual models, as evidenced by lower RMSE values. For instance, the ensemble of GraphCast (*gc*) and HRES (*hres*) achieves an RMSE of 1.6034, which is lower than the RMSE of either individual model (*gc*: 1.6518, *hres*: 1.7191). This represents a percentage improvement of approximately 2.93% relative to GraphCast.

Similarly, the ensemble of Aspire Meso (*meso*) and Pangu-Weather (*pangu*) yields an RMSE of 1.7277, which is a significant improvement over the *meso* model's RMSE of 1.9821. This corresponds to a percentage improvement of about 12.83% relative to *meso*.

However, the improvement is not uniform across all ensembles. In some cases, the ensemble does not outperform the better of the two individual models. For example, the ensemble of Pangu-Weather (*pangu*) and GraphCast (*gc*) results in an RMSE of 1.6644, which is higher than the RMSE of GraphCast alone (1.6518). This indicates a slight degradation in performance, with a percentage change of -0.76% relative to GraphCast.

These results suggest that while equally weighted ensembling can enhance model performance by combining the strengths of different models, it is not guaranteed to do so in all cases. The effectiveness of an ensemble depends on the complementarity of the models involved. In operational settings, ensemble members are often weighted based on their merit or optimized using historical data and machine learning algorithms. Such weighting schemes can potentially improve ensemble performance by assigning greater influence to more accurate models.

In conclusion, the equally weighted ensemble of certain models leads to a notable reduction in RMSE for the 10-meter wind speed variable, indicating improved predictive accuracy. The ensemble of GraphCast and HRES, in particular, shows the most significant improvement. Nonetheless, the results highlight the importance of weighting ensemble members appropriately, as some combinations may not yield the desired enhancement in performance when weighted equally.

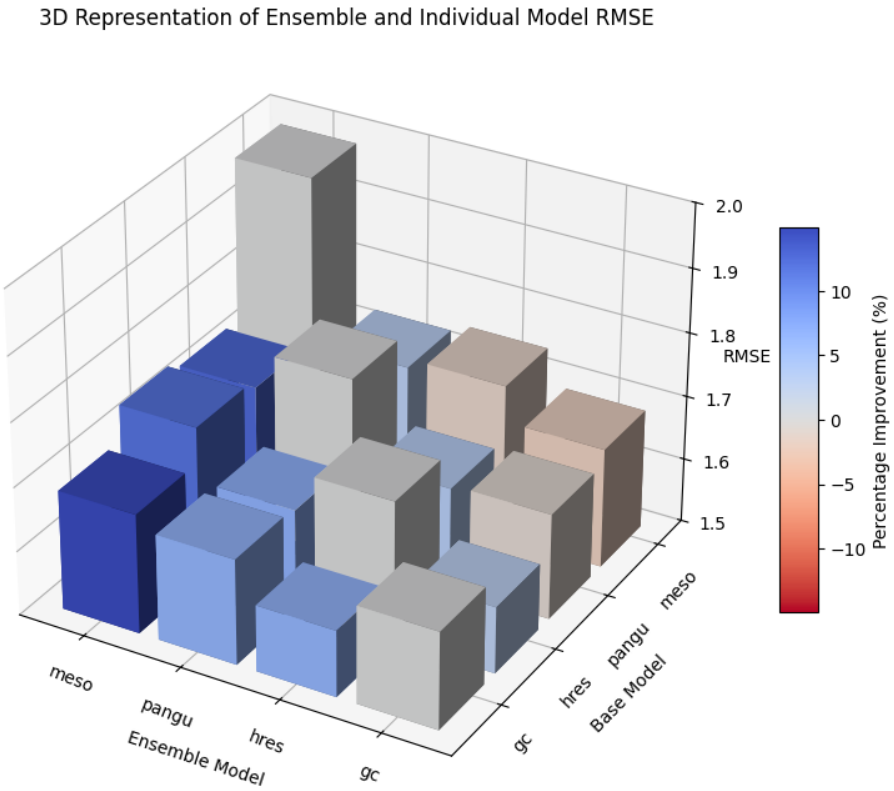


Figure 4.11: 3D representation of RMSE values for individual models and their equally weighted ensembles for the 10-meter wind speed variable. The X-axis represents the ensemble model, the Y-axis represents the base model, and the Z-axis indicates the RMSE value. The color gradient shows the percentage improvement relative to the base model.

4.8. Analysis of Large Eddy Simulation Model Performance

In this section, we extend our analysis to include the Aspire LES model, which was run at two specific locations among the 47 KNMI SYNOP stations: station 06203, P11-B (offshore), and station 06348, Cabauw (onshore). The LES model provides high-resolution insights into atmospheric dynamics, particularly at finer temporal scales and shorter forecast horizons. However, for consistency with the other models, we focus on the 10 m wind speed forecasts at lead times up to 90 hours.

To assess the performance of the LES model relative to other forecasting models, we employ Taylor diagrams, which succinctly summarize multiple statistical metrics—correlation coefficient, standard deviation, and root mean square deviation (RMSD)—in a single plot. Figure 4.12 displays Taylor diagrams for the 10 m wind speed at four lead times: 24, 48, 72, and 90 hours, comparing the LES model with other models including HRES, Aspire Meso, GraphCast, Pangu-Weather, and the ECMWF Analysis.

The Taylor diagrams reveal several key observations:

Firstly, the LES model does not outperform the other models in terms of RMSD and correlation coefficient. However, it exhibits a higher standard deviation compared to most models, indicating greater variability in its forecasts. This higher standard deviation persists despite the LES data being averaged hourly, suggesting significant potential for smoothing the LES outputs to reduce variability and potentially improve forecast skill.

An interesting comparison arises between Aspire Meso and Aspire LES. The LES model, while having a higher standard deviation, outperforms the mesocale model in terms of RMSD and correlation coefficient. This indicates that the LES model captures certain aspects of the wind speed dynamics better than the mesoscale model, despite its increased variability.

At the 90-hour lead time, we can distinguish three distinct groupings among the models:

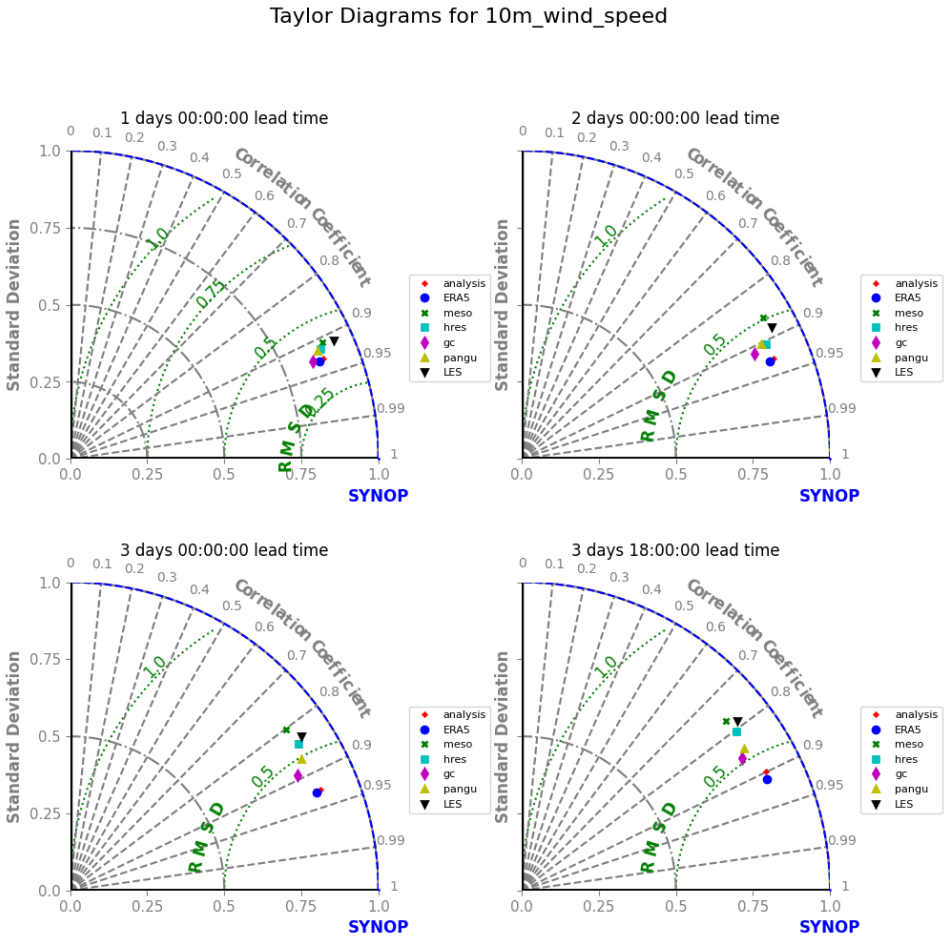


Figure 4.12: Taylor diagrams for the 10 m wind speed, showing model performance at four lead times: 24, 48, 72, and 90 hours. The diagrams compare different models (HRES, GraphCast, Pangu-Weather, Aspire Meso, Aspire LES, ECMWF Analysis) in terms of standard deviation, correlation coefficient, and RMSD, with "SYNOP" representing the reference observations.

1. **The (Re)Analysis Datasets**: ERA5 and ECMWF Analysis are grouped together, showing consistent performance across all lead times with RMSD around 0.3, standard deviation approximately 0.8, and correlation coefficient around 0.92. This grouping represents the highest performance in terms of alignment with observations.
2. **The AI-Based Models**: GraphCast and Pangu-Weather display very similar standard deviations to the (re)analysis datasets but with slightly worse RMSD and correlation coefficients. Pangu-Weather performs marginally worse than GraphCast, accompanied by a slightly higher standard deviation.
3. **The NWP Models**: HRES, Aspire Meso, and Aspire LES exhibit lower performance in terms of RMSD and correlation coefficient compared to the AI models and (re)analysis datasets. Notably, Aspire Meso and Aspire LES have higher standard deviations, suggesting more variability in their forecasts and, consequently, more opportunity for smoothing to enhance forecast skill.

Furthermore, all models show a decrease in correlation coefficient with increasing lead time, reflecting the common challenge in weather forecasting where predictive skill diminishes over longer forecast horizons. Despite the LES model's higher variability, its ability to outperform the mesocale model in some statistical metrics suggests that it captures certain atmospheric processes more effectively, particularly at the specific locations studied.

It is also important to note that the presented analysis does not include the initial six hours of the forecast horizon or the high temporal resolution data at 10-minute intervals that the LES model provides. These time frames and resolutions are particularly relevant for the LES model, which is designed to capture

fine-scale atmospheric processes during these periods. The exclusion of this data means that the full capabilities of the LES model are not entirely represented in the current analysis.

An additional note to be made is that Aspire is not optimized for surface wind speeds, is this is less relevant to wind power production, which generally focusses on 100m wind speed.

For a more comprehensive understanding of the LES model's performance, especially during the early forecast hours and at higher temporal resolutions, a detailed analysis is provided in Appendix B. This appendix delves deeper into the model's behavior, including its performance as a function of lead time.

4.9. Summary and Conclusions

In this chapter, we conducted a comprehensive evaluation of various weather forecasting models by comparing their outputs with observational data from 47 KNMI SYNOP stations across the Netherlands. The models assessed included GraphCast, Pangu-Weather, HRES, Aspire Meso, and Aspire LES, focusing on key atmospheric variables such as 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total accumulated precipitation over 6 hours.

Our analyses encompassed multiple approaches:

- **Visualization of Model Forecasts**: We examined the models' ability to capture significant weather events, specifically Storms Dudley and Eunice. The visualizations showed that while models effectively predicted the general patterns of wind speed increases associated with the storms, there were discrepancies in the precise timing and magnitude of peak winds. Notably, all models exhibited a timing error of approximately 6 hours during Storm Eunice, underpredicting peak wind speeds.
- **Time Series Analysis**: The time series plots at specific locations highlighted that models could capture general trends in mean sea level pressure and wind speed but struggled with the exact timing and intensity of severe weather events. Aspire Meso consistently underestimated mean sea level pressure, and all models underpredicted peak wind speeds during the storms.
- **Lead Time Analysis**: The evaluation of forecast errors over increasing lead times revealed that GraphCast generally outperformed other models, showing lower RMSE values across most variables and lead times. Diurnal patterns were evident in the bias of temperature, wind, and precipitation forecasts, reflecting the influence of daily cycles. The ensemble model, combining forecasts from different models, demonstrated improved accuracy, underscoring the benefits of ensemble forecasting.
- **Spatial Analysis of Forecast Errors**: Geographic patterns in forecast errors were identified, with higher temperature errors observed onshore and higher wind speed errors offshore. AI models like GraphCast and Pangu-Weather inherited error patterns from their training data (ERA5 and Analysis datasets), including biases such as underpredicting wind speeds offshore and overpredicting over land. The presence of wind farms also impacted forecast accuracy, indicating a need for models to account for such localized phenomena.
- **Scatter Plots**: The scatter density plots revealed that all models tended to underpredict higher wind speeds, with a slope less than one in the regression lines. GraphCast and HRES demonstrated tighter clustering around the ideal line, indicating better performance, while Aspire Meso showed marginally better predictions for high wind events.
- **Inter-model Similarity**: High Pearson correlation coefficients among models indicated strong agreement, particularly between AI models and traditional NWP models. However, correlations with observations were lower, suggesting that while models capture similar patterns, discrepancies exist when compared to real-world data.
- **Assessment of Ensemble Model Performance**: Equally weighted ensemble models generally exhibited improved performance over individual models, as evidenced by lower RMSE values. The ensemble of GraphCast and HRES showed the most significant improvement. However, in some cases, ensembles did not outperform the better of the two individual models, indicating the potential benefits of optimized weighting schemes in ensemble forecasting.

- **Analysis of Large Eddy Simulation Model Performance**: The Aspire LES model, evaluated using Taylor diagrams, did not outperform other models in terms of RMSD and correlation coefficient but showed higher standard deviation, indicating greater variability in forecasts. This suggests potential for smoothing the LES outputs to reduce variability and improve forecast skill.

Overall, the evaluation highlights that while current weather forecasting models demonstrate skill in predicting general weather patterns, challenges remain in accurately forecasting the timing and intensity of severe weather events. AI-based models like GraphCast and Pangu-Weather show promising performance but inherit biases from their training data, indicating a need for refinement in training datasets and methodologies. The ensemble approach enhances forecast accuracy, but equally weighted ensembles may not always yield the best results; optimized weighting based on model performance could further improve accuracy.

The spatial analysis underscores the importance of accounting for localized phenomena, such as wind farms, which significantly impact forecast accuracy in specific regions. Additionally, the models' tendency to underpredict high wind speeds and exhibit timing errors during severe events suggests that further improvements are needed in representing extreme weather phenomena.

Future work should focus on addressing these challenges by improving the representation of localized effects in models, refining training data to reduce inherited biases, and exploring advanced ensemble techniques with optimized weighting schemes. Enhancing the models' ability to accurately predict severe weather events will have significant benefits for weather-dependent industries and public safety.

5

Discussion

This chapter provides a comprehensive discussion of the research findings, their implications for weather forecasting, and addresses the limitations of the study. It also offers recommendations for future work.

5.1. Operational Considerations

Implementing weather forecasting models in operational settings requires careful consideration of computational costs, hardware requirements, and the specific needs of end-users. Models like GraphCast, Pangu-Weather, HRES, and Aspire each present unique challenges and advantages in practical applications.

5.1.1. Computational Costs and Hardware Requirements

GraphCast and **Pangu-Weather**, being AI-based global forecasting models, offer significant computational efficiency during inference compared to traditional numerical weather prediction (NWP) models. GraphCast's inference can be executed on modern GPUs with relatively low computational cost, though its memory footprint is substantial, necessitating access to high-end GPUs for operational use. Pangu-Weather boasts even faster inference times, generating a global 7-day forecast in just 1.4 seconds on a single GPU. However, both models have considerable training costs due to the vast amount of data and the complexity of their architectures, requiring substantial computational resources and time.

In contrast, **HRES**, the ECMWF's high-resolution deterministic model, demands significant computational resources for both development and operational runs. It relies on supercomputing facilities to solve complex physical equations at high spatial resolutions. The high operational costs and computational demands limit its accessibility for some users.

Aspire, designed for mesoscale and LES applications, balances computational cost with resolution but still requires considerable resources for running high-fidelity simulations, especially when resolving small-scale turbulence through Large Eddy Simulation (LES) nests. Its computational costs are lower when compared to global models due to its regional focus, but high-resolution LES runs can be computationally intensive.

Ensemble forecasting, which combines multiple model outputs, introduces additional computational considerations. While the computational cost of generating ensembles from pre-existing forecasts is minimal, the initial computational investment to generate individual forecasts from different models can be substantial, especially when including multiple high-resolution models.

5.1.2. Variable Selection and Relevance

The datasets utilized by these models encompass a range of meteorological variables critical for various applications. Variables like wind speed and direction at different altitudes are essential for wind energy forecasting, while temperature and precipitation forecasts are crucial for agriculture and water resource management. The inclusion of specific variables varies across models, affecting their applicability to

certain sectors.

For instance, **Pangu-Weather** does not include precipitation in its forecast outputs, limiting its utility for hydrological applications or regions where precipitation forecasting is critical. Additionally, some models may lack variables such as solar radiation or cloud cover, which are important for solar energy forecasting and climate studies.

A significant challenge lies in translating model outputs to specific heights or locations not directly represented in the model grids. For example, forecasting wind speeds at 100 meters above ground level is vital for wind farm operations but may not be directly available from the model outputs. Interpolation or the use of diagnostic models is often required to estimate conditions at these specific heights, introducing additional uncertainties.

5.2. Model Performance and Economic Implications

5.2.1. Performance Analysis of Models

The comparative analysis of the models indicates distinct performance trends. **GraphCast** consistently demonstrates superior accuracy over **HRES**, **Pangu-Weather**, and **Aspire Meso** across various meteorological variables and forecast lead times. This enhanced performance is evident in metrics such as Mean Absolute Error (MAE) and Bias, where GraphCast outperforms its counterparts.

Pangu-Weather also shows competitive performance, generally surpassing HRES and Aspire Meso in some variables and lead times. However, its omission of certain variables, like precipitation, limits its overall applicability.

All models exhibit baseline errors at initial forecast steps, primarily due to uncertainties in initial conditions and reliance on shared data sources like ERA5 reanalysis and ECMWF Analysis data for initialization. These errors highlight the intrinsic limitations in capturing the exact state of the atmosphere at the onset of the forecast period.

GraphCast

GraphCast's primary strength lies in its superior accuracy and computational efficiency during inference. Its ability to capture complex spatial and temporal patterns in atmospheric data contributes to its enhanced performance. However, despite these advantages, GraphCast suffers from similar issues as traditional models concerning spatial and diurnal biases. The model may not accurately capture localized phenomena or diurnal cycles, particularly in regions with complex terrain or microclimates.

A significant limitation of GraphCast is its black-box nature inherent to AI models. The lack of interpretability makes it challenging to diagnose and address specific weaknesses within the model. Unlike physics-based models, where adjustments to physical parameterizations can be made to improve performance in known problematic areas, modifying GraphCast to correct inherent problems is not straightforward. This limitation hinders the ability to refine the model based on physical understanding and targeted improvements.

Pangu-Weather

Pangu-Weather demonstrates competitive performance, often outperforming HRES and Aspire Meso in certain variables like temperature and geopotential height. Its use of a 3D Earth-specific transformer architecture allows it to capture spatial dependencies effectively. However, Pangu-Weather's exclusion of key variables such as precipitation limits its applicability, particularly for forecasting extreme weather events or hydrological applications.

Similar to GraphCast, Pangu-Weather relies heavily on data from models like HRES for initialization and training, potentially inheriting their biases. Its computational efficiency during inference is a notable advantage, but the substantial training cost and resource requirements present challenges for operational deployment.

HRES

HRES serves as a foundational numerical weather prediction model, utilizing sophisticated physical equations to simulate atmospheric processes. Its strengths include detailed representation of atmo-

spheric dynamics and comprehensive assimilation of observational data through four-dimensional variational (4D-Var) techniques. The 4D-Var data assimilation system is instrumental in generating the Analysis dataset, often considered as the ground truth in modeling studies.

However, HRES is computationally intensive, requiring significant resources to run at high resolutions. This computational cost limits its accessibility and frequency of updates in operational settings. Moreover, while HRES provides high-quality deterministic forecasts, the ECMWF places considerable emphasis on its ensemble prediction system (EPS) to improve probabilistic forecasting. The ensemble approach addresses the inherent uncertainties in initial conditions and model physics.

Despite being one of the most advanced global NWP models, discrepancies exist between HRES forecasts and actual observations. These differences highlight the limitations in model physics and data assimilation, especially in regions with sparse observational coverage or complex weather phenomena.

Aspire Meso

Aspire Meso, a mesoscale modeling system, exhibits comparatively worse performance in the evaluations. One contributing factor is the temporal downsampling from its native 10-minute intervals to 6-hour intervals, which results in a loss of variance and the smoothing of short-term variability critical for accurate forecasting.

The core design of Aspire Meso focuses on bridging the gap between coarse-gridded boundary conditions provided by global NWP models and the high-resolution requirements of localized simulations, particularly through LES nests. This specialization makes it adept at capturing small-scale processes in limited areas but less effective when evaluated over broader regions or longer lead times.

A significant challenge with Aspire Meso is the computational tradeoff between spatial resolution and the extent to which turbulence is resolved versus parameterized. High-resolution simulations that resolve turbulence require substantial computational resources, limiting their practicality for operational use over large domains or extended forecast periods.

Aspire LES

The **Aspire LES** model focuses on high-resolution, localized simulations using Large Eddy Simulation (LES) to resolve small-scale turbulent processes in the atmospheric boundary layer. In this study, Aspire LES was applied to two specific station locations: P11-B (offshore) and Cabauw (onshore), providing detailed forecasts at these sites.

Performance Analysis:

Aspire LES demonstrated superior performance in capturing fine-scale atmospheric features and turbulence compared to other models at the specific stations. The high spatial and temporal resolution allowed for better representation of local weather phenomena, such as boundary layer dynamics and wind shear, which are critical for applications like wind energy forecasting.

At the offshore station P11-B, Aspire LES showed improved accuracy in predicting wind speed and direction, outperforming other models, especially during periods of complex atmospheric conditions. Similarly, at the onshore station Cabauw, Aspire LES provided more accurate temperature and wind profiles near the surface.

Strengths:

- *High Resolution:* Aspire LES's ability to resolve small-scale turbulence leads to more accurate and detailed forecasts at specific locations.
- *Local Adaptability:* The model can be tailored to specific sites, making it valuable for localized forecasting needs, such as wind farm operations or urban weather forecasting.

Limitations:

- *Computational Cost:* The high-resolution simulations require substantial computational resources, limiting the practicality of Aspire LES for operational use over large domains or extended forecast periods.

- *Limited Spatial Extent:* Aspire LES's localized nature means it cannot provide forecasts over broader regions without significant computational investment.
- *Data Requirements:* The model relies on accurate boundary conditions from mesoscale models like Aspire Meso, inheriting any errors or biases from them.

Despite these limitations, Aspire LES offers valuable insights into local atmospheric conditions that are not captured by global models like GraphCast or Pangu-Weather. Its detailed simulations are particularly beneficial for applications that require high-resolution data, albeit at a higher computational expense.

5.2.2. Performance of Ensemble Forecasts

The inclusion of ensemble forecasting methods in this study reveals that combining forecasts from multiple models can improve overall forecast accuracy. Ensembles constructed by averaging outputs from models like GraphCast, HRES, Pangu-Weather, and Aspire Meso generally exhibit lower MAE and Bias compared to individual models, particularly for variables where models have complementary strengths.

For instance, the ensemble combining **GraphCast** and **Pangu-Weather** demonstrates improved accuracy in temperature forecasts, leveraging the strengths of both AI models. Similarly, ensembles including **HRES** can benefit from its detailed physical modeling, enhancing forecasts for variables like precipitation, which Pangu-Weather lacks.

However, the benefits of ensemble forecasting are somewhat constrained by the models' reliance on similar data sources and underlying biases. The ensembles do not achieve the full potential of error reduction due to the correlated errors among the models. Additionally, simple equal-weighted averaging may not be the most effective method, and more sophisticated weighting schemes based on model performance could yield better results.

5.2.3. Discussion on Imbalance Costs

The analysis provides a rough estimate of imbalance costs due to wind speed forecast errors for the Gemini wind farm. As summarized in Table 5.1, even the baseline model (HRES) has a significant average error of 1.153 m/s at the 24-hour lead time, resulting in an annual imbalance cost of approximately €12,614,400. The **GraphCast** model, with a slightly lower error, reduces this cost by about €213,400, while **Pangu-Weather** shows an increase in cost due to higher error. The **Aspire Meso** model, with a larger error, increases the cost by more than €1,470,179. Notably, the four-model ensemble offers the lowest cost among all models, yielding a reduction of about €609,670 compared to the baseline.

These estimates underscore the financial impact of forecast accuracy, highlighting that even small improvements can lead to meaningful cost reductions. Ensemble forecasts, by further reducing errors, have the potential to decrease imbalance costs more significantly, highlighting the economic benefits of adopting ensemble forecasting strategies.

Table 5.1: Imbalance Cost Impact for Different Weather Forecasting Models

Model	Wind Speed Error (m/s)	Imbalance Cost (€)	Cost Difference (€)
Baseline (HRES)	1.153	12,614,400	–
GraphCast	1.134	12,400,996	-213,404
Pangu-Weather	1.251	13,682,563	+1,068,163
Aspire Meso	1.288	14,084,579	+1,470,179
Ensemble	1.097	12,004,730	-609,670

5.3. Implications of Neural Scaling Laws for Weather Forecasting Models

The development of AI-based weather forecasting models like GraphCast and Pangu-Weather has been influenced by advances in deep learning architectures and training methodologies. An important

consideration in the field of machine learning is how model performance scales with increases in model size, dataset size, and computational resources, as described by Kaplan et al. (2020) in their work on neural scaling laws for language models. This section discusses the implications of these scaling laws for weather forecasting models, particularly in light of the observed performance of GraphCast and Pangu-Weather, their parameter counts, training compute usage, and the intermodel correlation in errors.

5.3.1. Scaling Laws and Model Performance

The scaling laws proposed by Kaplan et al. (2020) indicate that model performance improves predictably with increases in model parameters, dataset size, and compute used during training, following power-law relationships. Larger models trained on more data with greater computational effort tend to perform better, assuming none of these factors are bottlenecks. However, this improvement follows a trend of diminishing returns, where each additional unit of resource yields smaller gains in performance.

5.3.2. Model Size and Training Compute

In the context of weather forecasting models, **GraphCast** consists of 36.7 million parameters, while **Pangu-Weather** has 256 million parameters. Despite Pangu-Weather being approximately seven times larger, GraphCast demonstrates superior performance across various meteorological variables and forecast lead times. Additionally, the training cost for Pangu-Weather is significantly higher, involving a cluster of 192 NVIDIA Tesla V100 GPUs over 16 days for each model variant, whereas GraphCast was trained using 32 Cloud TPU v4 devices over four weeks.

This disparity raises questions about the applicability of the scaling laws in the domain of weather forecasting. If larger models and more compute generally lead to better performance, as suggested by the scaling laws, why does GraphCast, a smaller model, outperform the larger Pangu-Weather?

5.3.3. Data Quality and Model Architecture

One possible explanation lies in the quality of the training data and the efficiency of the model architectures. Both GraphCast and Pangu-Weather are trained on the ERA5 reanalysis dataset, which, despite being comprehensive, has inherent limitations and biases. The scaling laws assume that data quality is sufficient and that the model capacity can be effectively utilized. In this case, the data may not provide enough additional information for a larger model like Pangu-Weather to leverage effectively, leading to diminishing returns in performance.

Furthermore, GraphCast employs a novel architecture utilizing Graph Neural Networks (GNNs) and a multimesh representation, enabling it to capture both local and global atmospheric patterns efficiently. Pangu-Weather, while utilizing a sophisticated 3D Earth-specific Transformer architecture, may not exploit the available data as effectively as GraphCast. This suggests that architectural efficiency and suitability to the problem domain can have a significant impact on performance, potentially outweighing the benefits of simply scaling up model size.

5.3.4. Intermodel Correlation and Data Bottlenecks

The high intermodel correlation observed between GraphCast and Pangu-Weather indicates that both models make similar errors, likely due to being trained on the same data source (ERA5). This suggests a data bottleneck, where the limitations of the training data prevent models from improving beyond a certain point, regardless of their size or complexity. In such cases, increasing model size may not lead to better performance, as the models are constrained by the quality and diversity of the data they are trained on.

5.3.5. Limitations of Scaling in Weather Forecasting

The application of scaling laws in weather forecasting models appears to be limited by several factors:

- **Data Quality and Diversity**: The ERA5 dataset, while extensive, may not contain enough diversity or resolution to support significant gains from larger models. The presence of biases and errors in the reanalysis data can limit the maximum achievable performance.
- **Model Capacity vs. Data Complexity**: If the complexity of the atmospheric data does not match

the increased capacity of larger models, the additional parameters may not contribute to better performance.

- **Overfitting and Generalization**: Larger models are more prone to overfitting, especially when the training data is limited in diversity. Without sufficient high-quality data, larger models may not generalize better than smaller ones.
- **Architectural Efficiency**: Efficient model architectures that are well-suited to capturing the complexities of atmospheric dynamics can lead to performance improvements without necessarily increasing model size. GraphCast's use of GNNs may provide a more effective way of modeling atmospheric processes than Pangu-Weather's transformer-based approach.

5.3.6. Implications for Future Model Development

The observations from this study suggest that simply scaling up model size and compute does not guarantee better performance in weather forecasting models. Instead, focusing on:

- **Improving Data Quality**: Enhancing the quality and resolution of training data, possibly by incorporating more observational data or higher-resolution reanalyses, could help models better utilize increased capacity.
- **Architectural Innovation**: Developing model architectures that are better suited to capturing the complexities of atmospheric dynamics can lead to performance improvements without necessarily increasing model size.
- **Data-Efficient Training**: Employing training strategies that maximize the effective use of available data, such as transfer learning or data augmentation, can improve model performance.
- **Understanding Model Limitations**: Recognizing that there may be asymptotic limits to performance imposed by data constraints, and that beyond a certain point, scaling models further yields diminishing returns.

5.3.7. Conclusion

The scaling laws for neural language models provide valuable insights into how model performance can be improved by increasing model size, data, and compute. However, in the domain of weather forecasting, these laws may not fully apply due to data limitations and the specific challenges of modeling atmospheric processes. The superior performance of GraphCast, despite its smaller size compared to Pangu-Weather, highlights the importance of efficient model architectures and the effective use of available data. Future work should focus on improving data quality, exploring innovative architectures, and understanding the limits of model scaling in weather forecasting.

5.4. Model Diversity

A notable observation from the study is the reliance of all evaluated models on HRES data for training, initialization, or boundary conditions. This dependence results in overlapping error patterns and limits the diversity of forecast outputs. The similarity matrices presented earlier highlight the correlations between model errors, underscoring the lack of independence among the models.

This reliance on a single source affects the robustness of the forecasting system. Errors or biases inherent in HRES propagate through the dependent models, reducing the overall reliability of the forecasts. Additionally, the lack of diversity may stifle innovation in model development, as alternative data sources and modeling approaches are underutilized.

Ensemble forecasting benefits from diversity among its members. However, the current ensembles' effectiveness is limited due to the correlated errors stemming from shared data sources and similar model architectures. Incorporating models trained on different data sets or using fundamentally different approaches could enhance ensemble performance by reducing error correlations.

Diversifying training data and models could enhance performance by introducing different perspectives and error characteristics. Incorporating data from other global models, observational networks, or reanalysis datasets could provide a more robust training set for AI models like GraphCast and Pangu-Weather.

5.5. Limitations

5.5.1. Limitations of AI Forecasting

AI-based forecasting models like GraphCast and Pangu-Weather offer significant advantages but also present specific limitations that warrant careful consideration. While computationally efficient during inference, the memory footprint of these models is substantial, necessitating the use of state-of-the-art GPUs. This requirement may limit accessibility for some operational users, particularly in resource-constrained environments.

Global AI models exhibit inflexibility when adapting to local scales. Models like GraphCast and Pangu-Weather require full global input, processing, and output, even if the area of interest is geographically limited. This operational constraint results in unnecessary computational overhead and challenges in tailoring the models to specific regional applications.

The dependence on HRES for training and initialization affects the independent performance of AI models. Since GraphCast and Pangu-Weather are trained on data derived from ERA5, which in turn relies heavily on models like HRES, there is a possibility of asymptotic behavior in forecast skill. Regardless of the model architecture, improvements may plateau due to limitations in the underlying training data quality and diversity.

This pattern mirrors observations in the development of large language models (LLMs), where architectural enhancements yield marginal gains compared to improvements derived from data quality and diversity. In LLMs, techniques like reinforcement learning with human feedback have led to significant advancements, but such approaches are infeasible at a global scale in weather forecasting due to the sheer volume and complexity of atmospheric data.

However, this limitation opens opportunities for globally trained models that can be locally fine-tuned. By adapting global AI models with localized data, it may be possible to enhance performance in specific regions without the need for retraining the entire model. This modularity could improve flexibility and applicability in diverse operational contexts.

The black-box nature of AI models poses challenges in understanding the underlying physical processes represented within them. This lack of interpretability makes it difficult to diagnose errors or biases and to trust the model outputs, especially in critical situations like extreme weather events. The inability to fully understand when and why the model may be wrong undermines confidence in its forecasts.

Despite these limitations, AI models like GraphCast and Pangu-Weather demonstrate excellent performance at extremely low inference costs compared to traditional NWP models. Their ability to rapidly produce forecasts with competitive accuracy offers significant benefits, particularly in scenarios where computational resources or time are constrained.

5.5.2. Limitations of This Study

While the study provides valuable insights into the performance of various weather forecasting models, several limitations must be acknowledged:

- **Temporal Limitations:** The study employs a short forecast window with coarse time intervals, which may overlook short-term variability and transient weather phenomena. The use of 6-hour intervals, in particular, may smooth out important fluctuations that occur on shorter timescales, limiting the assessment of models' capabilities in capturing rapid atmospheric changes.
- **Spatial Limitations:** Focusing on the Netherlands, a region with relatively uniform terrain and climate, restricts the diversity of conditions evaluated. This geographic limitation may not reflect the models' performance in areas with complex terrain, such as mountainous regions, or in different climatic zones, such as tropical or arctic environments.
- **Variable Limitations:** The exclusion of important meteorological variables like solar radiation and cloud cover narrows the scope of the evaluation. Pangu-Weather's omission of precipitation in its outputs further limits the assessment of models across all relevant variables.
- **Data and Model Limitations:** Dependence on specific datasets, such as SYNOP observations, may introduce biases related to data quality and coverage. Additionally, the rigid configurations of

the models evaluated prevent exploration of their full potential or adaptability. For instance, modifications to model physics or assimilation techniques were not considered, which could influence performance outcomes.

- **Scope and Generalizability:** Given these limitations, caution is warranted in generalizing the findings beyond the specific context of the study. Additional validation in diverse settings and with extended variables is necessary to fully assess the models' strengths and weaknesses. Overextending conclusions without such validation could lead to inaccurate assumptions about model performance in different operational scenarios.

5.6. Implications for Weather Forecasting

The findings from this study have several broader implications for the field of weather forecasting:

- **AI-Based Forecasting Performs:** After thorough assessment of both GraphCast, Pangu-Weather, and NWP alternatives, we can say that AI-based models generally perform better, at lower cost. Not just against the assumed ground truth (analysis) but also against observations. This further validates data-driven approaches in weather forecasting.
- **Global vs. Localized Forecasting:** A trade-off exists between the broad coverage provided by global models and the detailed insights offered by localized models. Global AI models like GraphCast and Pangu-Weather deliver comprehensive forecasts efficiently but may lack the resolution and specificity required for local applications. Conversely, localized physical models capture fine-scale processes but are computationally intensive and limited in spatial extent.
- **Need for Hybrid Approaches:** The limitations and strengths of both AI-based and physical models suggest a need for hybrid approaches. Combining global AI models with localized physical models could leverage the efficiency of AI while incorporating detailed physical representations where necessary. Such integration may enhance overall forecast accuracy and applicability across scales. Additionally, it might break through any asymptotic behavior following from ERA5.
- **Ensemble Forecasting Enhances Accuracy:** The study demonstrates that ensemble forecasting, even with simple equal-weighted averaging, can improve forecast accuracy. By combining outputs from different models, ensembles can mitigate individual model biases and errors, leading to more reliable forecasts.
- **Potential for Cost-effective Probabilistic Forecasting:** AI models offer potential for cost-effective probabilistic forecasting by rapidly generating ensemble members through perturbations or by using ensemble techniques within the AI framework. This capability could improve the accessibility of probabilistic forecasts, which are crucial for risk assessment and decision-making in various sectors.
- **Challenges in Short-term Forecasting:** Temporal resolution limitations in models like GraphCast and Pangu-Weather pose challenges for short-term forecasting. The inability to capture rapid atmospheric changes on timescales shorter than the model's output interval can hinder the model's utility in scenarios requiring high temporal fidelity, such as severe weather warnings or energy grid management.

5.7. Recommendations and Future Work

To enhance AI-based weather models and address identified limitations, several recommendations are proposed for future research.

5.7.1. Incorporate More Observational Data

Integrating additional observational data into model training could reduce baseline errors and enhance local accuracy. The opportunity exists to leverage globally trained models on datasets like ERA5, which effectively capture large-scale processes and long-range connections at a relatively low computational cost. Local fine-tuning using higher-resolution data could further improve performance in specific regions.

5.7.2. Enhance Model Modularity and Adaptability
There is architectural flexibility within models like GraphCast and Pangu-Weather that can be exploited to improve modularity and adaptability. By designing models that can be easily retrained or fine-tuned for different regions or applications, the utility and applicability of AI models can be significantly expanded.

5.7.3. Improve Temporal Resolution
Retraining models with higher time-resolved data would enhance their ability to capture short-term variability and improve performance in short-range forecasting. This enhancement is crucial for applications requiring high temporal fidelity.

5.7.4. Expand Variable Range
Including a broader range of meteorological variables, such as solar radiation, cloud cover, and precipitation (in the case of Pangu-Weather), would increase the models' applicability across various sectors. A more comprehensive set of outputs would meet the needs of diverse end-users and applications.

5.7.5. Emphasize Validation Against Observational Data
Shifting the focus of validation from reanalysis datasets to actual observational data would provide a more accurate assessment of model performance. This approach acknowledges the discrepancies between reanalysis products and real-world observations, leading to more reliable evaluations.

5.7.6. Explore Advanced Ensemble Techniques
Developing more sophisticated ensemble forecasting techniques, such as weighted averaging based on model performance or machine learning ensemble methods, could further improve forecast accuracy. Exploring methods to efficiently generate and optimize ensembles with AI models is a promising avenue for future research.

5.8. Conclusions
This study has evaluated the performance of various weather forecasting models, highlighting the superior accuracy and computational efficiency of AI-based models like GraphCast and Pangu-Weather. The findings demonstrate the potential of AI to enhance weather forecasting accuracy while reducing computational costs during inference.

The inclusion of ensemble forecasting shows promise in further improving forecast accuracy, although limitations due to correlated errors among models persist. Addressing these challenges through diversification of models and data sources can enhance ensemble performance.

However, limitations related to model flexibility, spatial and temporal resolution, and interpretability have been identified. Addressing these challenges is crucial for maximizing the benefits of AI in operational settings. Balancing model sophistication with practical considerations, such as hardware requirements and adaptability to local scales, will enhance the utility of these models.

Future work should build upon these findings by incorporating more diverse and higher-resolution data, improving model architectures for greater modularity, and exploring hybrid approaches that combine the strengths of AI and physical models. By addressing the limitations and expanding the applications of AI in weather forecasting, we can advance the field and improve the accuracy and reliability of forecasts critical for decision-making across various sectors.

6

Conclusions

This thesis has undertaken a comprehensive evaluation of AI-based weather forecasting models, specifically GraphCast and Pangu-Weather, in comparison to traditional numerical weather prediction (NWP) models like the High-Resolution Model (HRES) from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Aspire models developed by Whiffle. By focusing on the Netherlands—a region of significant interest due to its rapid wind energy development and availability of dense observational data—we aimed to assess the performance, operational considerations, and economic implications of these models in an operational setting.

6.1. Summary of Research

6.1.1. Theoretical Framework of Forecasting Models

The thesis began by exploring the theoretical foundations of the four distinct weather forecasting models:

- **GraphCast**: An AI-driven model utilizing Graph Neural Networks (GNNs) for medium-range global weather forecasting. It employs an encoder-processor-decoder architecture with a multimesh representation, capturing complex atmospheric interactions to deliver accurate 10-day forecasts.
- **Pangu-Weather**: Another AI-based global forecasting model that leverages a 3D Earth-specific Transformer architecture. It excels in rapid inference times, producing a global 7-day forecast in approximately 1.4 seconds on a single GPU.
- **HRES**: A deterministic NWP model by ECMWF that relies on the fundamental laws of physics, utilizing spectral transform methods, hybrid vertical coordinates, and semi-Lagrangian semi-implicit time integration to simulate atmospheric processes at high resolutions.
- **Aspire Meso and LES**: A mesoscale model and a Large Eddy Simulation (LES) model designed for high-resolution, turbulence-resolving simulations. They leverage GPU-based computation to perform detailed atmospheric modeling, particularly valuable for applications like wind energy assessments.

6.1.2. Research Design and Methodology

The research employed a robust design to compare the models:

- **Data Selection**: Utilized observational data from 47 KNMI SYNOP stations in the Netherlands, covering variables such as 2 m temperature, mean sea level pressure, 10 m wind speed and direction, and total precipitation over the last 6 hours.
- **Time Frame**: Generated and analyzed forecasts for the entire year of 2022, ensuring that GraphCast's training data did not overlap with the evaluation period.

- **Forecast Lead Time**: Chose a 96-hour (4-day) forecast window to balance computational efficiency with the ability to capture the breakdown of forecast skill over time.
- **Model Configurations**:
  - **GraphCast**: Employed the operational model with 0.25° resolution, initialized using HRES analysis data, and generated forecasts using autoregressive methods.
  - **Pangu-Weather**: Utilized pre-generated operational forecasts from WeatherBench2, aligned to match the evaluation criteria.
  - **HRES**: Used historic forecast data at 0.25° resolution from WeatherBench, downsampling from its original 0.1° resolution.
  - **Aspire Meso**: Configured the mesoscale model to cover the Netherlands with a 2000-meter horizontal grid spacing, focusing on capturing mesoscale phenomena. Aspire LES was also employed for high-resolution simulations at specific stations.
- **Evaluation Metrics**: Assessed model performance using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Bias, Standard Deviation, and Pearson Correlation Coefficient.
- **Dimensions of Analysis**: Analyzed results across initialization times, lead times, and station locations to provide a comprehensive evaluation.

6.1.3. Analysis of Forecast Performance

The comparative analysis revealed several key findings:

- **GraphCast's Superior Performance**: GraphCast consistently outperformed HRES, Pangu-Weather, and Aspire Meso across all variables and lead times, demonstrating its effectiveness in capturing complex atmospheric patterns with lower RMSE and Bias values.
- **Baseline Errors**: All models exhibited similar baseline errors at the initial forecast step (T = 6 hours), highlighting uncertainties in initial conditions and shared data sources like ERA5 reanalysis and ECMWF Analysis data.
- **Temporal and Spatial Biases**: Diurnal patterns in biases were observed, with models showing varying tendencies to overestimate or underestimate variables like temperature and wind speed depending on the time of day and location. Higher temperature errors were noted onshore, while wind speed errors were higher offshore.
- **Aspire Meso's Limitations**: Aspire Meso showed higher variance and consistent elevation bias in mean sea level pressure, partly due to temporal downsampling and its focus on resolving small-scale turbulence.
- **Inter-Model Similarity**: High correlation between GraphCast, Pangu-Weather, HRES, and the ECMWF Analysis indicated a lack of diversity in model approaches, as all relied heavily on HRES data for training or initialization. This led to overlapping error patterns and limited the effectiveness of ensemble forecasting.
- **Aspire LES Performance**: Aspire LES provided superior performance in capturing fine-scale atmospheric processes at specific stations, particularly in resolving small-scale turbulence critical for applications like wind energy forecasting. However, its high computational cost and limited spatial extent restrict its practicality for broader operational use.
- **Economic Implications**: Differences in wind speed forecast accuracy had significant financial impacts on wind energy projects like the Gemini wind farm. GraphCast reduced imbalance costs compared to HRES, while Aspire Meso increased costs due to higher forecast errors.

6.2. Implications for Weather Forecasting

6.2.1. Validation of AI-Based Forecasting

The findings validate the potential of AI-based models like GraphCast and Pangu-Weather in operational weather forecasting:

- **Accuracy and Efficiency**: GraphCast demonstrated superior accuracy over traditional NWP models while offering computational efficiency during inference. Pangu-Weather also showed competitive performance, though it lacks precipitation forecasts.
- **Operational Benefits**: The reduced computational costs and faster inference times make AI models attractive for operational settings, enabling more frequent updates and real-time applications. However, the substantial training costs and hardware requirements remain challenges.
- **Economic Advantages**: Improved forecast accuracy can lead to significant cost savings in sectors like renewable energy, as evidenced by the reduction in imbalance costs for wind farms using GraphCast.

6.2.2. Challenges and Limitations

Despite the advantages, several challenges were identified:

- **Black-Box Nature**: AI models lack interpretability, making it difficult to diagnose and address specific weaknesses or biases.
- **Dependence on Training Data**: Reliance on data like ERA5 and HRES for training limits the diversity of error patterns and may lead to asymptotic behavior in forecast improvements.
- **Inflexibility in Local Adaptation**: Global AI models require full global input and output, reducing their adaptability to local scales and specific regional phenomena.
- **Hardware Requirements**: Substantial memory footprints necessitate high-end GPUs, potentially limiting accessibility in resource-constrained environments.
- **Variable Limitations**: Models like Pangu-Weather exclude critical variables such as precipitation, limiting their applicability in certain sectors.
- **Model Diversity**: The high inter-model correlation due to shared data sources underscores the need for diversification in model training data and architectures to improve ensemble performance.

6.2.3. Implications of Neural Scaling Laws and Model Architecture

The study explored the implications of neural scaling laws for weather forecasting models:

- **Model Size vs. Performance**: Despite being smaller, GraphCast outperformed the larger Pangu-Weather model, suggesting that architectural efficiency and effective data utilization are more critical than merely increasing model size.
- **Data Bottlenecks**: High inter-model correlations indicate that shared data sources limit performance improvements, emphasizing the need for diversified and higher-quality training datasets.
- **Architectural Efficiency**: GraphCast's use of Graph Neural Networks (GNNs) and multimesh representations may offer more effective modeling of atmospheric processes compared to Pangu-Weather's Transformer-based architecture.
- **Limitations of Scaling**: The findings challenge the applicability of neural scaling laws in weather forecasting, indicating that simply increasing model size and computational resources does not guarantee better performance. Factors like data quality, architectural innovation, and domain-specific model design play significant roles.

6.2.4. Model Diversity and Ensemble Forecasting

The study highlighted the need for increased model diversity:

- **Diversity in Training Data**: Incorporating data from other global models and observational networks could enhance model robustness and reduce correlated errors.
- **Ensemble Approaches**: Utilizing ensembles composed of models with varied architectures and data sources can provide a more comprehensive depiction of forecast uncertainty. However, current ensembles were limited by the high inter-model correlation due to shared data sources.
- **Hybrid Models**: Combining the strengths of AI-based and physical models may offer improved accuracy and applicability across different scales and regions.

6.2.5. Economic Implications
The economic analysis underscores the financial impact of forecast accuracy:

- **Imbalance Costs**: Even small improvements in wind speed forecast accuracy can lead to substantial cost reductions in wind energy operations, highlighting the economic value of accurate forecasting models.
- **Cost-Benefit of Ensembles**: Ensemble forecasting models, by further reducing forecast errors, have the potential to decrease imbalance costs more significantly, reinforcing the economic benefits of adopting ensemble strategies.

6.2.6. Operational Benefits and Challenges
AI models offer operational advantages such as reduced computational costs and faster inference times, making them suitable for real-time forecasting needs. However, challenges related to hardware requirements, model interpretability, and adaptability to local scales must be addressed to fully realize their potential in operational settings.

6.3. Limitations of the Study
Several limitations must be acknowledged:

- **Temporal Resolution**: The use of 6-hour intervals may overlook short-term variability and transient weather phenomena, limiting the assessment of models' capabilities in capturing rapid atmospheric changes.
- **Geographical Scope**: Focusing solely on the Netherlands limits the generalizability of the findings to regions with different climatic and geographical characteristics, such as mountainous or tropical regions.
- **Variable Selection**: Exclusion of variables like solar radiation and cloud cover narrows the evaluation scope and may not reflect the full capabilities of the models.
- **Data Dependence**: The study's reliance on specific datasets may introduce biases related to data quality and coverage, potentially affecting model performance assessments.
- **Model Configurations**: Rigid model configurations prevented exploration of their full potential or adaptability through modifications to physical parameterizations or data assimilation techniques.
- **Ensemble Limitations**: The ensembles employed were equally weighted and limited by high inter-model correlations, suggesting that more sophisticated weighting or diversity in model selection could yield better performance.

6.4. Recommendations for Future Work
To address the identified limitations and enhance the utility of AI-based weather models, the following recommendations are proposed:

6.4.1. Incorporate More Observational Data
Integrating additional observational data can:

- **Reduce Baseline Errors**: Enhance the accuracy of initial conditions and improve model performance by leveraging more diverse and high-resolution data sources.
- **Enhance Local Accuracy**: Fine-tuning globally trained models with higher-resolution local data can improve forecasts for specific regions, addressing localized phenomena more effectively.
- **Diversify Training Sets**: Reduce over-reliance on a single data source, leading to more robust models with varied error characteristics.

6.4.2. Enhance Model Modularity and Adaptability
Enhancing model architectures to allow for:

- **Regional Fine-Tuning**: Adapt models to specific regions without retraining the entire global model, improving local forecast accuracy.

- **Application-Specific Configurations**: Tailor models to meet the needs of different sectors or applications, such as renewable energy or agriculture.
- **Hybrid Integration**: Combine AI models with physical models to leverage the strengths of both approaches, enhancing overall forecast accuracy and reliability.

6.4.3. Improve Temporal and Spatial Resolution

Improving resolution can:

- **Capture Short-Term Variability**: Better represent rapid atmospheric changes essential for short-range forecasting and real-time applications.
- **Increase Applicability**: Meet the requirements of applications needing high temporal fidelity, such as severe weather warnings and energy grid management.
- **Resolve Local Phenomena**: Improve the representation of microclimates and complex terrains, enhancing local forecast accuracy.

6.4.4. Expand Variable Range

Including additional meteorological variables can:

- **Broaden Applicability**: Address the needs of sectors like solar energy forecasting, climate studies, and hydrological applications.
- **Enhance Model Completeness**: Provide a more comprehensive depiction of atmospheric conditions, enabling multi-faceted applications.
- **Improve Inter-variable Relationships**: Capture interactions between different atmospheric parameters, leading to more accurate and coherent forecasts.

6.4.5. Emphasize Validation Against Observations

Focusing on actual observations rather than reanalysis datasets can:

- **Improve Reliability**: Provide a more accurate assessment of model performance in real-world conditions, ensuring that models are truly capturing atmospheric dynamics.
- **Identify Model Biases**: Highlight discrepancies between model outputs and actual atmospheric states, guiding targeted improvements.
- **Guide Model Improvements**: Inform adjustments to address specific weaknesses, enhancing overall forecast accuracy and reliability.

6.4.6. Explore Advanced Ensemble Forecasting Techniques

Developing ensemble techniques can:

- **Enhance Reliability**: Provide probabilistic forecasts that account for uncertainties, improving decision-making in risk-sensitive applications.
- **Improve Decision-Making**: Offer a range of possible outcomes, aiding in risk assessment and strategic planning across various sectors.
- **Leverage Computational Efficiency**: Utilize AI models' rapid inference to generate ensembles without prohibitive computational costs, making ensemble forecasting more accessible and scalable.

6.5. Final Conclusion

This thesis has demonstrated the significant potential of AI-based models like GraphCast and Pangu-Weather to advance weather forecasting, offering improved accuracy and computational efficiency over traditional NWP models. The superior performance of GraphCast, despite its smaller size compared to Pangu-Weather, underscores the importance of architectural efficiency and effective data utilization over mere scaling. The study highlights that simply increasing model size and computational resources does not guarantee better performance, emphasizing the critical roles of data quality, model architecture, and domain-specific design.

The high inter-model correlation due to shared data sources points to a data bottleneck, suggesting that future advancements should focus on diversifying training datasets and incorporating more observational data to break through performance plateaus. Enhancing model modularity and adaptability can address challenges related to local adaptation and operational flexibility, making AI models more practical for a variety of applications.

The economic analysis reinforces the tangible benefits of accurate weather forecasting, particularly in sectors like renewable energy where forecast errors can lead to substantial financial losses. By adopting more accurate models and exploring advanced ensemble techniques, organizations can improve operational efficiency and reduce costs.

In conclusion, the integration of AI in weather forecasting presents a promising avenue for innovation. Balancing the strengths of AI models with the depth of physical understanding inherent in traditional NWP models will be key to advancing the field. Addressing the identified limitations through the recommended strategies will enable the development of more robust, flexible, and comprehensive forecasting systems, ultimately meeting the evolving needs of society in the face of changing climatic conditions.

References

Battaglia, P., Pascanu, R., Lai, M., Rezende, D., & Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, 4509–4517.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks, 1–40. http://arxiv.org/abs/1806.01261

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3
Pangu Weather model.

Bjerknes, V. (1904). *Weather forecasting as a problem in mechanics and physics*. The Carnegie Institution of Washington.

Bremnes, J. B., Nipen, T. N., & Seierstad, I. A. (2023). Evaluation of forecasts by a global data-driven weather model with and without probabilistic post-processing at Norwegian stations. (1), 1–9. http://arxiv.org/abs/2309.01247

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. http://arxiv.org/abs/2104.13478
book recommended by graphcast author.

Charney, J. G. (1948). On the scale of atmospheric motions. *Geofysiske Publikasjoner*, *17*, 3–17.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, *6*(1), 1–27. https://doi.org/10.1038/s41612-023-00512-1
Ensemble forecasting with 3 different lead times.

Coiffier, J. T. A. .-. T. T. .-. (2011). Fundamentals of numerical weather prediction (NV - 1 online resource). https://doi.org/LK-https://tudelft.on.worldcat.org/oclc/774696886

De Burgh-Day, C. O., & Leeuwenburg, T. (2023). Machine learning for numerical weather and climate modelling: A review. *Geoscientific Model Development*, *16*(22), 6433–6477. https://doi.org/10.5194/gmd-16-6433-2023

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204. https://doi.org/10.1016/S0377-2217(98)00380-4

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

ECMWF. (n.d.). *Medium-range forecasts* [Accessed: 2024-09-24]. https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts

ECMWF. (2013). Part II : Data Assimilation IFS DOCUMENTATION – Cy40r1 Operational implementation 22. (November), 1–102.

ECMWF. (2020a). IFS DOCUMENTATION – Cy47r1 Operational implementation 30 June 2020 PART VI : TECHNICAL AND. (June), 1–241.

ECMWF. (2020b). Part I : Observations IFS DOCUMENTATION – Cy47r1 Operational implementation 30 June 2020 PART I : OBSERVATIONS. (June), 1–82.

ECMWF. (2020c). Part Iii: Dynamics and Numerical Procedures Revision History. (June), 1–40.

ECMWF. (2020d). Part IV : Physical Processes IFS DOCUMENTATION – Cy47r1 Operational implementation 30 June 2020 PART IV : PHYSICAL PROCESSES. (June). https://www.ecmwf.int/en/elibrary/20198-ifs-documentation-cy47r3-part-iv-physical-processes

ECMWF. (2020e). Part VII : ECMWF Wave Model IFS DOCUMENTATION – Cy47r1 Operational implementation 30 June 2020 PART VII : ECMWF WAVE MODEL. (June), 1–114.

ECMWF. (2021). *Ecmwf opens new data center to house atos supercomputer in italy* [Accessed: 2024-09-24]. https://www.datacenterdynamics.com/en/news/ecmwf-opens-new-data-center-to-house-atos-supercomputer-in-italy/

ECMWF. (2023a). *Ecmwf high-performance computing facility boosts forecasts and research* [Accessed: 2024-09-24]. https://www.ecmwf.int/en/about/what-we-do/computing/our-facilities

ECMWF. (2023b). IFS Documentation CY48R1 - Part VII: ECMWF Wave Model. *IFS Documentation CY48R1*, (7), 1–84.

ECMWF. (2024). Copernicus Climate Change Service provides new tools for users — ecmwf.int [[Accessed 16-10-2024]].

Elliott, G. (2011). Averaging and The Optimal Combination of Forecasts. *University of California, San Diego.*, (858), 1–30. http://weber.ucsd.edu/$%5Csim$gelliott/AveragingOptimal.pdf

Fabbri, A., Roman, T., Abbad, J., & Quezada, V. (2005). Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market. *IEEE Transactions on Power Systems*, *20*(3), 1440–1446. https://doi.org/10.1109/TPWRS.2005.852148

GetDeploying. (2024, November). Nvidia a100 - price comparison [Accessed: 2024-10-09].

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248–249. https://doi.org/10.1126/science.1115255

Google Cloud Platform. (2024). Google cloud pricing [[Accessed 16-10-2024]].

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., & Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*. https://doi.org/10.1002/qj.3803

International Renewable Energy Agency. (2020). *Advanced forecasting of variable renewable power generation*. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2020/Jul/IRENA_Advanced_weather_forecasting_2020.pdf?la=en&hash=8384431B56569C0D8786C9A4FDD56864443D10AF

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. http://arxiv.org/abs/2001.08361

Keisler, R. (2022). Forecasting Global Weather with Graph Neural Networks, 1–16. http://arxiv.org/abs/2202.07575
graphcast.

Kolmogorov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Doklady Akademii Nauk SSSR*, *30*, 301–304.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023a). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1422. https://doi.org/10.1126/science.adi2336
Google Deepmind original Graphcast paper.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023b). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1422. https://doi.org/10.1126/science.adi2336
Graphcast supplementary.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., & Rabier, F. (2024). AIFS - ECMWF's data-driven forecasting system. http://arxiv.org/abs/2406.01465

Liu, C. C., Hsu, K., Peng, M. S., Chen, D. S., Chang, P. L., Hsiao, L. F., Fong, C. T., Hong, J. S., Cheng, C. P., Lu, K. C., Chen, C. R., & Kuo, H. C. (2024). Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific. *npj Climate and Atmospheric Science*, *7*(1), 1–12. https://doi.org/10.1038/s41612-024-00769-0

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Kashinath, K., Kautz, J., & Pritchard, M. (2023). Generative Residual Diffusion Modeling for Km-scale Atmospheric Downscaling. http://arxiv.org/abs/2309.15214
CorDiff paper by nvidia.

McNally, A., Lessig, C., Lean, P., Boucher, E., Alexe, M., Pinnington, E., Chantry, M., Lang, S., Burrows, C., Chrust, M., Pinault, F., Villeneuve, E., Bormann, N., & Healy, S. (2024). Data driven weather forecasts trained and initialised directly from observations. http://arxiv.org/abs/2407.15586

Nielsen, H. A., Nielsen, T. S., Madsen, H., San Isidro Pindado, M. J., & Marti, I. (2007). Optimal combination of wind power forecasts. *Wind Energy*, *10*(5), 471–482. https://doi.org/10.1002/we.237

Olivetti, L., & Messori, G. (2024). Do data-driven models beat numerical models in forecasting weather extremes? a comparison of ifs hres, pangu-weather and graphcast. *EGUsphere*, *2024*, 1–35. https://doi.org/10.5194/egusphere-2024-1042

Owens, B. (2022, August). *Section 2.5 model data assimilation, 4d-var* [Forecast User Guide, Last updated September 4, 2024]. European Centre for Medium-Range Weather Forecasts (ECMWF). https://confluence.ecmwf.int/display/FUG/Section+2.5+Model+Data+Assimilation%2C+4D-Var

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., & Engelke, S. (2024). Validating deep-learning weather forecast models on recent high-impact extreme events. https://arxiv.org/abs/2404.17652

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., & Anandkumar, A. (2022). FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. http://arxiv.org/abs/2202.11214
NVIDIA FourCastNet paper.

PBL. (2022). Definitieve Correctiebedragen 2021. https://www.pbl.nl/sites/default/files/downloads/2022-pbl-definitieve-correctiebedragen-2021-4809.pdf

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2023). GenCast: Diffusion-based ensemble forecasting for medium-range weather. (1050), 1–46. http://arxiv.org/abs/2312.15796
Google Deepmind GenCast paper: update to Graphcast.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., & Sha, F. (2024). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. https://arxiv.org/abs/2308.15560

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench [e2020MS002405 2020MS002405]. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405. https://doi.org/https://doi.org/10.1029/2020MS002405

Richardson, L. F. (1922). *Weather prediction by numerical process*. Cambridge University Press.

Rozema, W., Bae, H. J., Moin, P., & Verstappen, R. (2015). Minimum-dissipation models for large-eddy simulation. *Physics of Fluids*, *27*(8), 085107. https://doi.org/10.1063/1.4928700

Schalkwijk, J., Jonker, H. J. J., Siebesma, A. P., & Meijgaard, E. V. (2015). *Weather forecasting using gpu-based large-eddy simulations* (tech. rep.). Whiffle B.V. https://whiffle.nl/wp-content/uploads/Weather-Forecasting-Using-GPU-Based-Large-Eddy-Simulations-1.pdf

Siebesma, A. P., Wyszogrodzki, A. A., et al. (2010). The dales model: A large-eddy simulation approach to study atmospheric boundary layer dynamics. *Geoscientific Model Development*, *3*(2), 415–444. https://doi.org/10.5194/gmd-3-415-2010

Sørensen, J. N. (2022). 2.07 - aerodynamic analysis of wind turbines. In T. M. Letcher (Ed.), *Comprehensive renewable energy (second edition)* (Second Edition, pp. 172–193). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-819727-1.00127-8

Stull, R. B. (1988). *An introduction to boundary layer meteorology*. Kluwer Academic. https://doi.org/LK-https://tudelft.on.worldcat.org/oclc/898834789

Sweeney, C., Bessa, R. J., Browell, J., & Pinson, P. (2020). The future of forecasting for renewable energy. *WIREs Energy and Environment*, *9*(2), e365. https://doi.org/https://doi.org/10.1002/wene.365

Team, L., & Meta, A. @. (2024). The Llama 3 Herd of Models. *arXiv preprint*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere [e2020MS002109 10.1029/2020MS002109].

Journal of Advances in Modeling Earth Systems, *12*(9), e2020MS002109. https://doi.org/https://doi.org/10.1029/2020MS002109

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models [e2021MS002502 2021MS002502]. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002502. https://doi.org/https://doi.org/10.1029/2021MS002502

Whiffle. (2024, July). *Aspire documentation* [Confidential Matter]. Version 13.0.0. Whiffle Holding BV. https://www.whiffle.nl/

Wu, L., Cui, P., Pei, J., & Zhao, L. (2020). *Graph neural networks*.

A

Supplementary Figures and Maps

A.1. Introduction

This appendix contains supplementary figures, maps, and visualizations that provide additional insights into the forecast performance analysis presented in Chapter 4. The sections are structured similarly to the main chapter for easy reference.

A.2. Supplementary Lead Time Analysis

A.2.1. Lead Time Plots by Metric

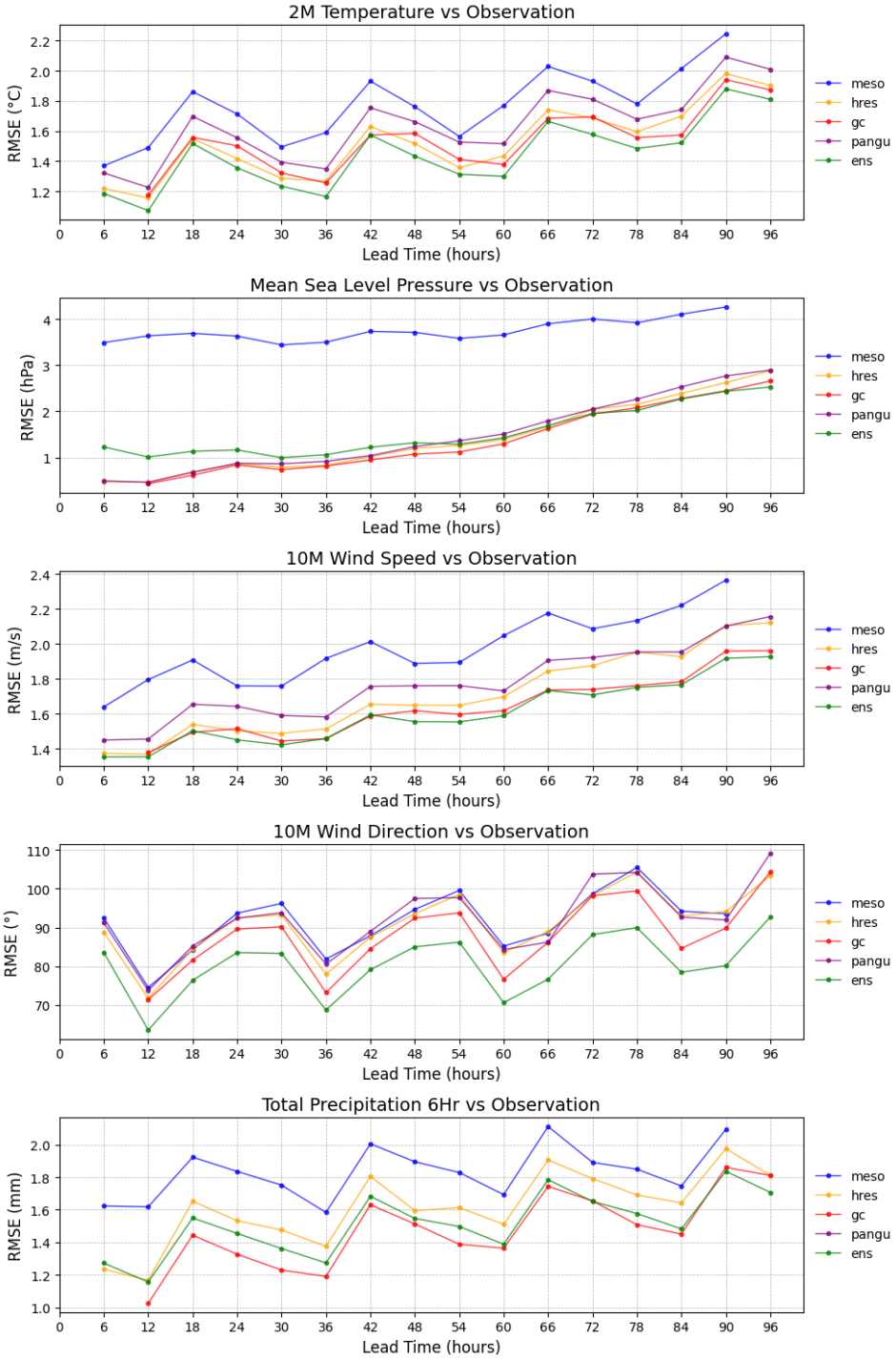


Figure A.1: Lead time RMSE as a function of lead time for various weather variables and models.

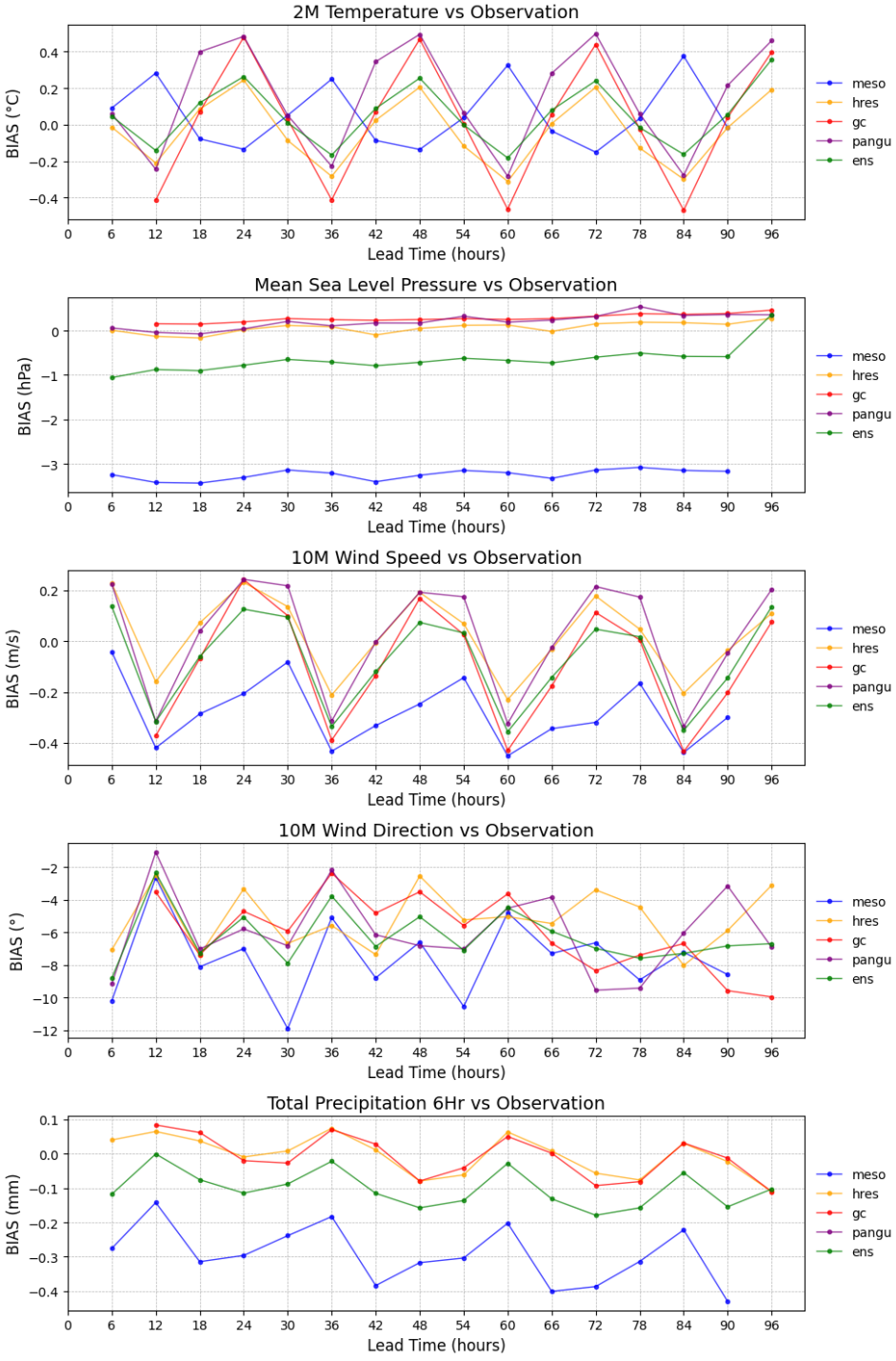


Figure A.2: Lead time bias as a function of lead time for various weather variables and models.

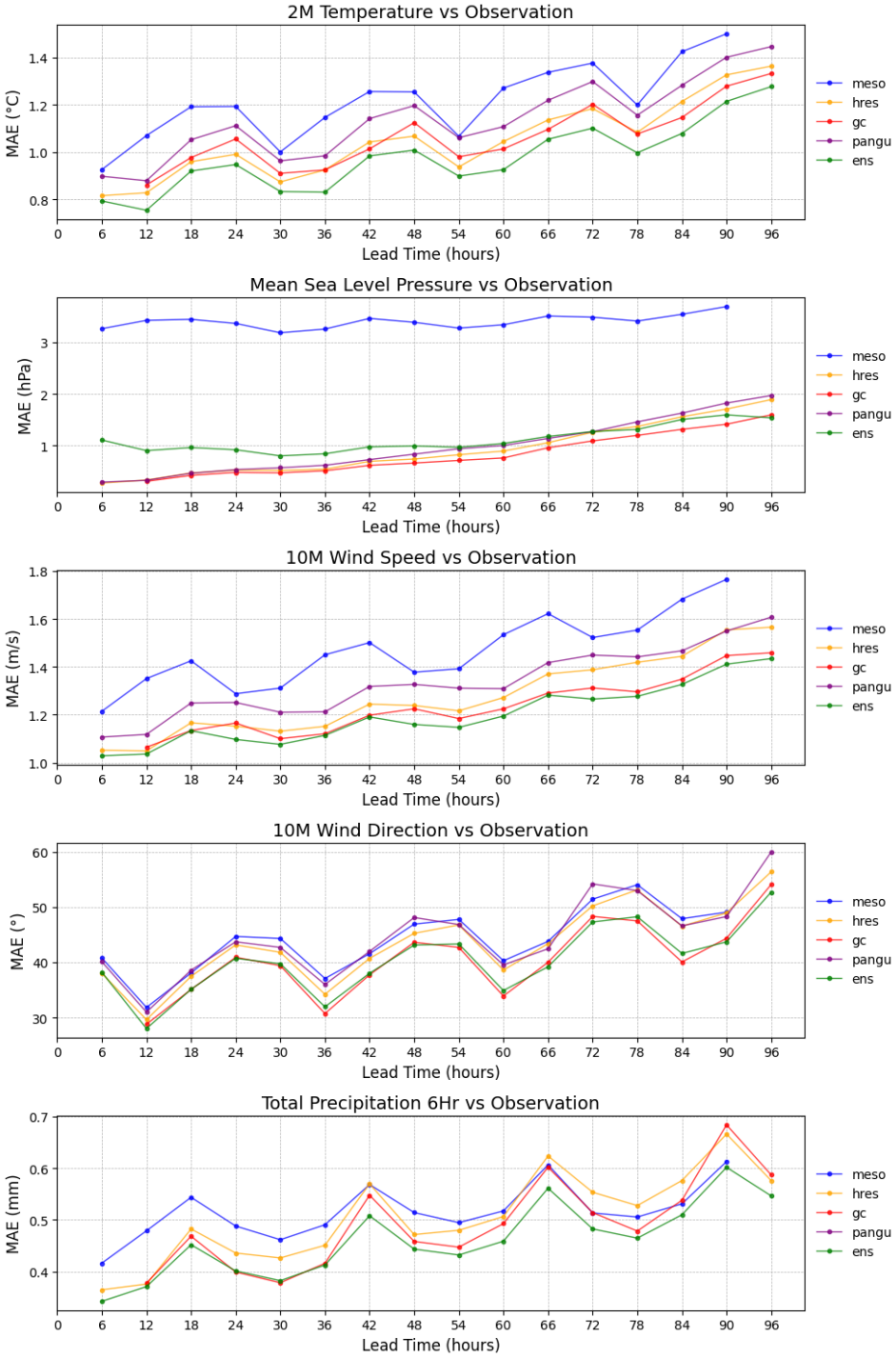


Figure A.3: Lead time MAE as a function of lead time for various weather variables and models.

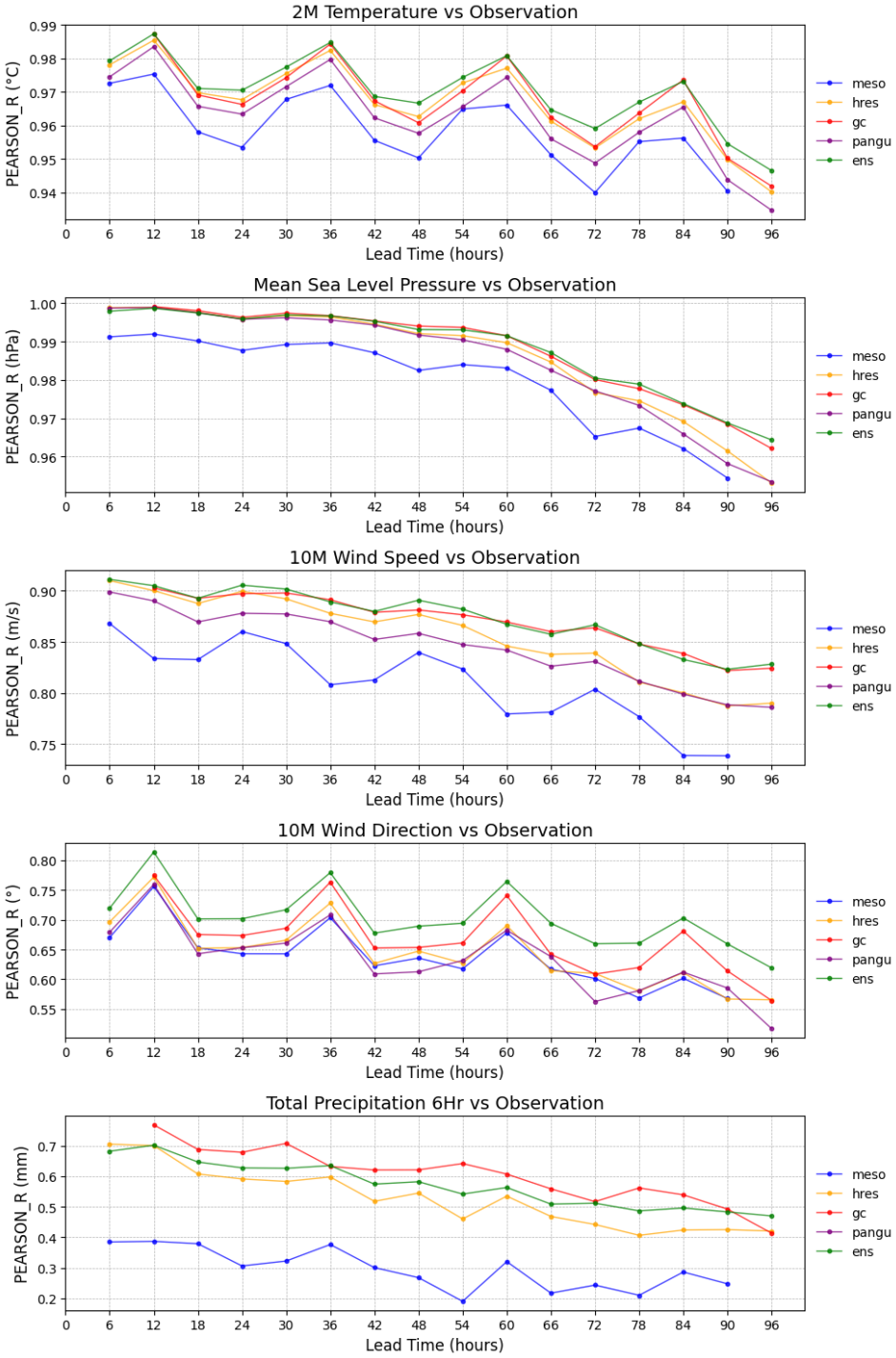


Figure A.4: Lead time Pearson correlation coefficient as a function of lead time for various weather variables and models.

A.3. Supplementary Spatial Analysis

A.3.1. Geographic Map Plots

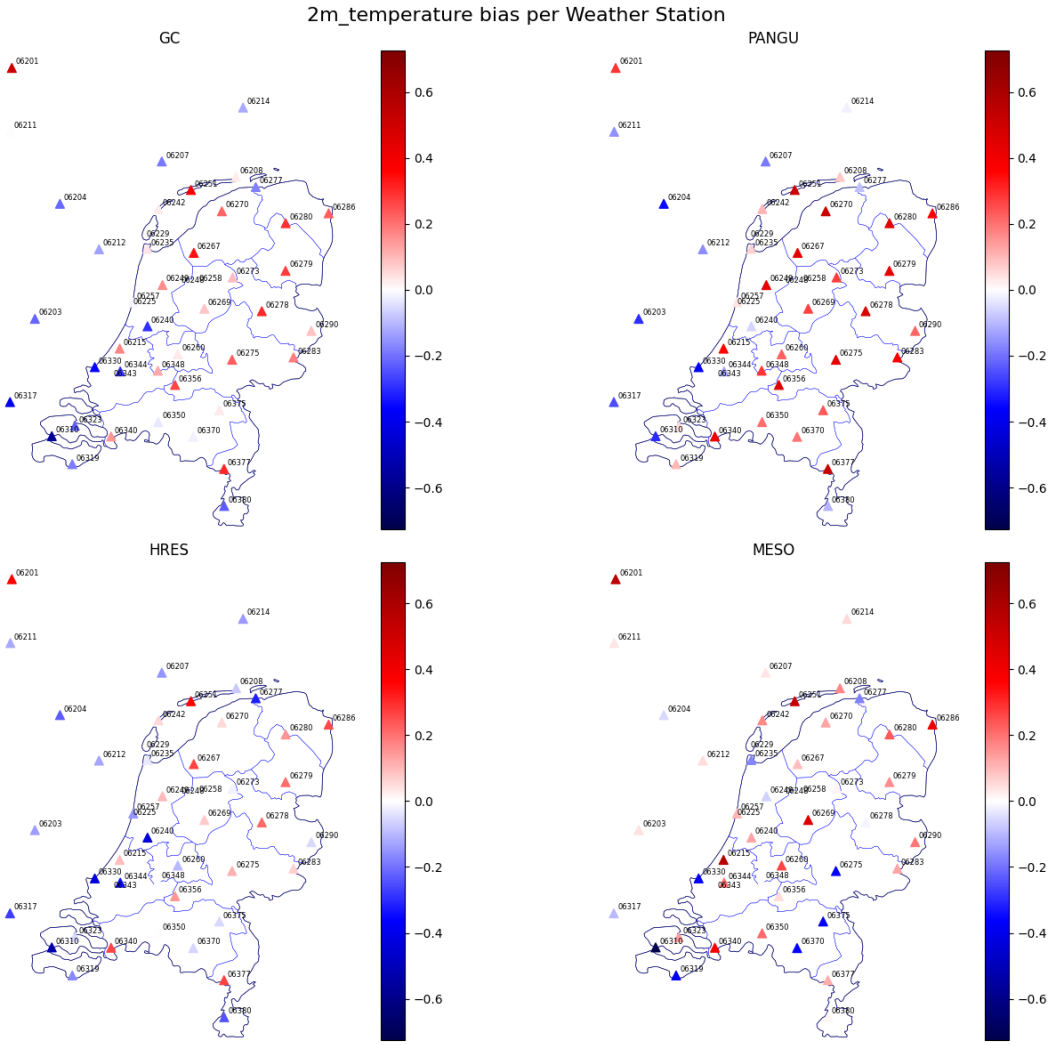


Figure A.5: 2m temperature bias per weather station for various models.

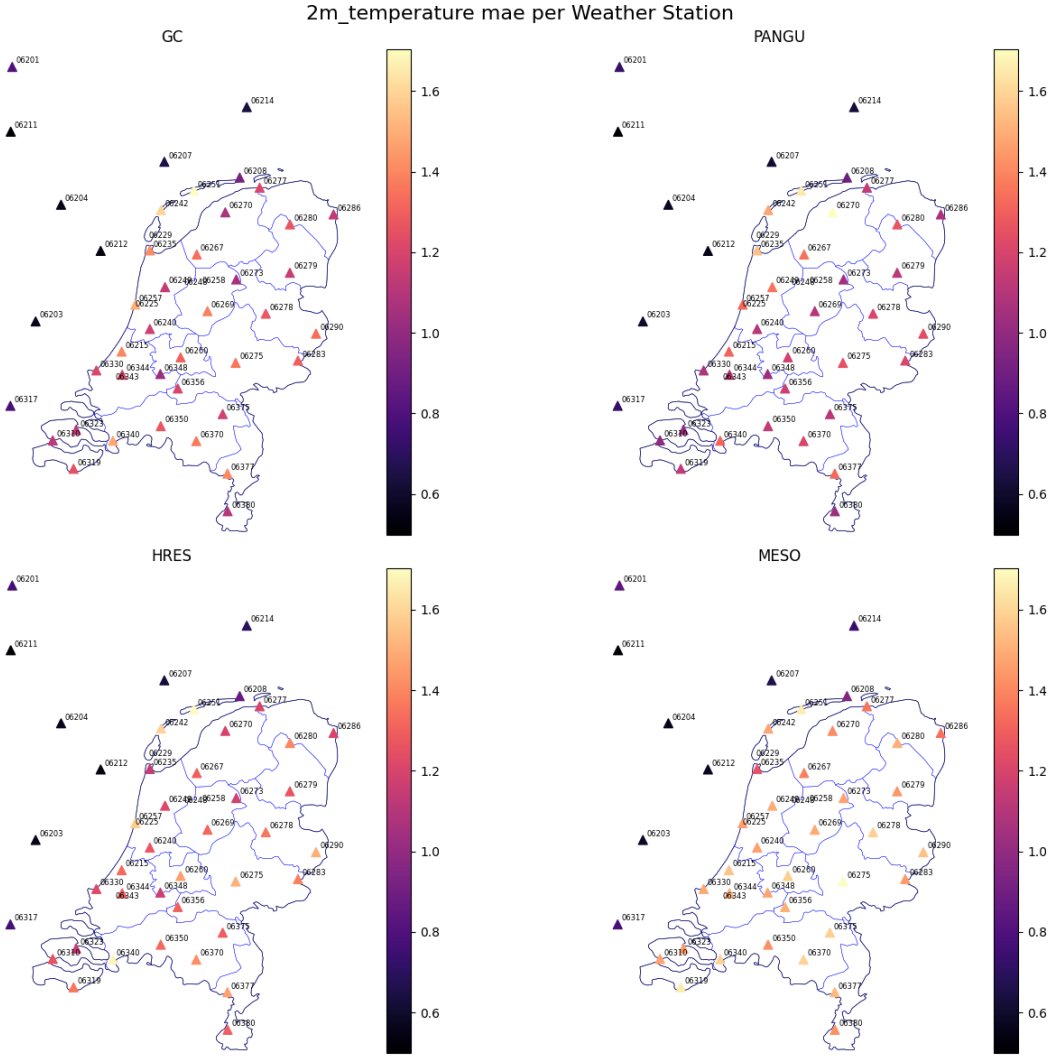


Figure A.6: 2m temperature MAE per weather station for various models.

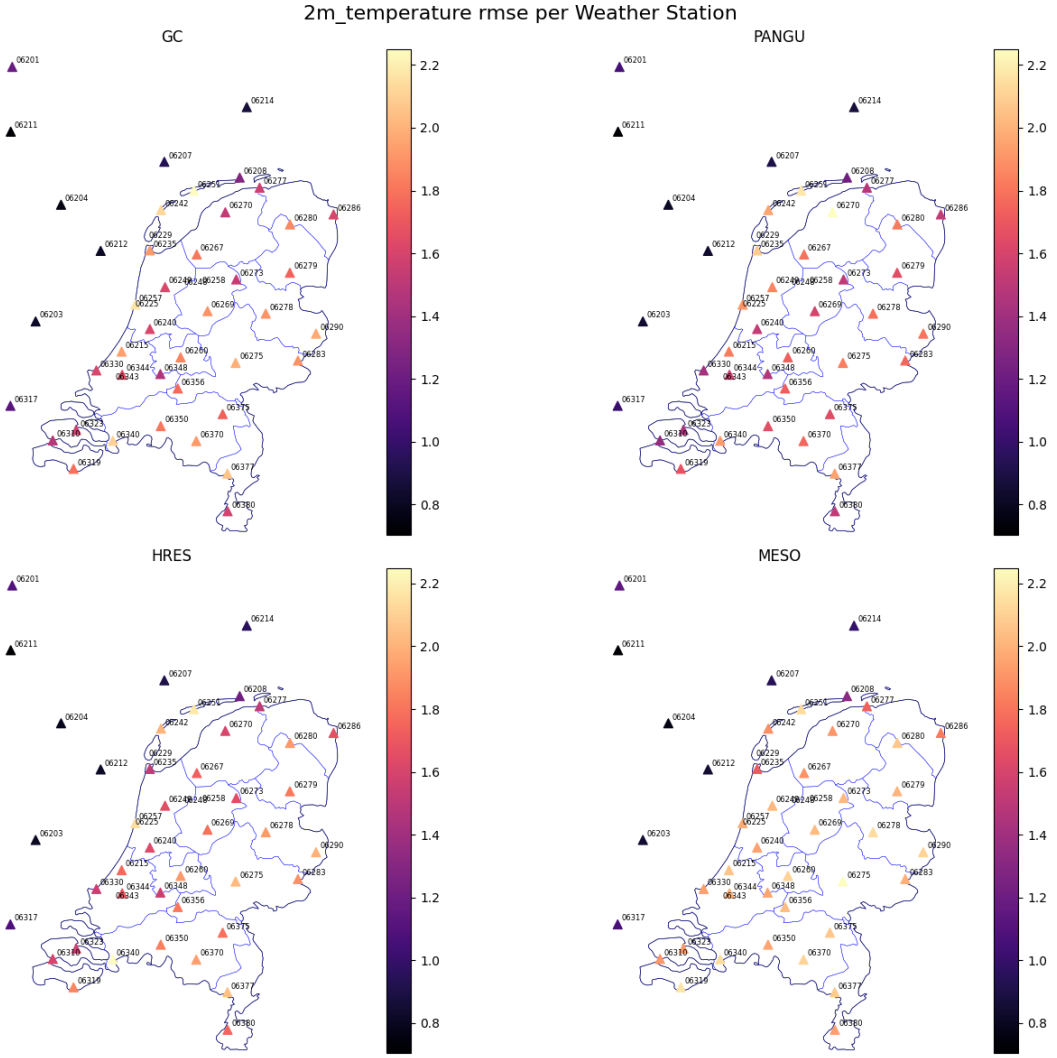


Figure A.7: 2m temperature RMSE per weather station for various models.

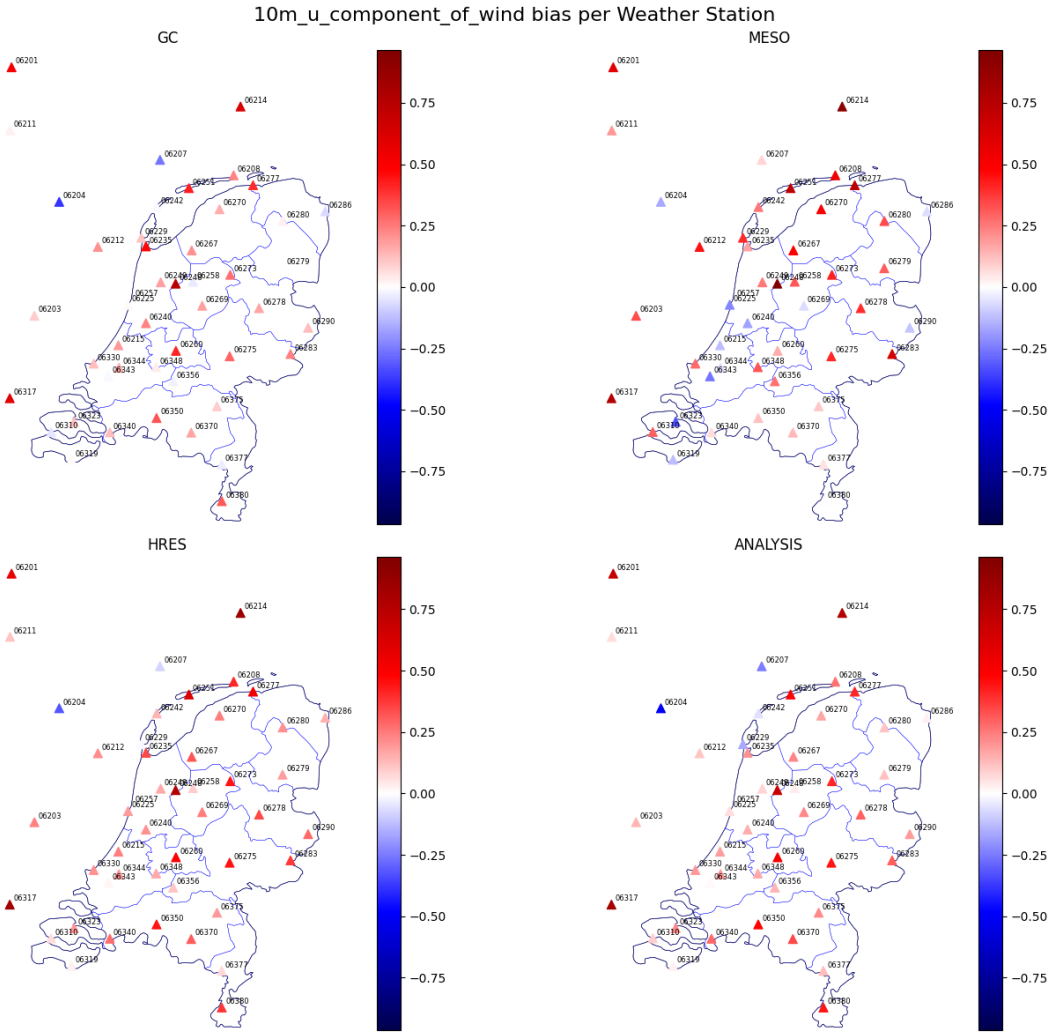


Figure A.8: 10m u-component of wind bias per weather station for various models.

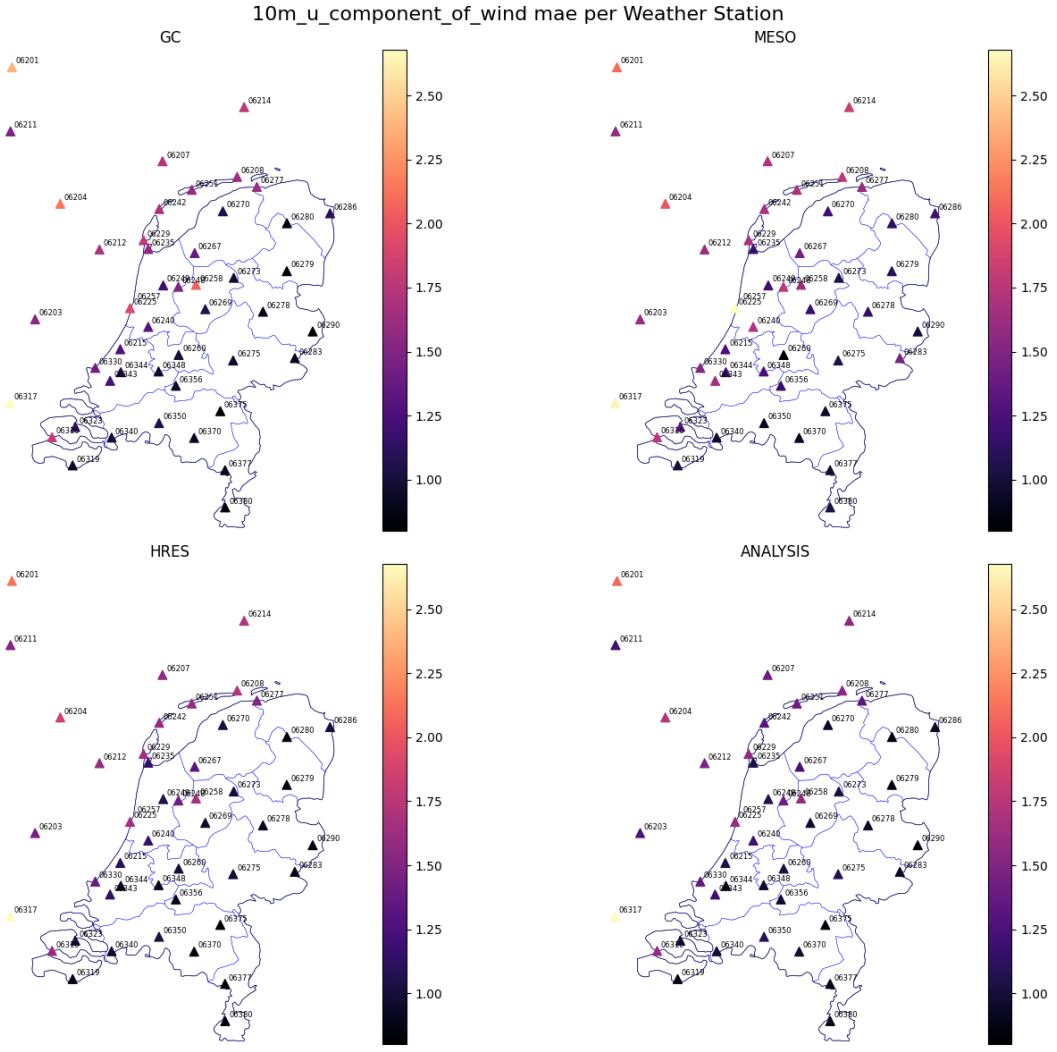


Figure A.9: 10m u-component of wind MAE per weather station for various models.

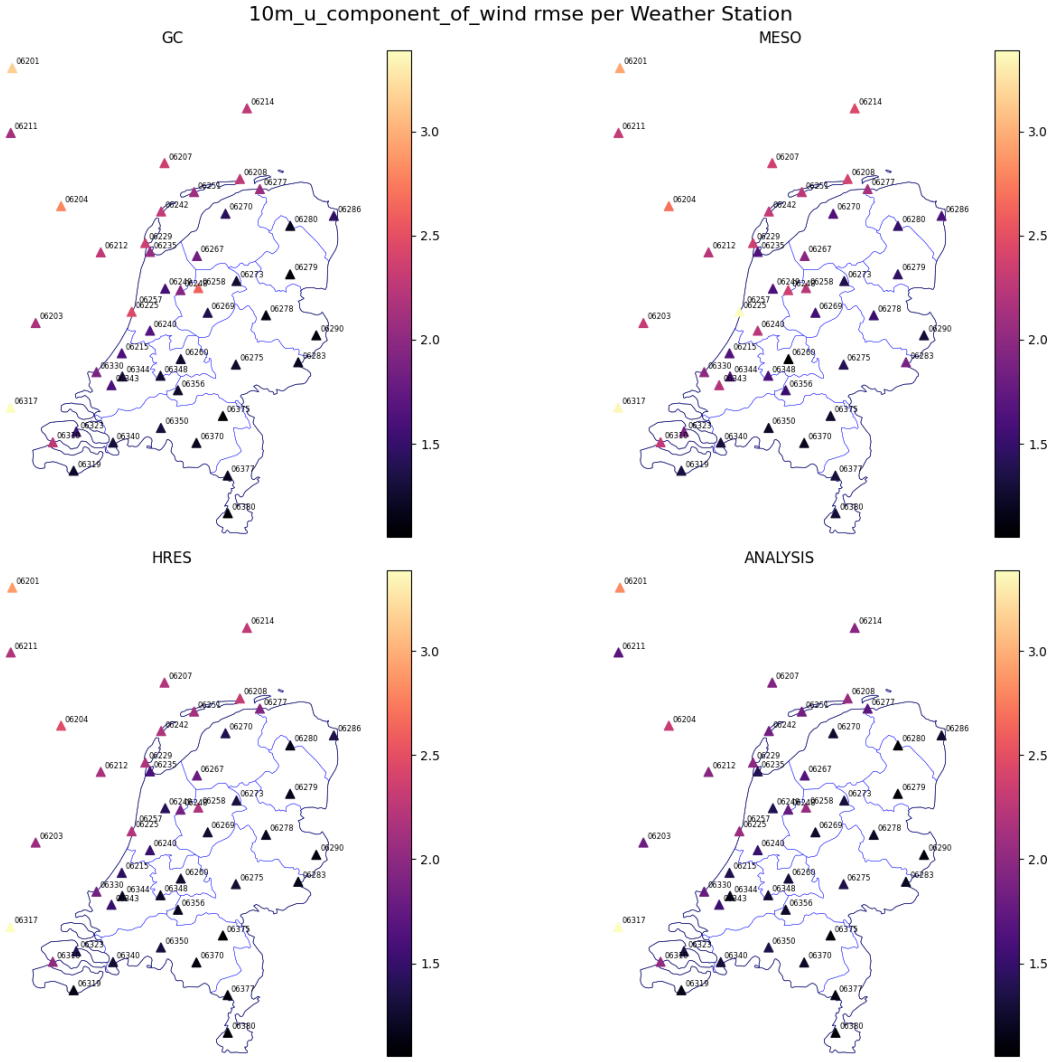


Figure A.10: 10m u-component of wind RMSE per weather station for various models.

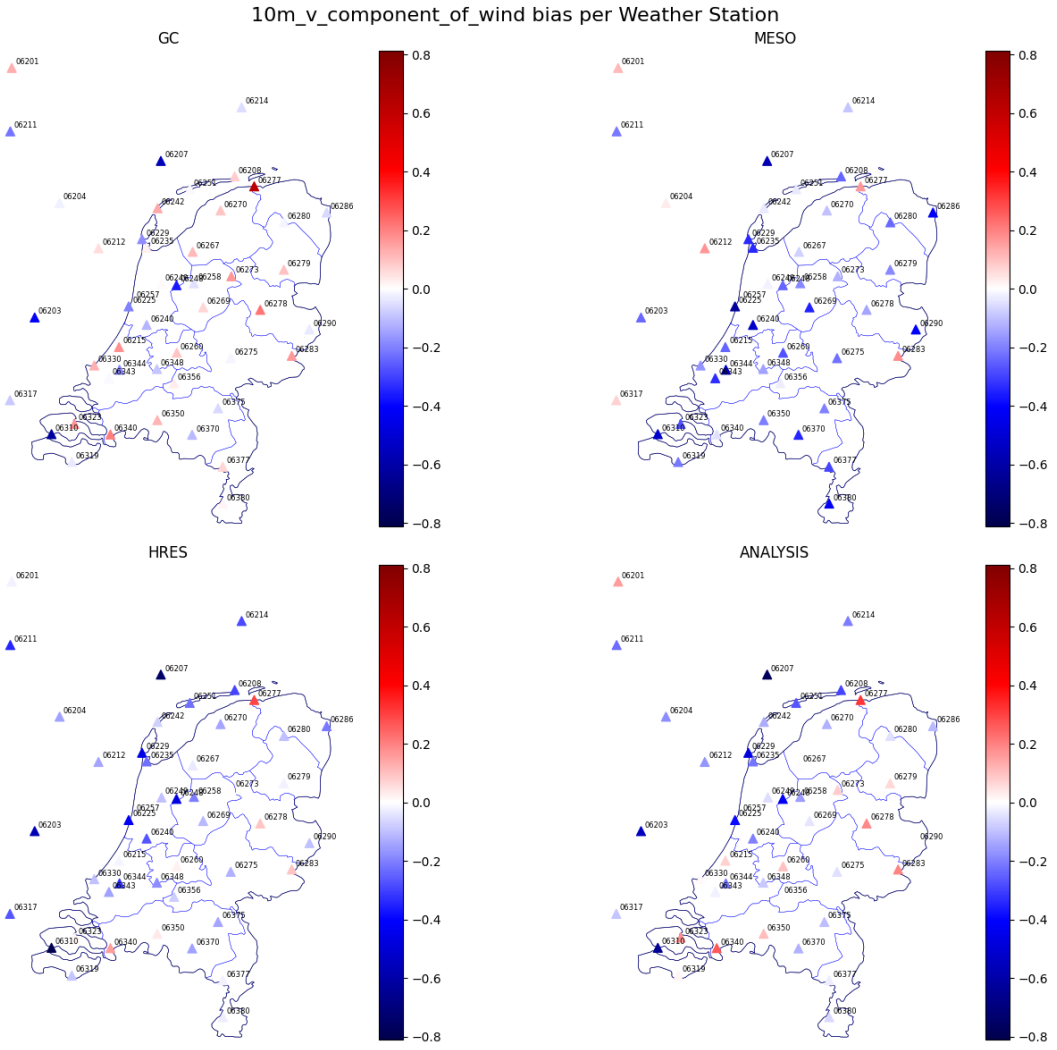


Figure A.11: 10m v-component of wind bias per weather station for various models.

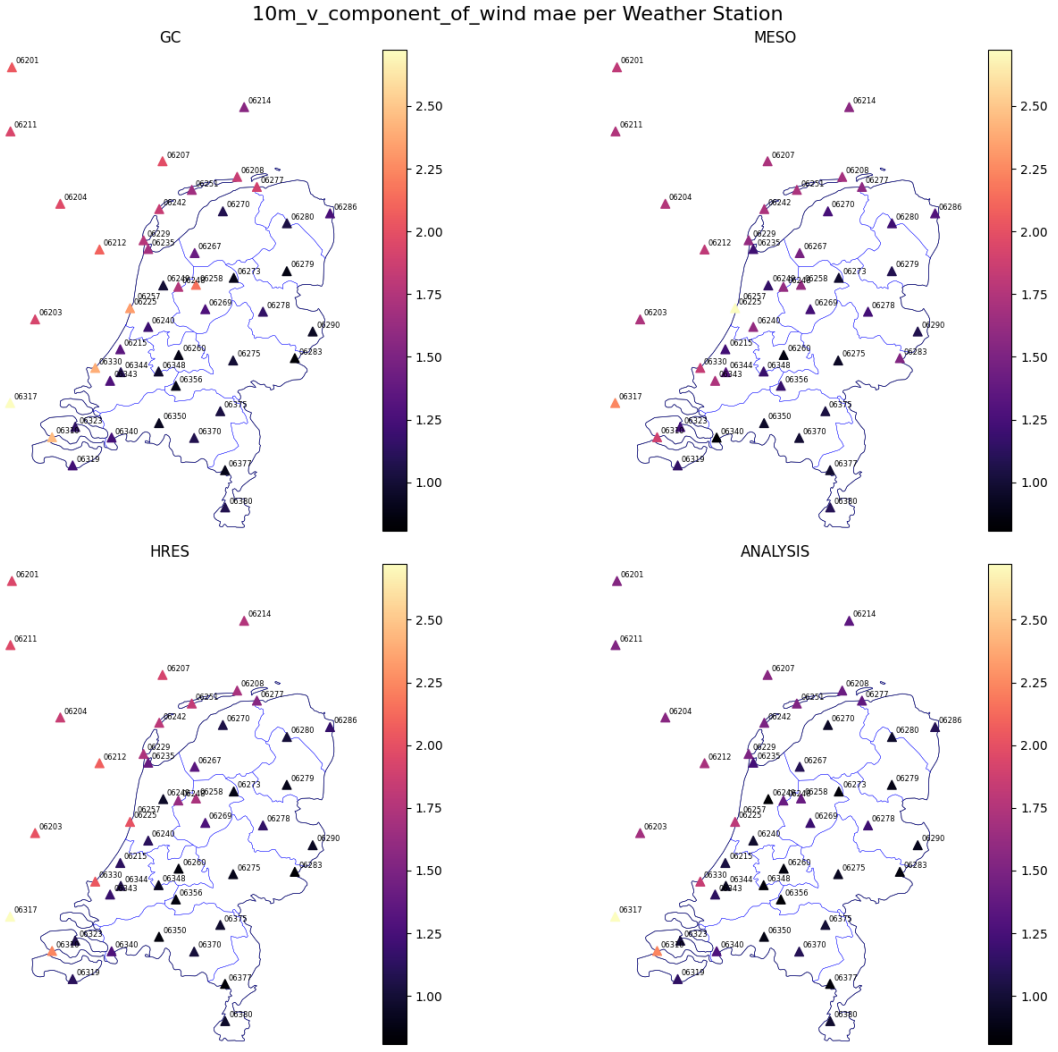


Figure A.12: 10m v-component of wind MAE per weather station for various models.

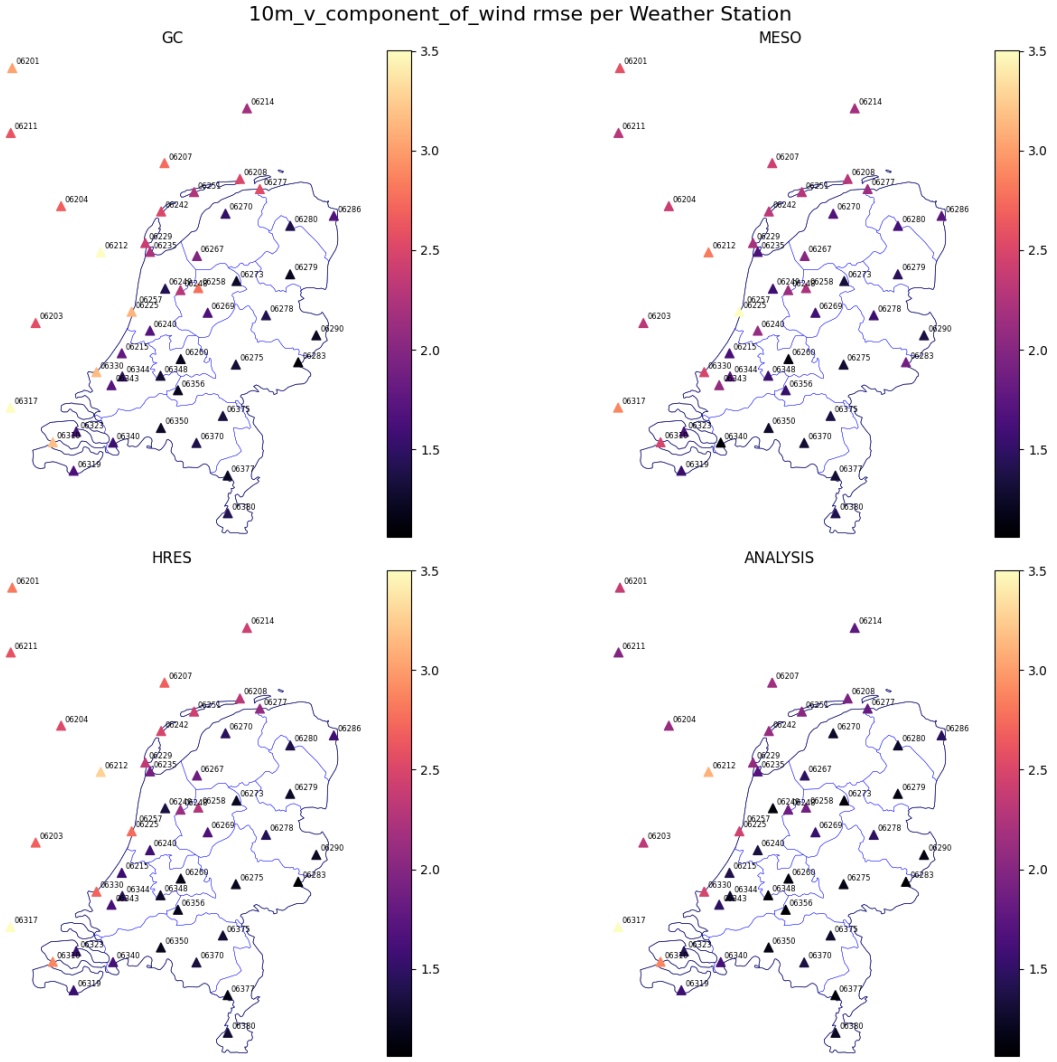


Figure A.13: 10m v-component of wind RMSE per weather station for various models.

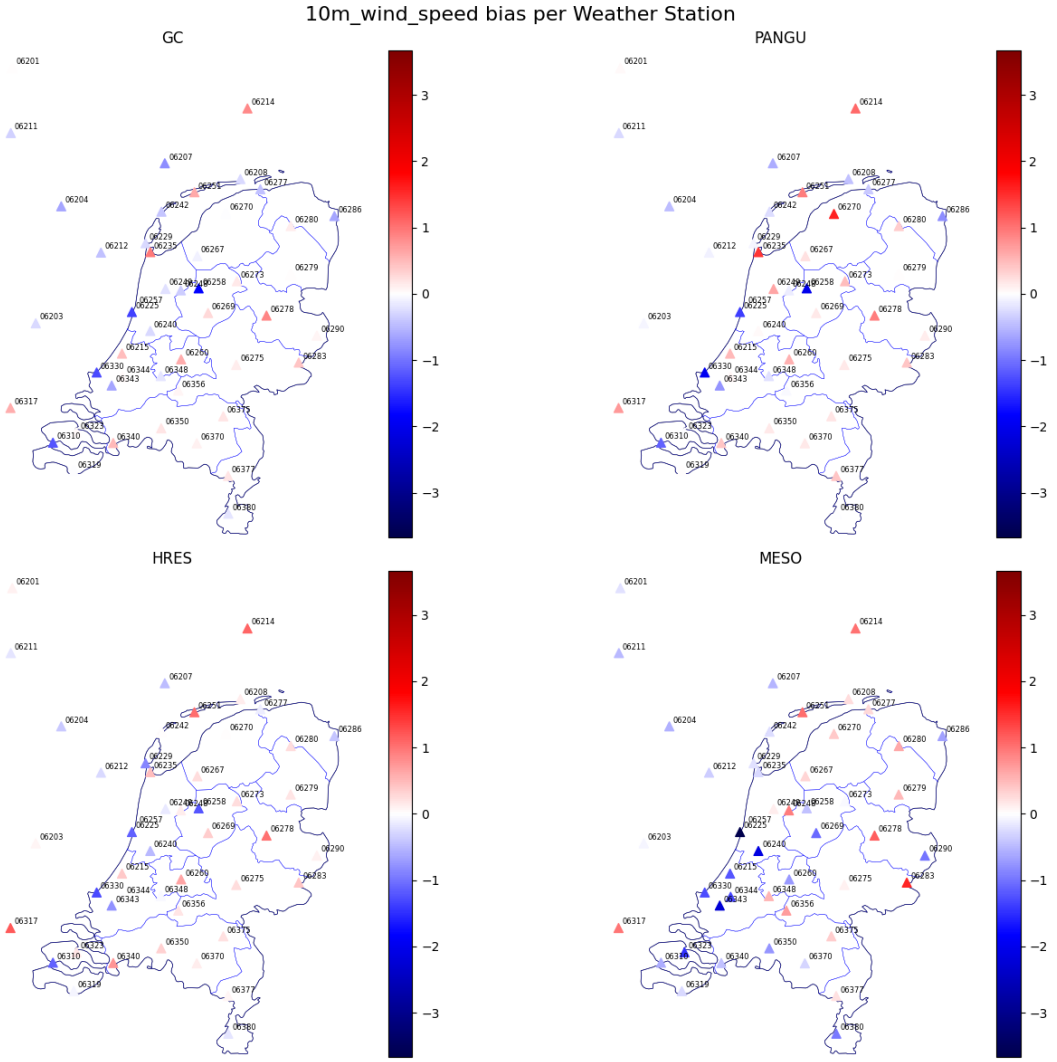


Figure A.14: 10m wind speed bias per weather station for various models.

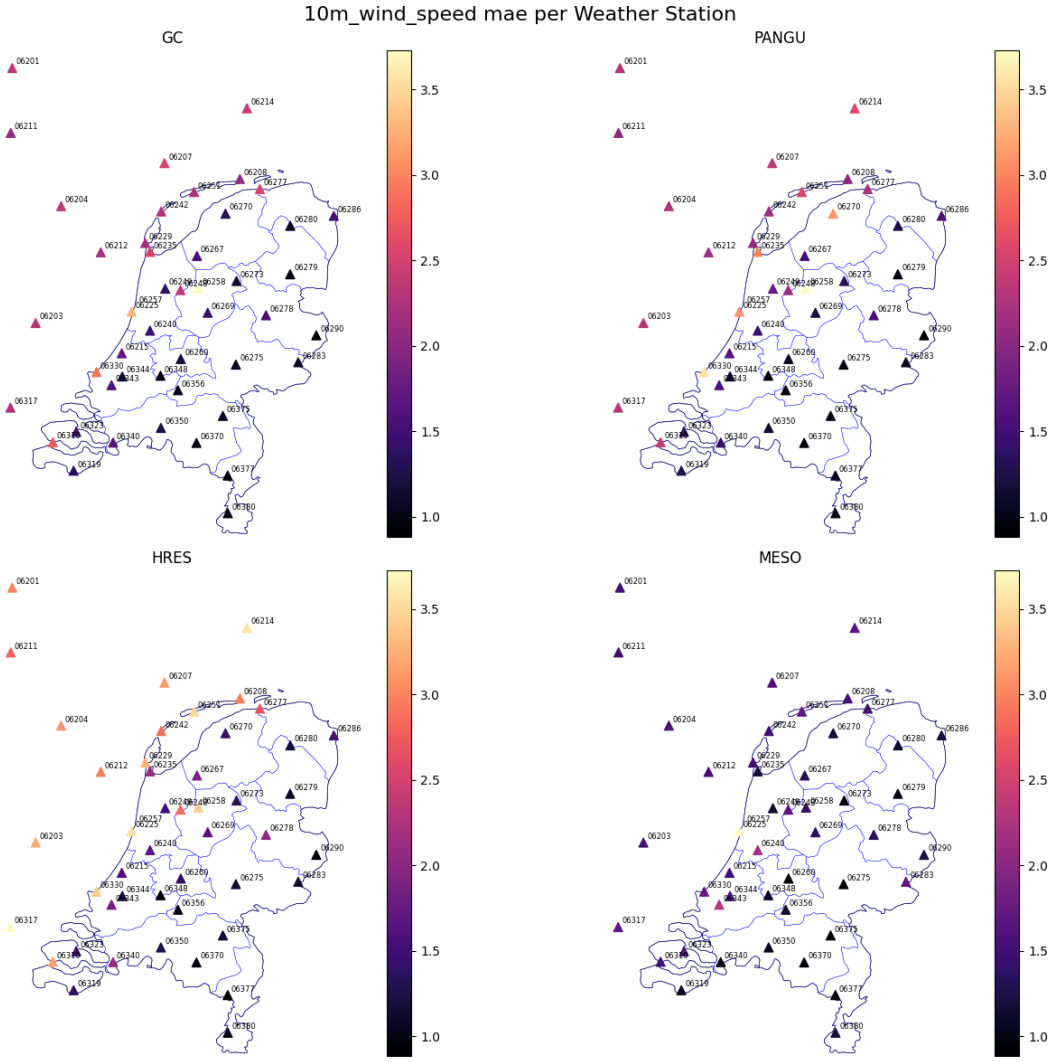


Figure A.15: 10m wind speed MAE per weather station for various models.

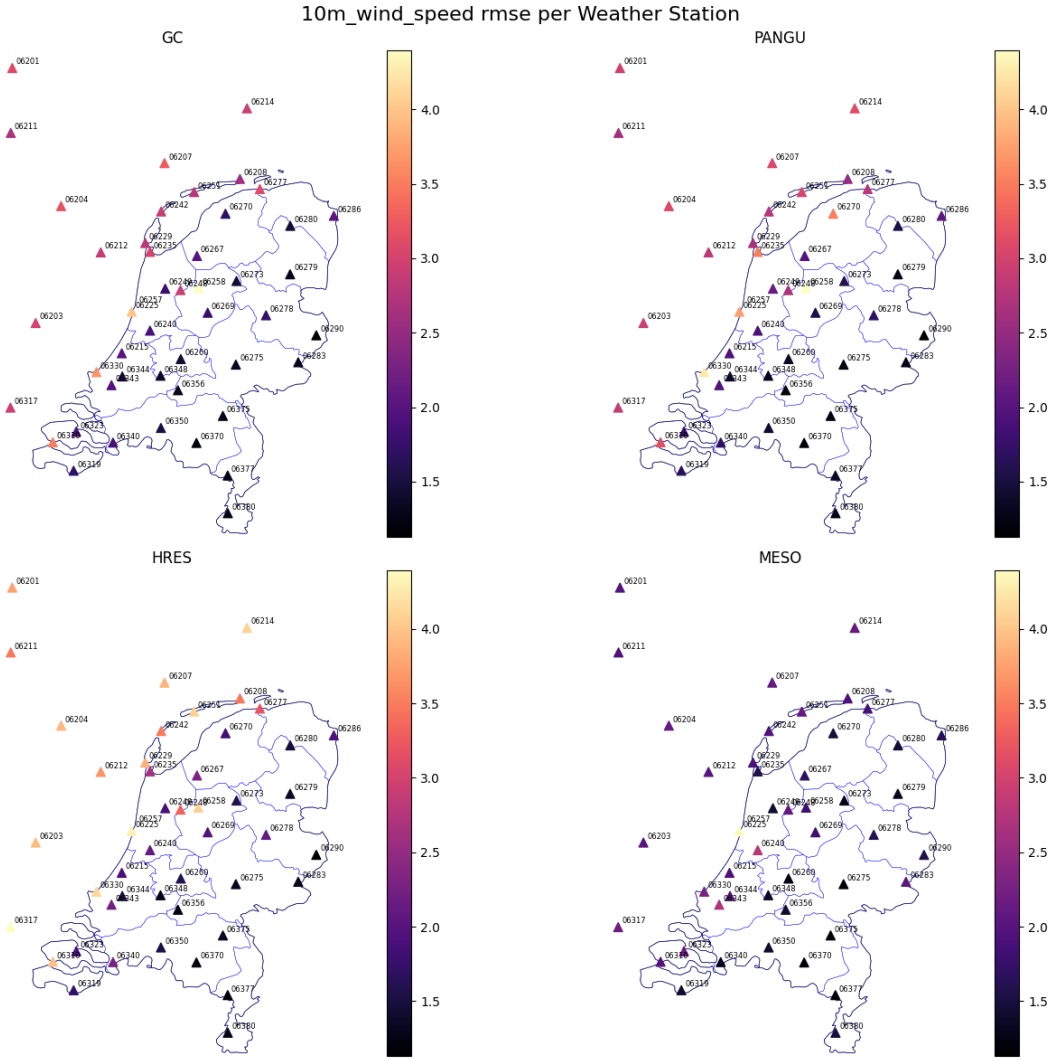


Figure A.16: 10m wind speed RMSE per weather station for various models.

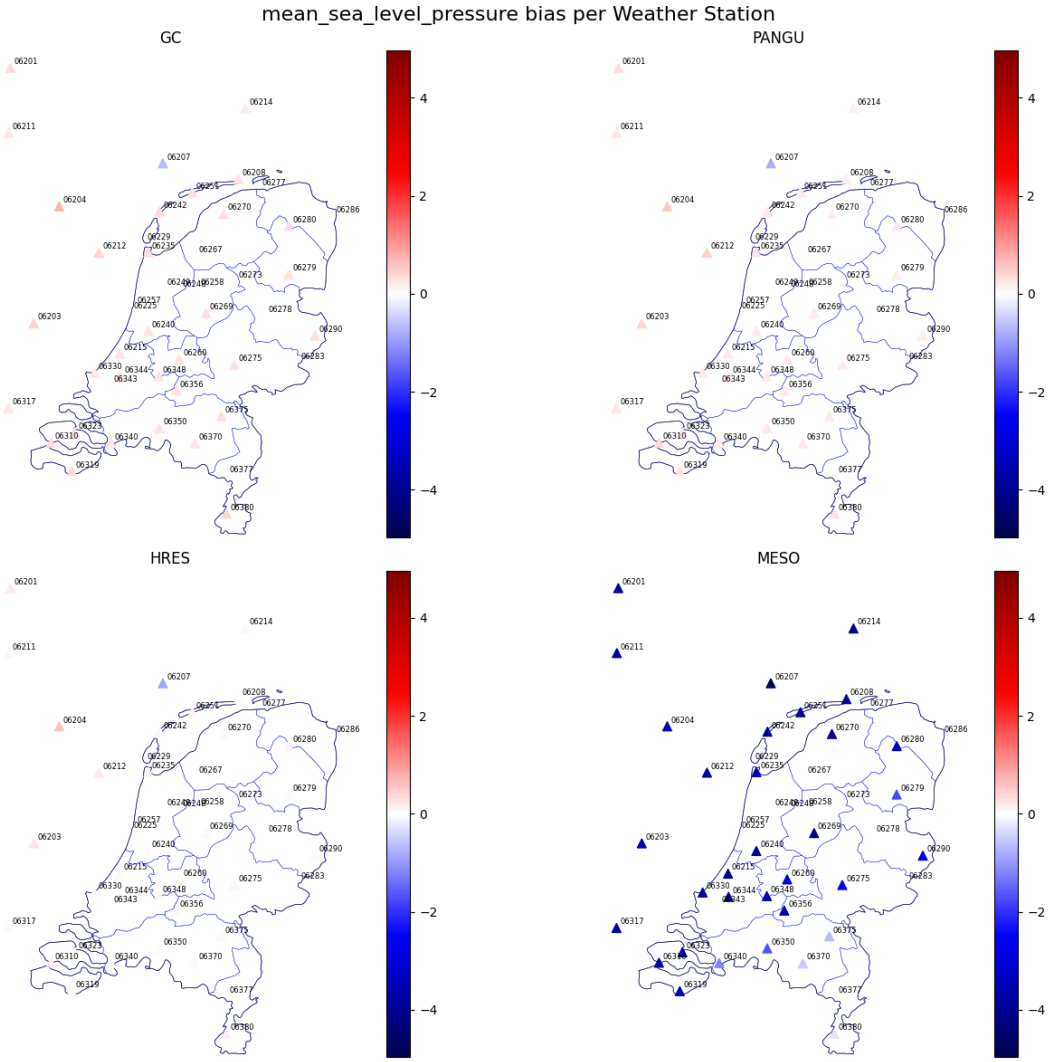


Figure A.17: Mean sea level pressure bias per weather station for various models.

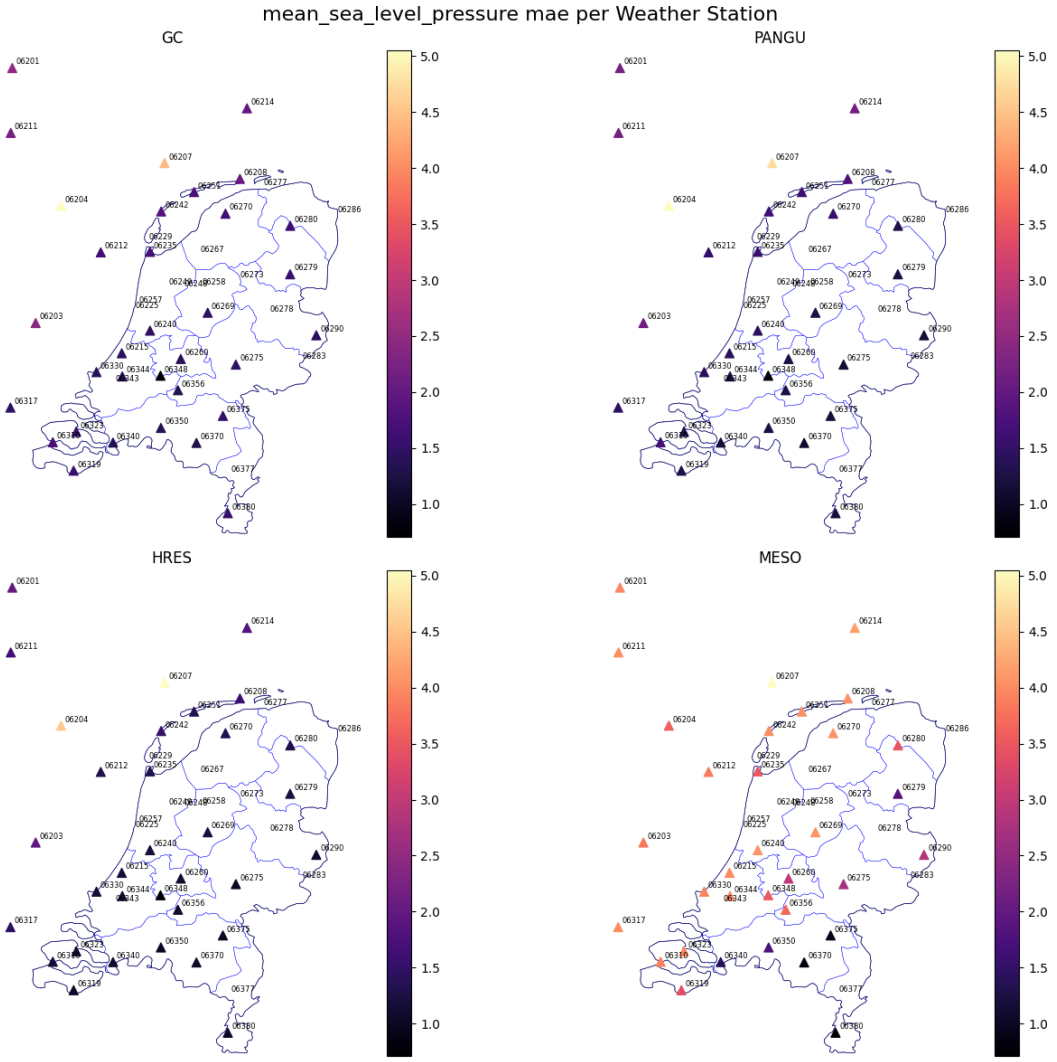


Figure A.18: Mean sea level pressure MAE per weather station for various models.

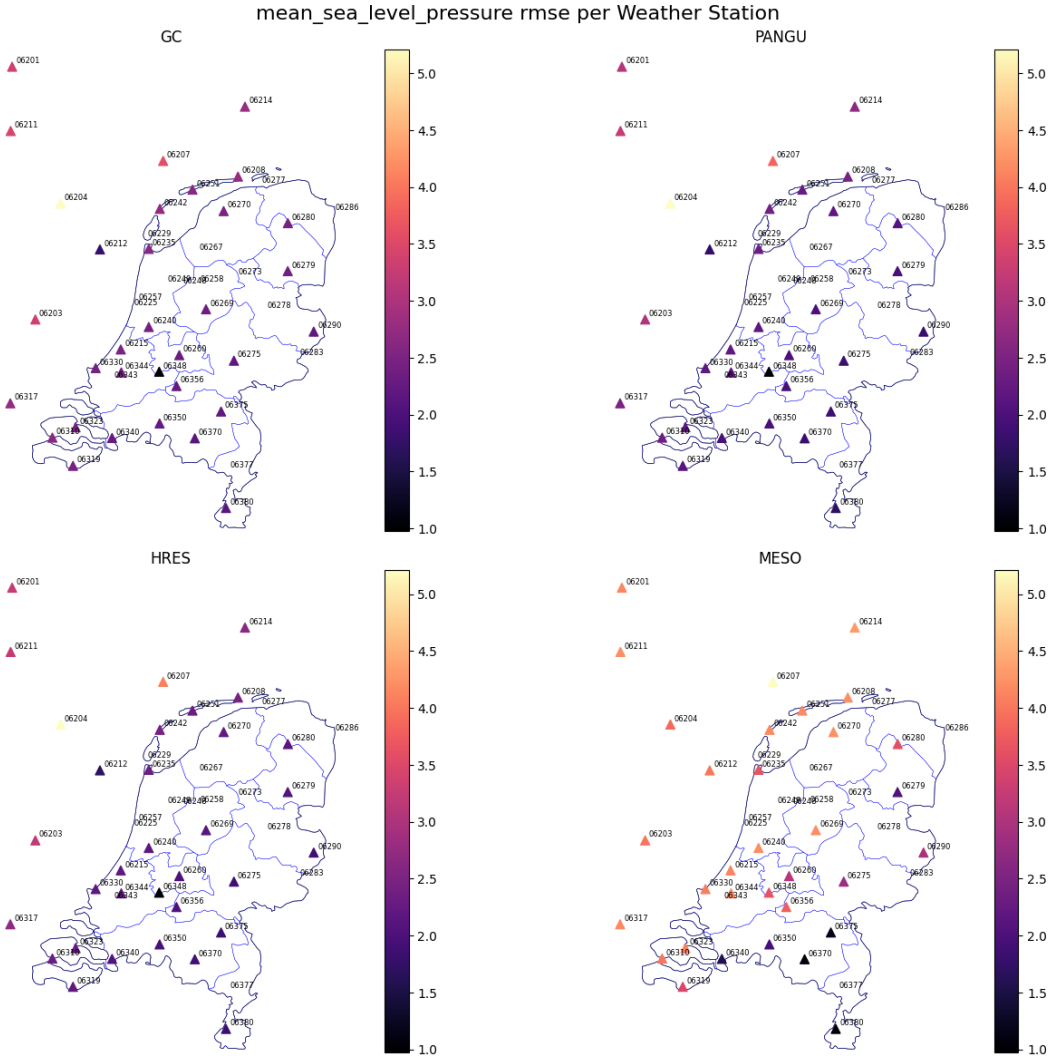


Figure A.19: Mean sea level pressure RMSE per weather station for various models.

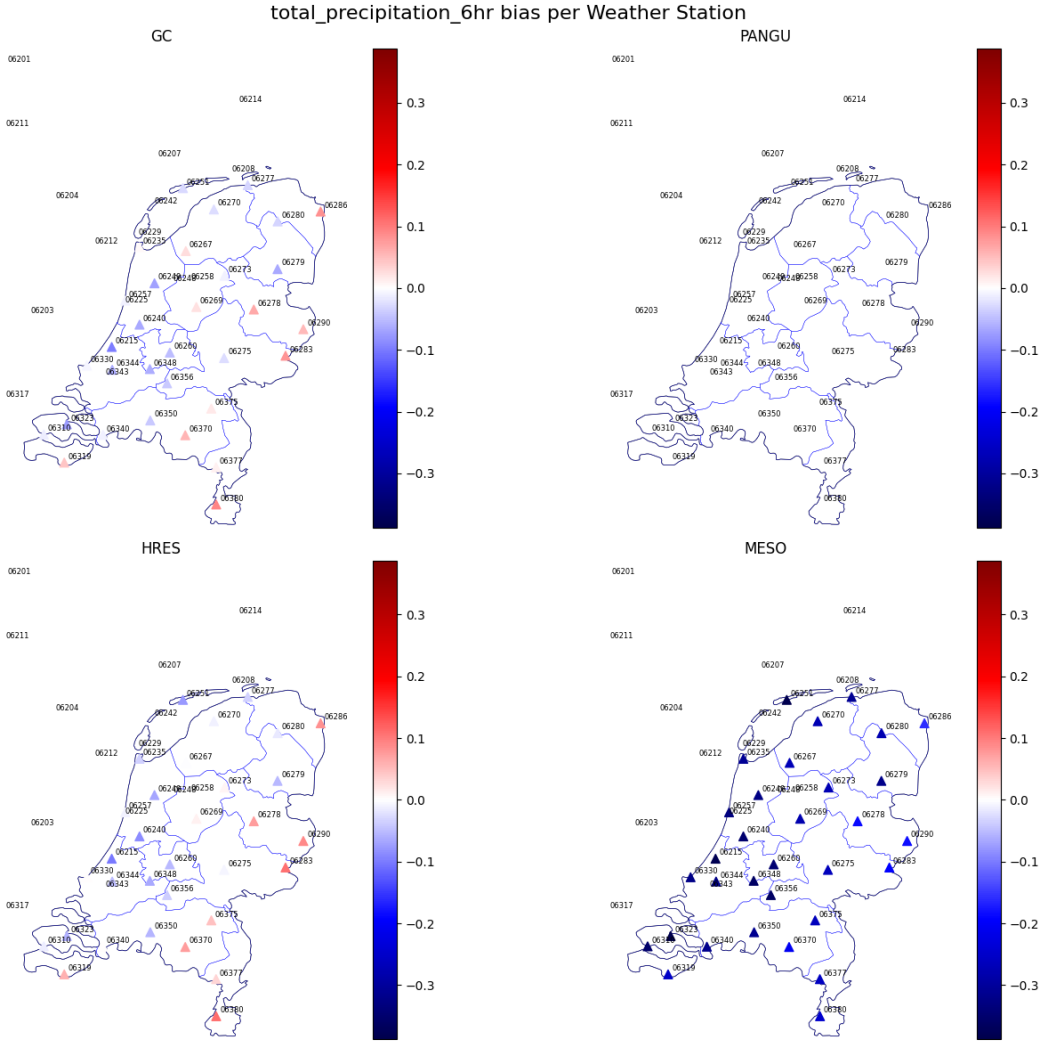


Figure A.20: Total precipitation (6hr) bias per weather station for various models.

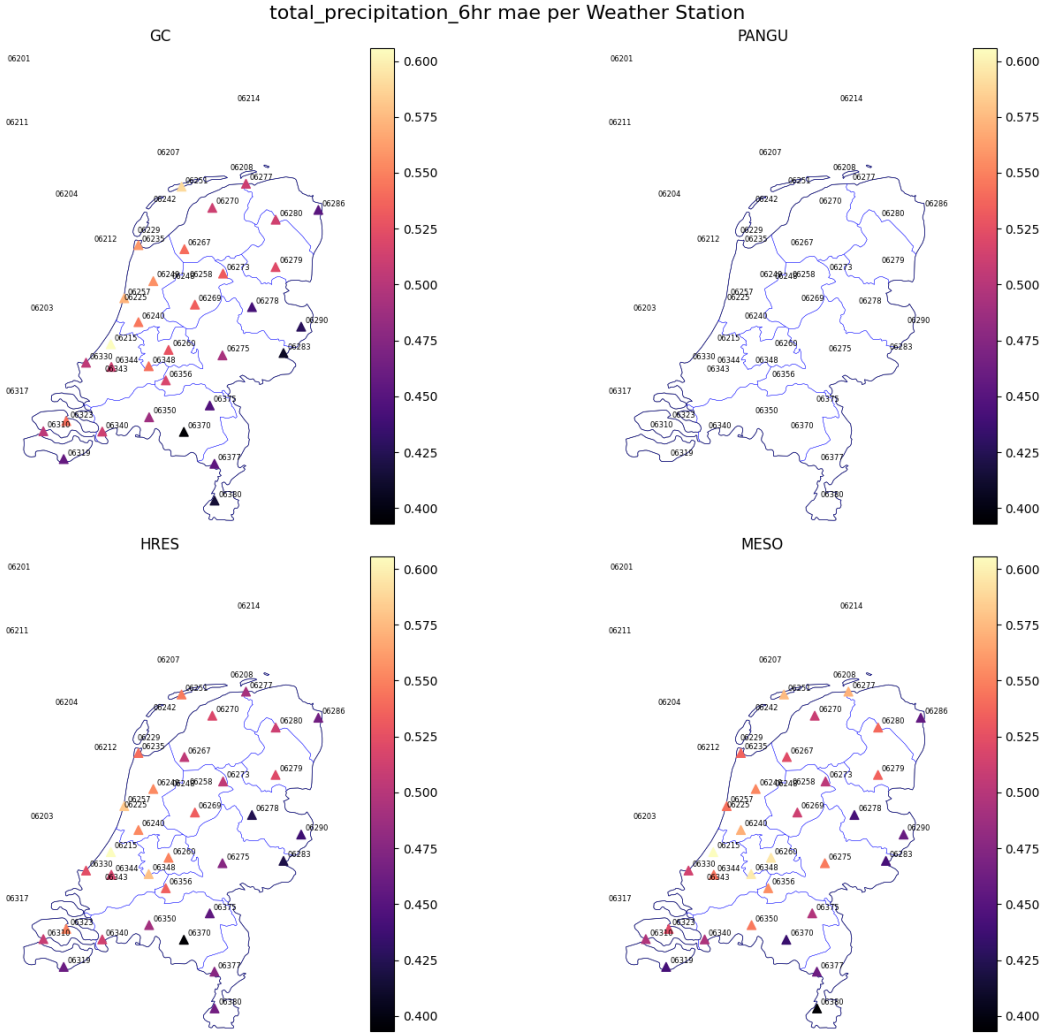


Figure A.21: Total precipitation (6hr) MAE per weather station for various models.

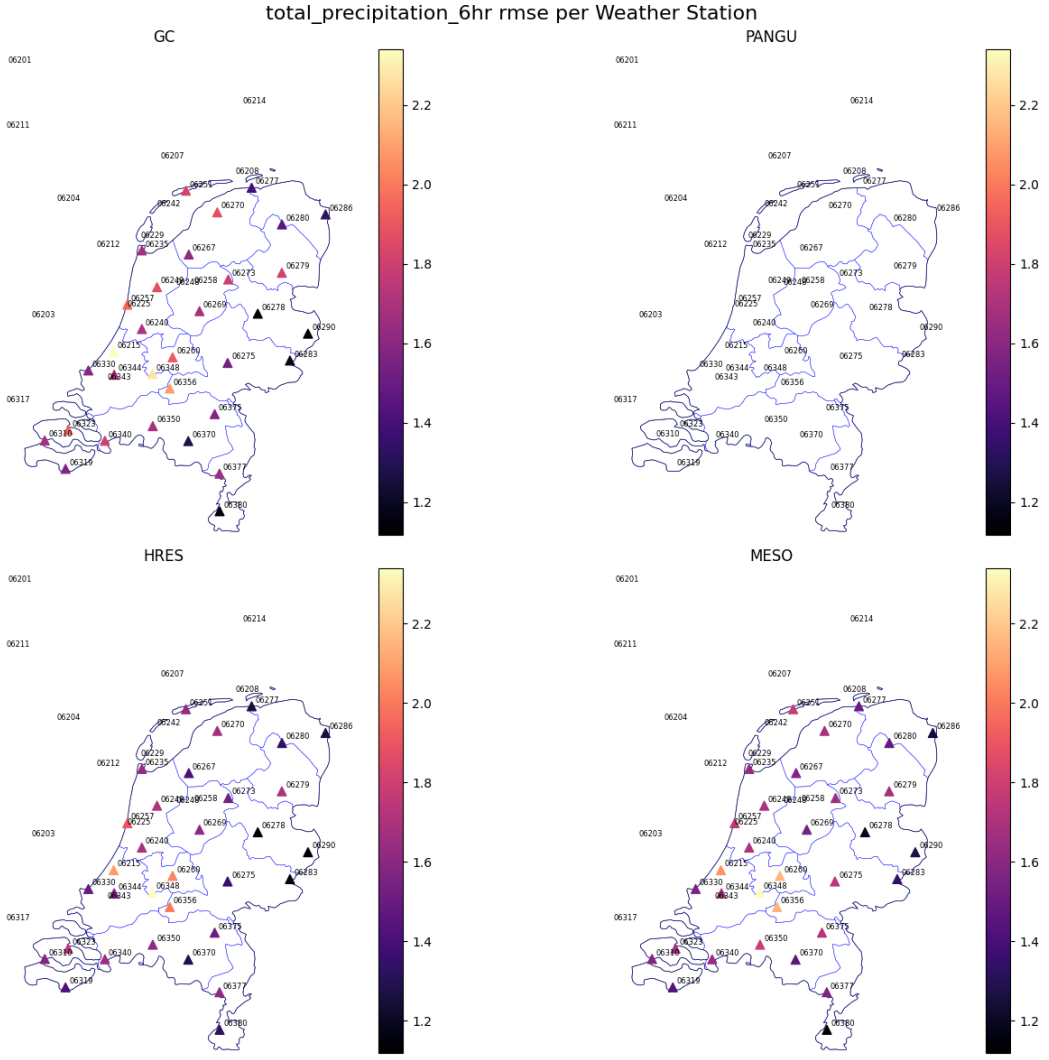


Figure A.22: Total precipitation (6hr) RMSE per weather station for various models.

A.4. Supplementary Taylor Diagrams

A.4.1. Taylor Diagrams

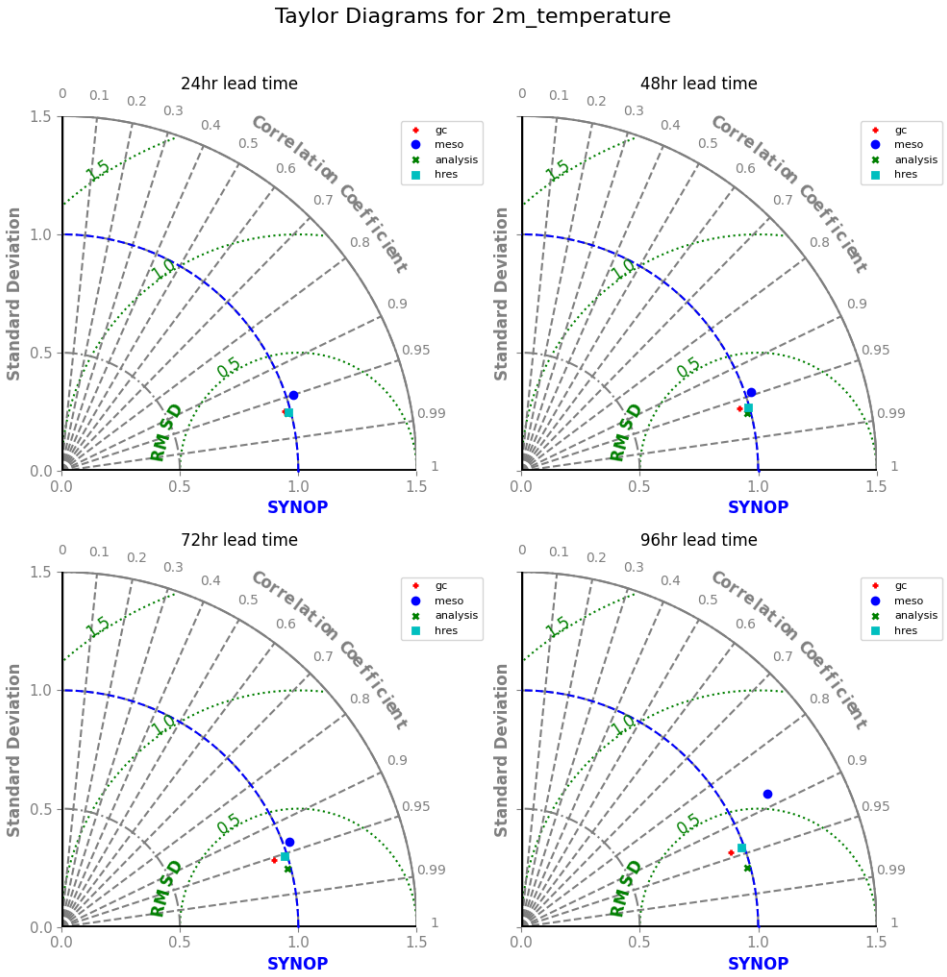


Figure A.23: Taylor diagram for 2m temperature, showing model performance at lead times of 24, 48, 72, and 96 hours.

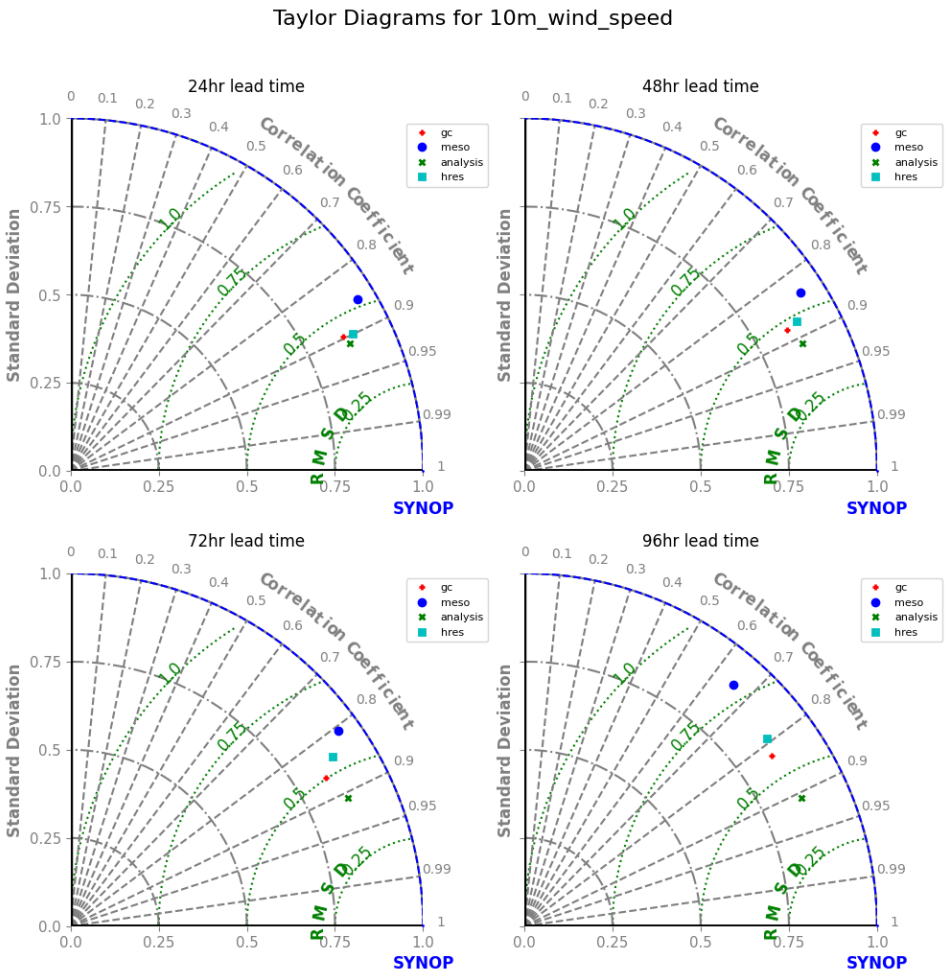


Figure A.24: Taylor diagram for 10m wind speed, showing model performance at lead times of 24, 48, 72, and 96 hours.

A.5. Supplementary Inter-model Similarity Analysis

A.5.1. Model Similarity Matrices

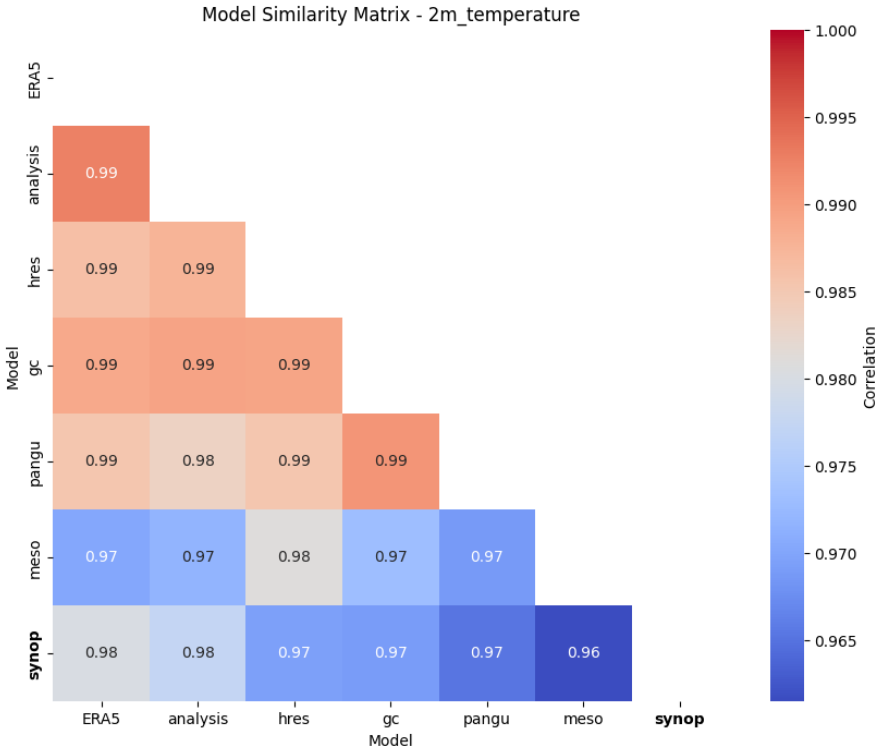


Figure A.25: Model similarity matrix for 2m temperature, showing the Pearson correlation coefficients between different models and observations.

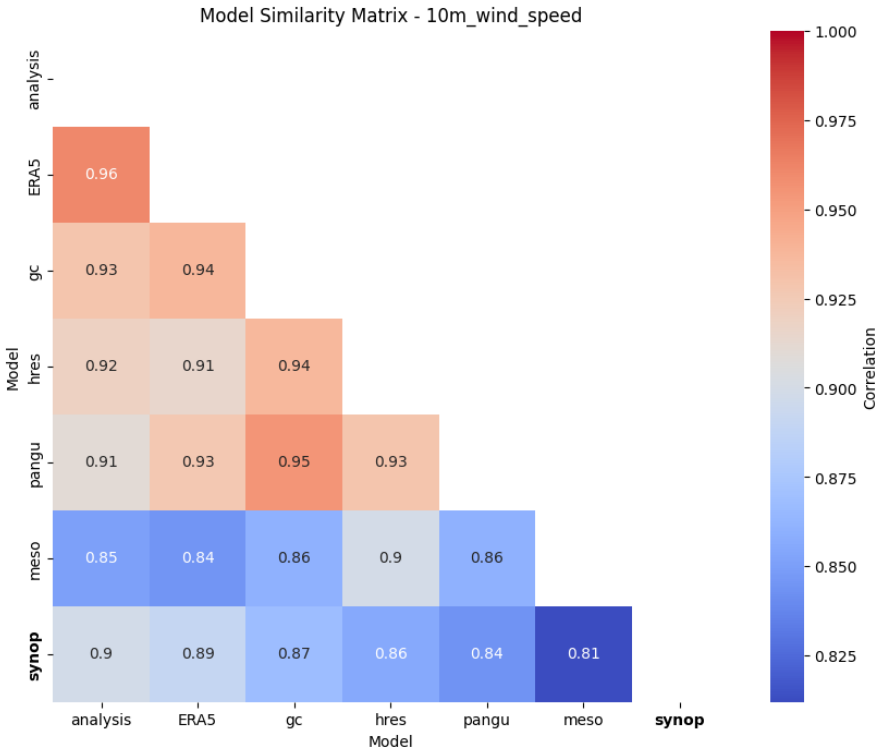


Figure A.26: Model similarity matrix for 10m wind, showing the Pearson correlation coefficients between different models and observations.

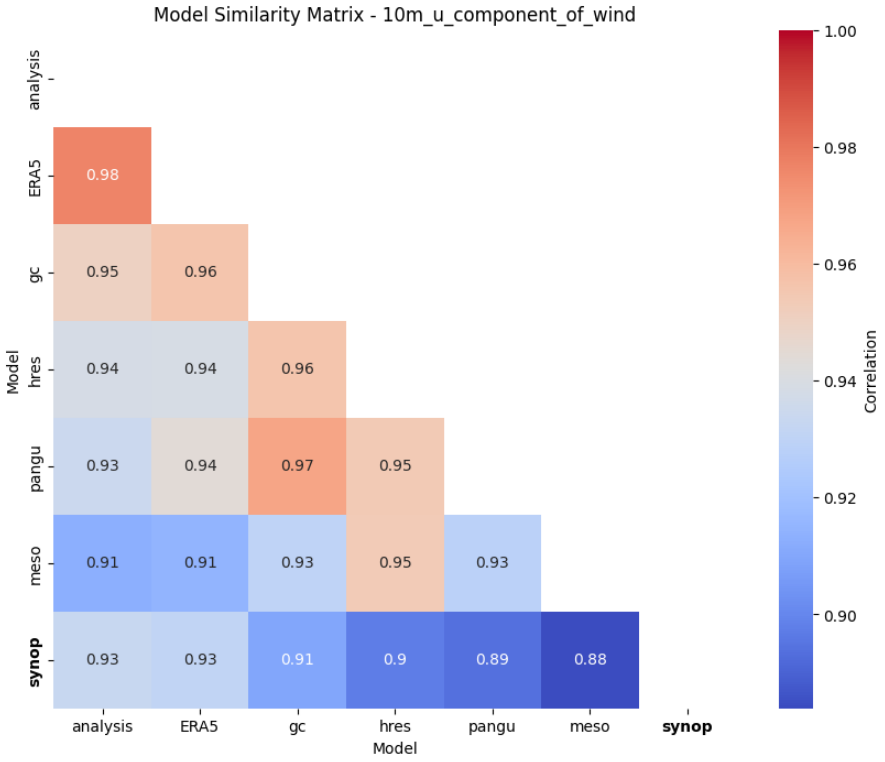


Figure A.27: Model similarity matrix for 10m u-component of wind, showing the Pearson correlation coefficients between different models and observations.

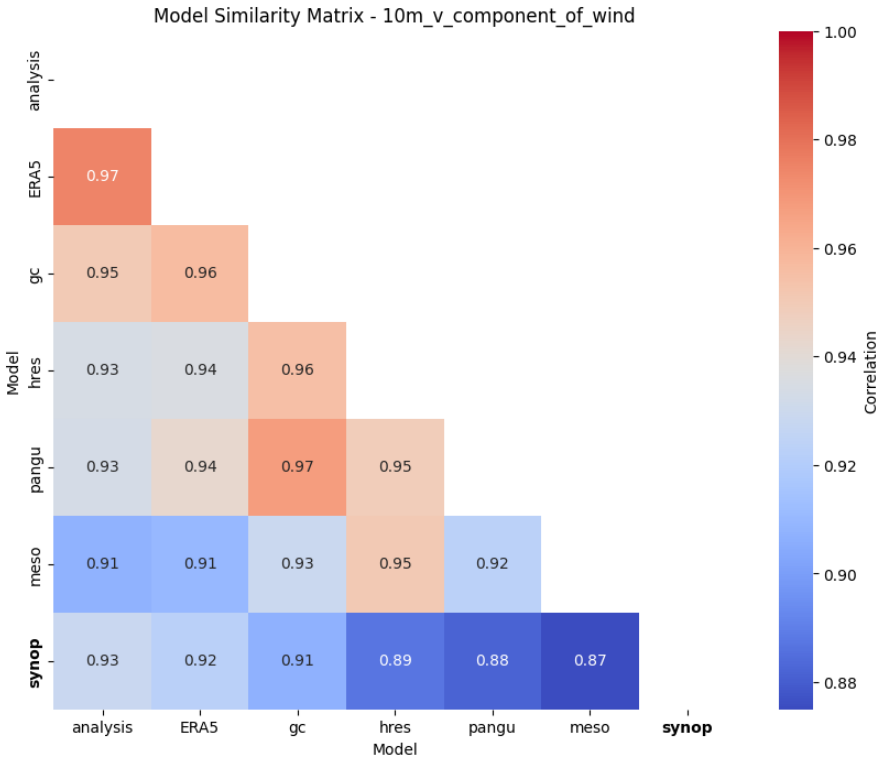


Figure A.28: Model similarity matrix for 10m v-component of wind, showing the Pearson correlation coefficients between different models and observations.

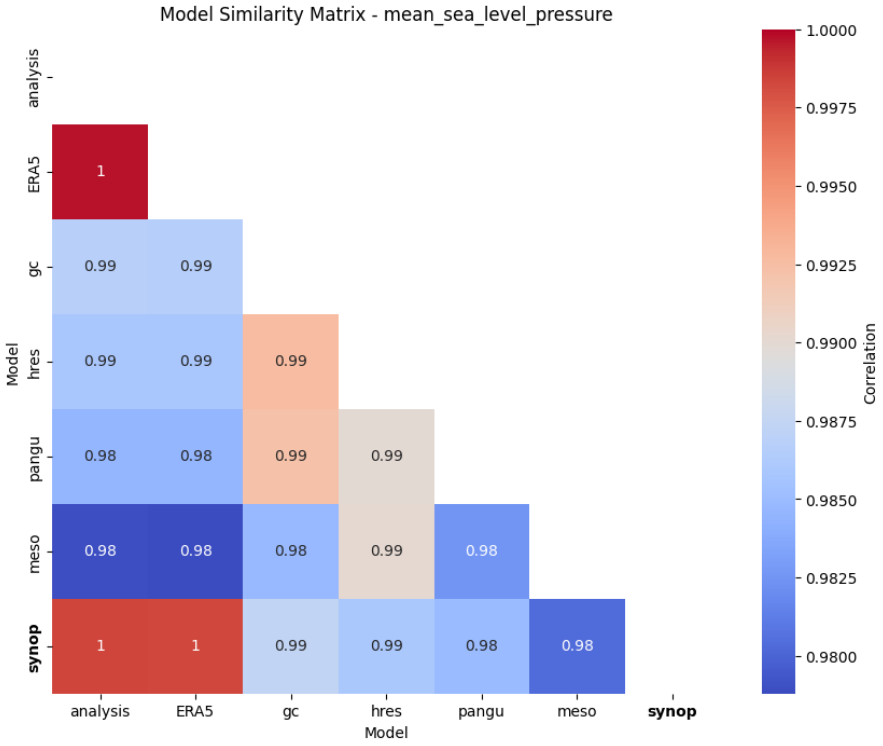


Figure A.29: Model similarity matrix for mean sea level pressure, showing the Pearson correlation coefficients between different models and observations.

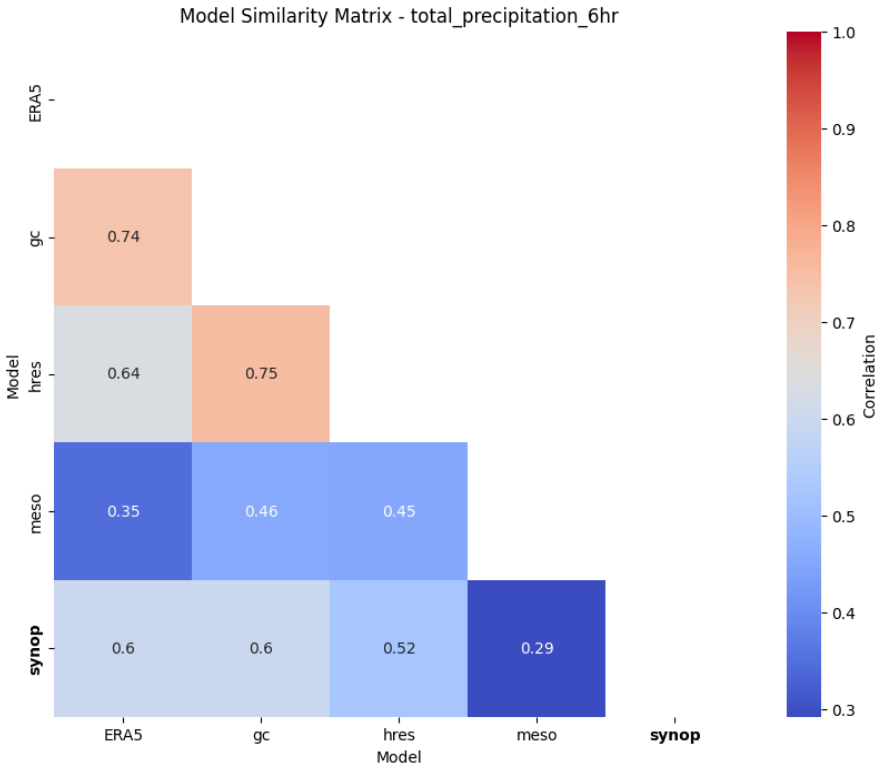


Figure A.30: Model similarity matrix for total precipitation (6hr), showing the Pearson correlation coefficients between different models and observations.

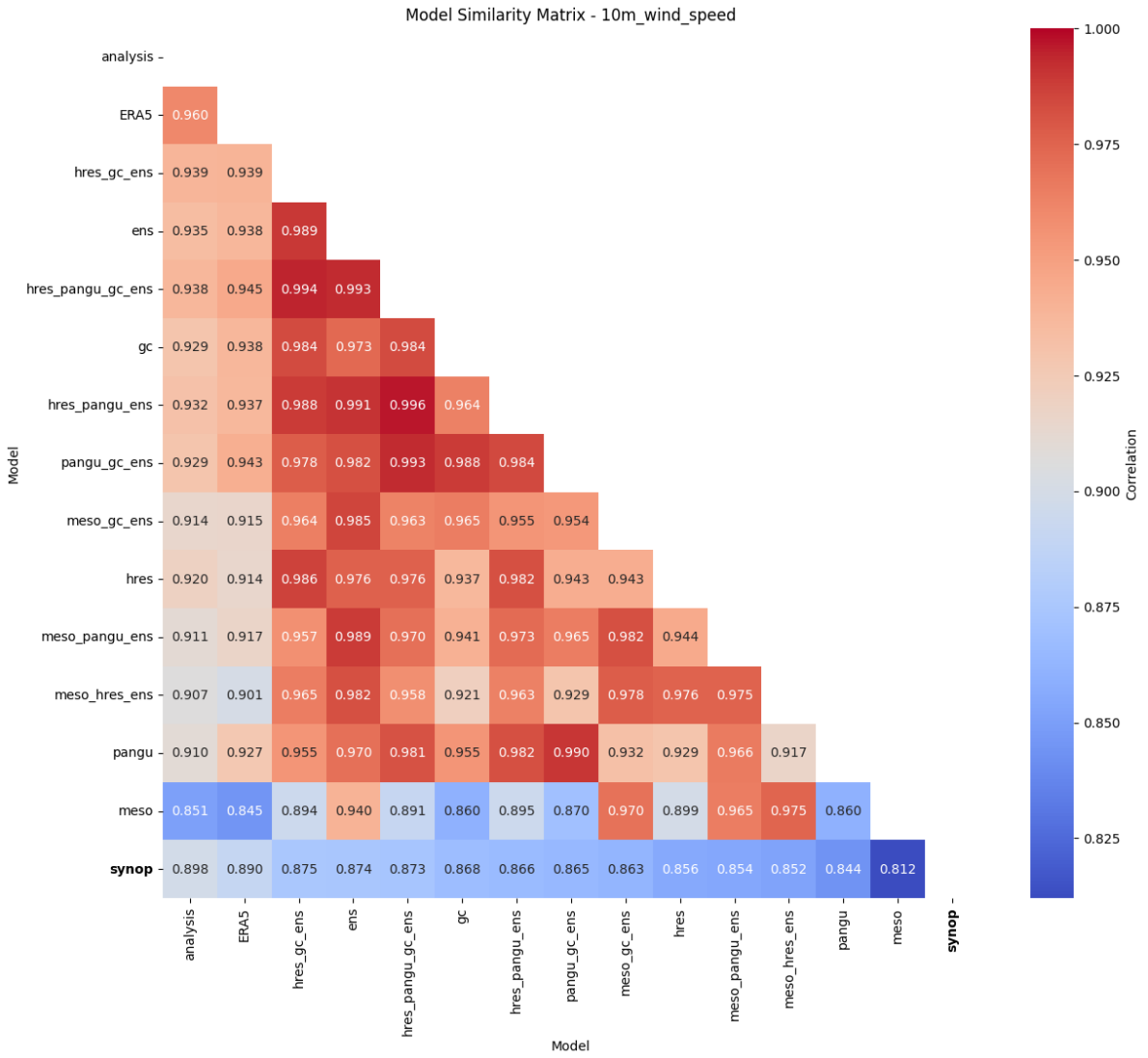


Figure A.31: Model similarity matrix for 10 m wind speed, showing the Pearson correlation coefficients between different models, observations and all ensembles.

A.6. Supplementary Scatter Plots

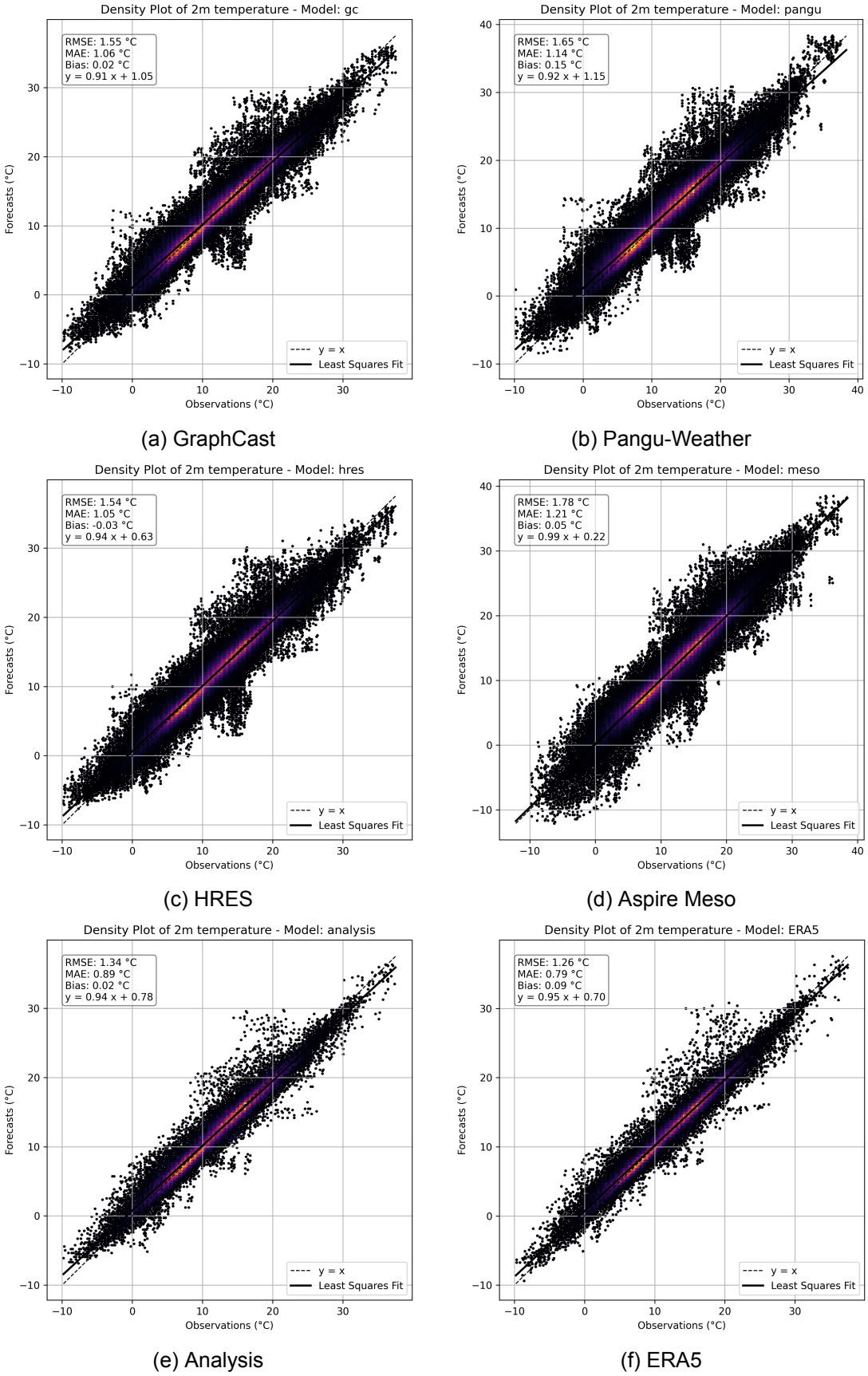


Figure A.32: Scatter plots of observations vs. forecasts for 2m Temperature using different weather models.

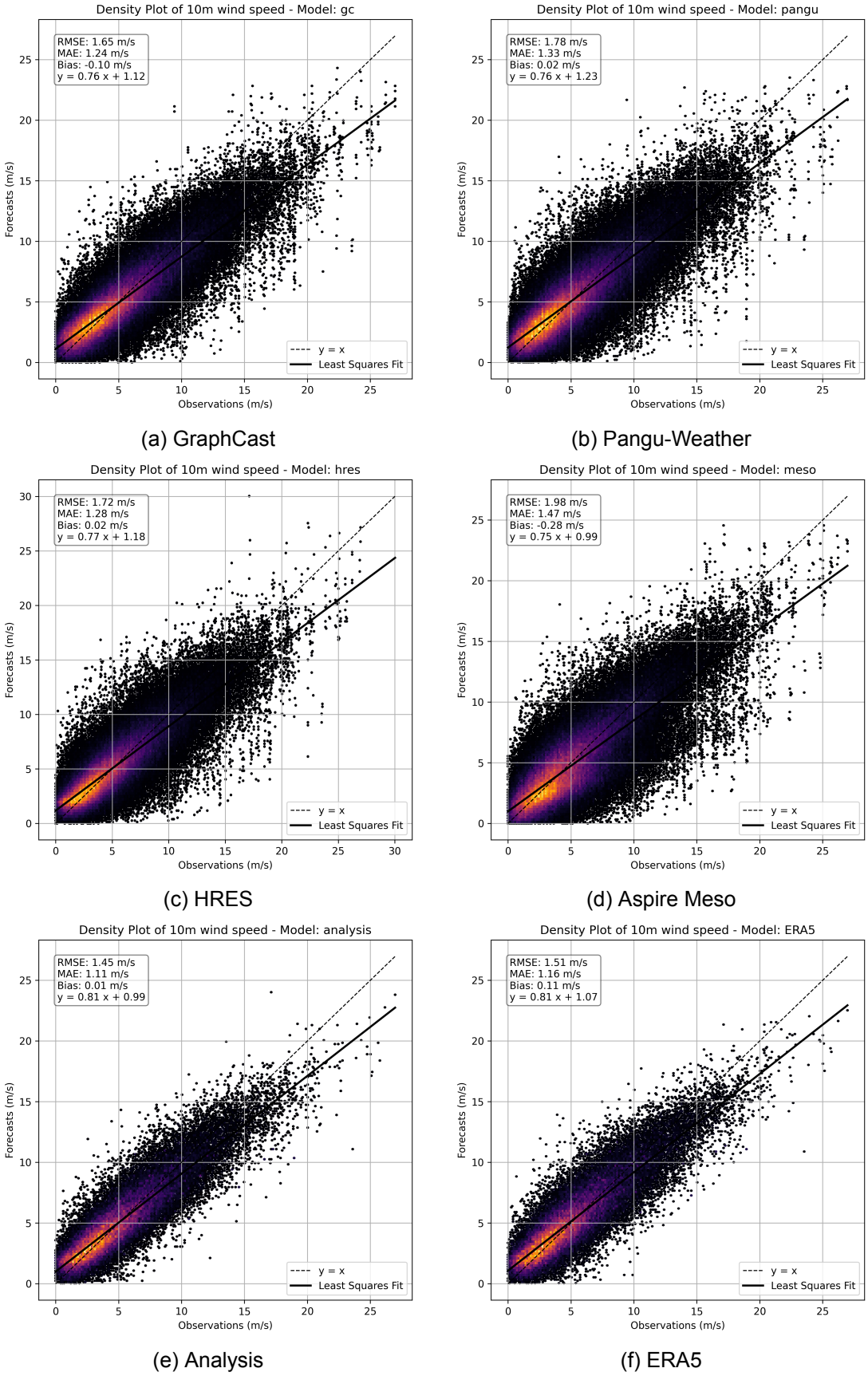
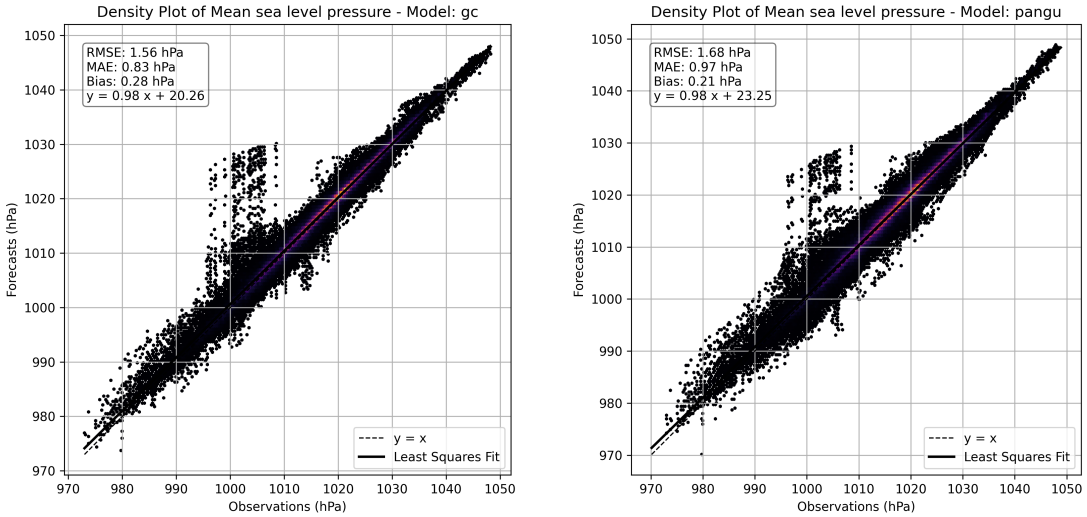
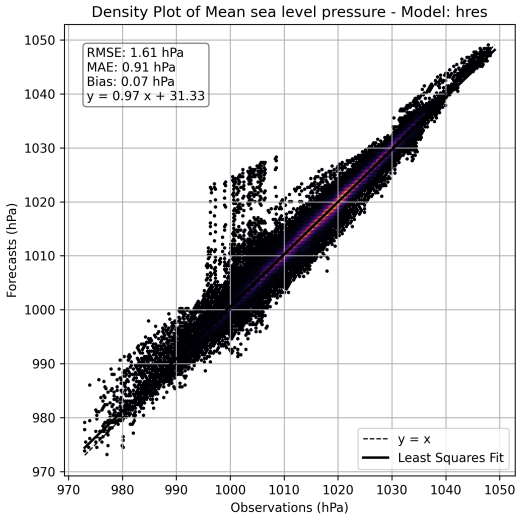


Figure A.33: Scatter plots of observations vs. forecasts for 10m Wind Speed using different weather models.

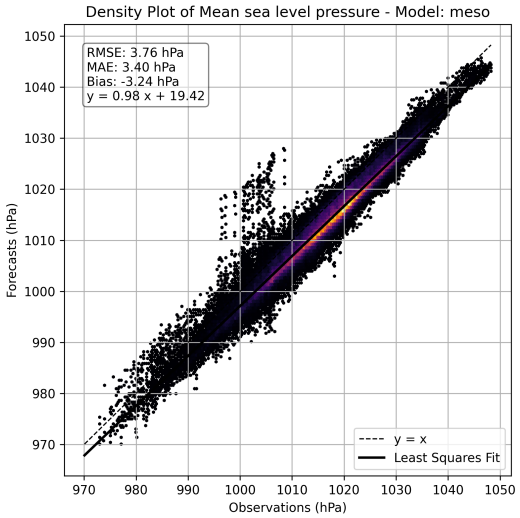


(a) GraphCast

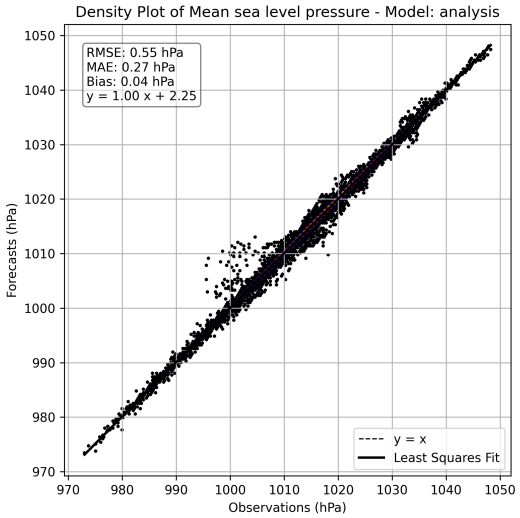
(b) Pangu-Weather



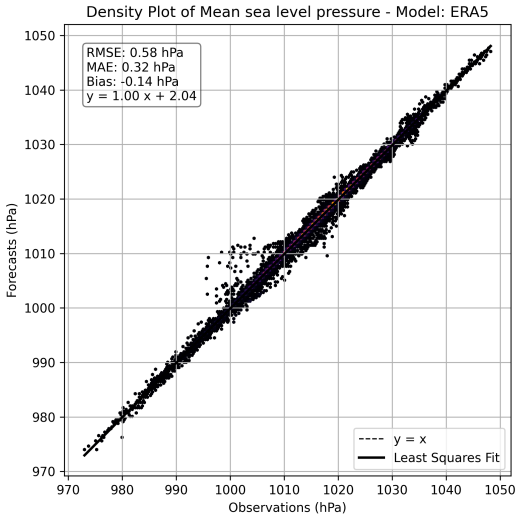
(c) HRES



(d) Aspire Meso



(e) Analysis



(f) ERA5

Figure A.34: Scatter plots of observations vs. forecasts for Mean Sea Level Pressure using different weather models.

A.7. Supplementary Time Series Plots

A.7.1. Time Series Plots

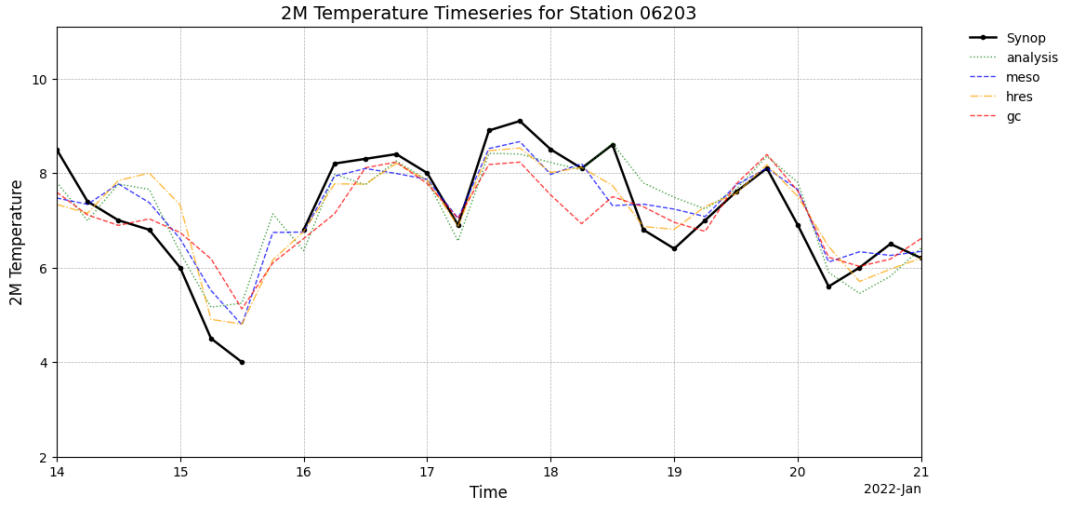


Figure A.35: Time series plots of 2m temperature at station 06203 from January 14 to January 21, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

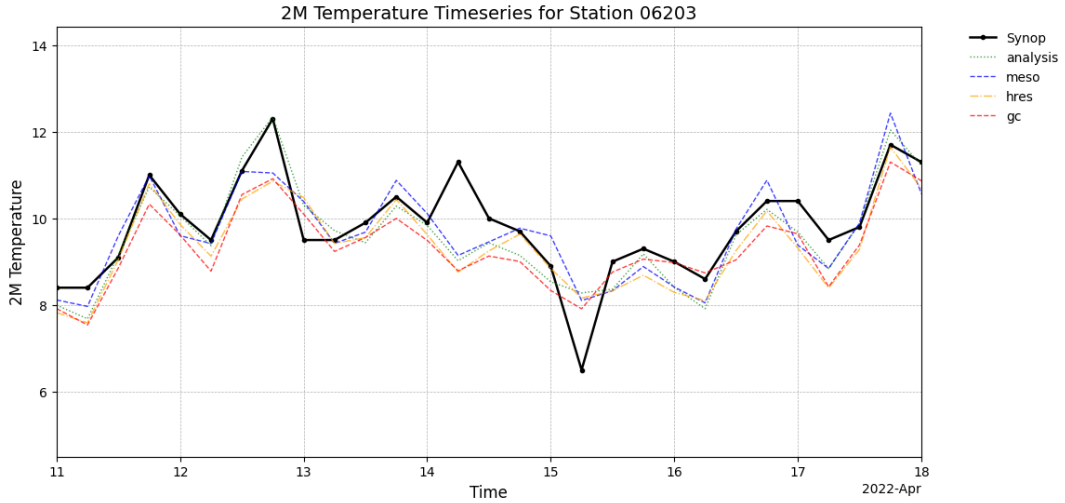


Figure A.36: Time series plots of 2m temperature at station 06203 from April 11 to April 18, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

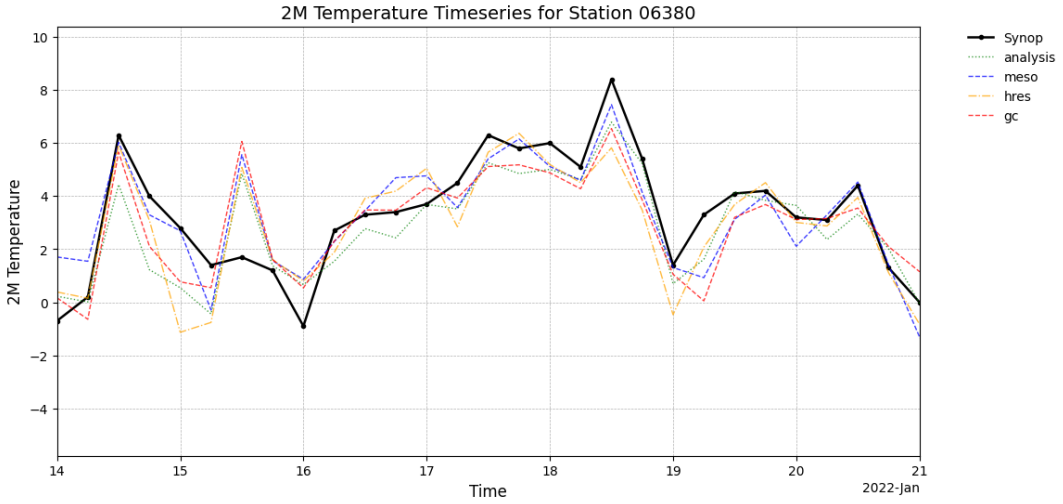


Figure A.37: Time series plots of 2m temperature at station 06380 from January 14 to January 21, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

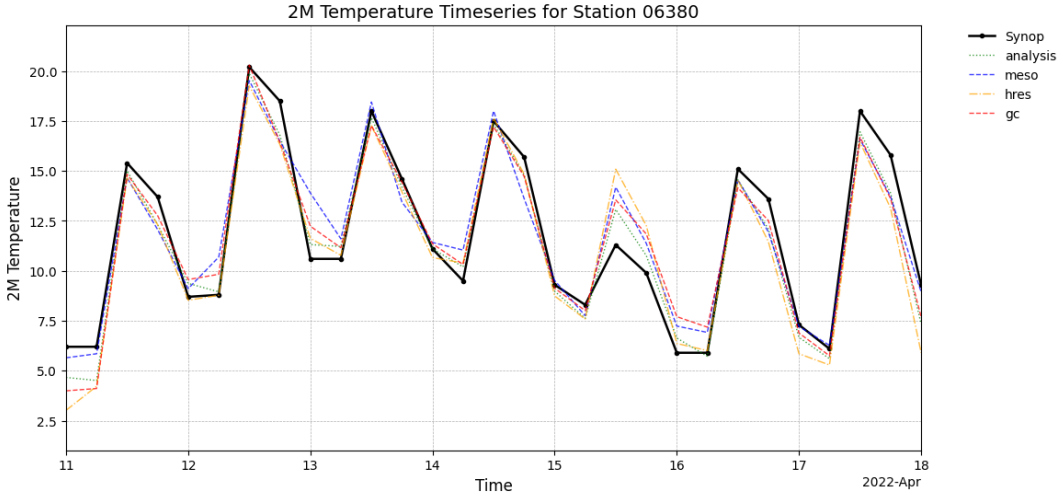


Figure A.38: Time series plots of 2m temperature at station 06380 from April 11 to April 18, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

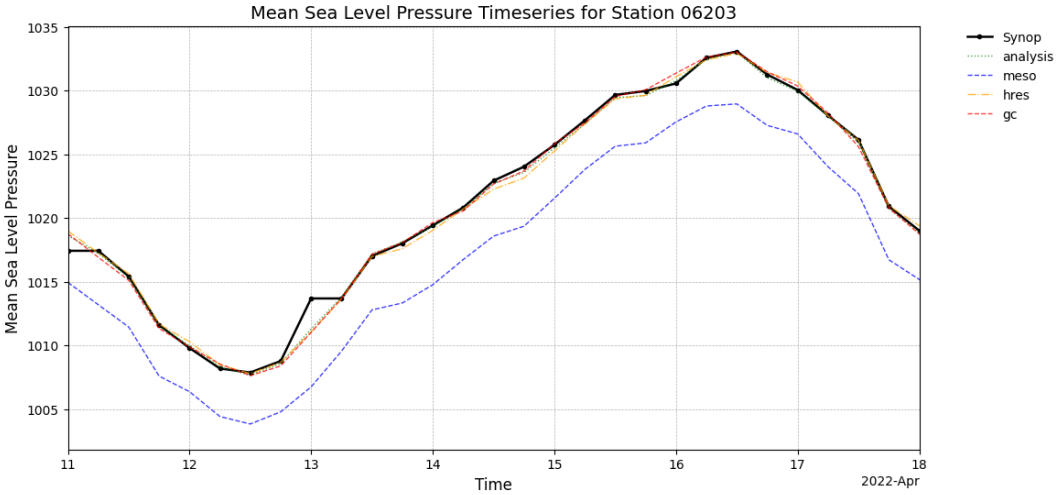


Figure A.39: Time series plots of mean sea level pressure at station 06203 from April 11 to April 18, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

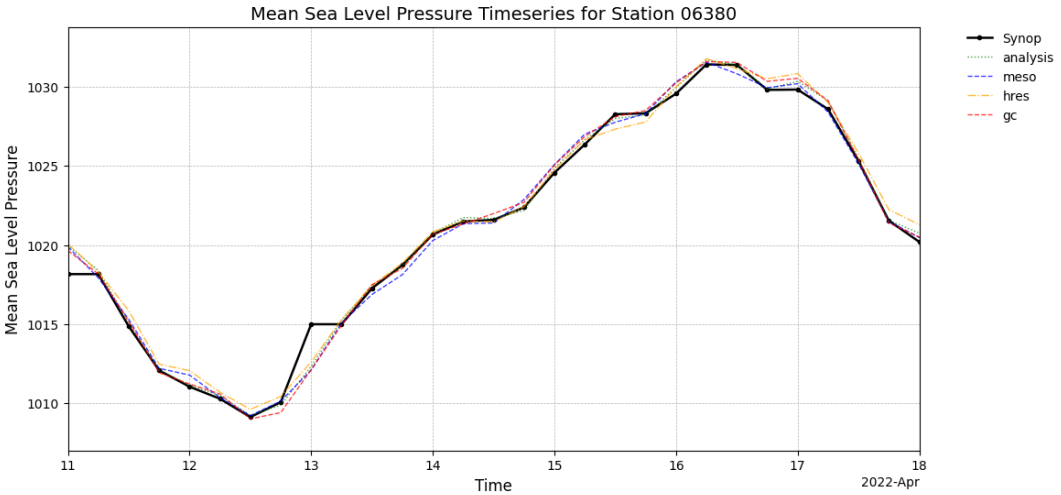


Figure A.40: Time series plots of mean sea level pressure at station 06380 from April 11 to April 18, 2022. Observations are shown in black, GraphCast forecasts in dotted green, HRES in dashed orange, and Aspire Meso in dashed blue.

B

Detailed Analysis of LES Model Performance

In this section, we extend our analysis to include the Aspire Large Eddy Simulation (LES) model, which was run at two specific locations among the 47 KNMI SYNOP stations: station 06203, P11-B (offshore), and station 06348, Cabauw (onshore). The choice of these stations enables an examination of model performance both in an offshore environment, characterized by open water conditions, and an onshore environment with more complex surface interactions.

The following figures depict the lead-time Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Bias metrics for various weather variables and models. Each figure illustrates how the Aspire LES model performs over different lead times, providing insight into its strengths and limitations for short-range and extended forecasts. Our aim in this detailed analysis is to highlight the distinct behavior of the model in the two locations, reflecting the variability in atmospheric processes and surface interactions that can influence forecast accuracy.

The results depicted in these figures demonstrate the differences in error propagation between the offshore P11-B station and the onshore Cabauw station. In particular, we observe that the Aspire LES model's performance is influenced by the station environment, with distinct lead-time dependency trends visible for each metric. The onshore environment of Cabauw, with its complex terrain and boundary layer processes, tends to introduce greater variability in forecast accuracy compared to the relatively homogeneous surface conditions offshore at P11-B.

This analysis emphasizes the importance of location-specific model evaluations, as localized atmospheric dynamics can significantly affect the predictive skill of weather models. Th

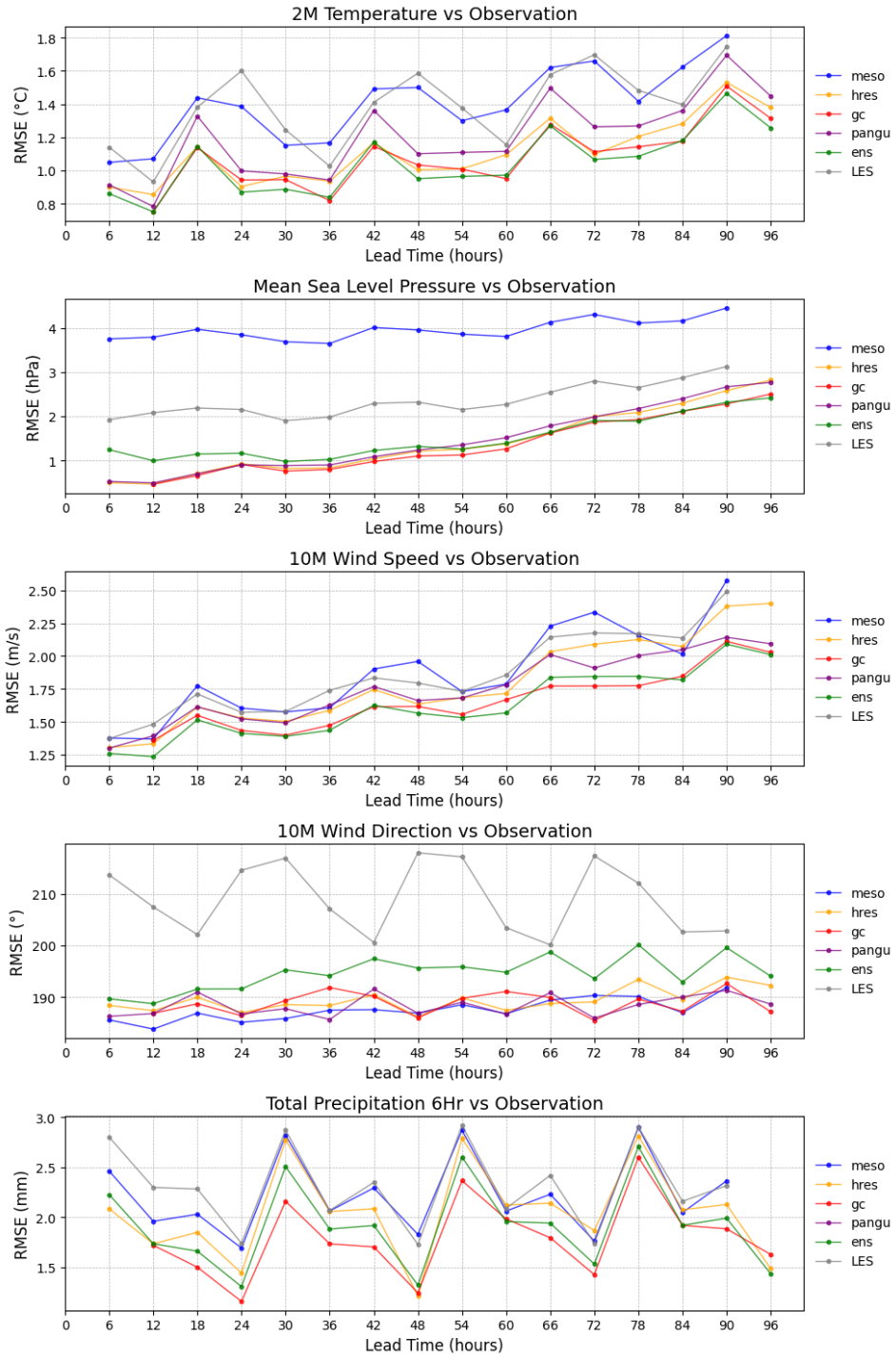


Figure B.1: Lead time RMSE as a function of lead time for various weather variables and models at station 06203 (P11-B) and station 06348 (Cabauw). The plot illustrates the evolving error in forecasts with increasing lead time, emphasizing the relative performance differences in offshore versus onshore settings.

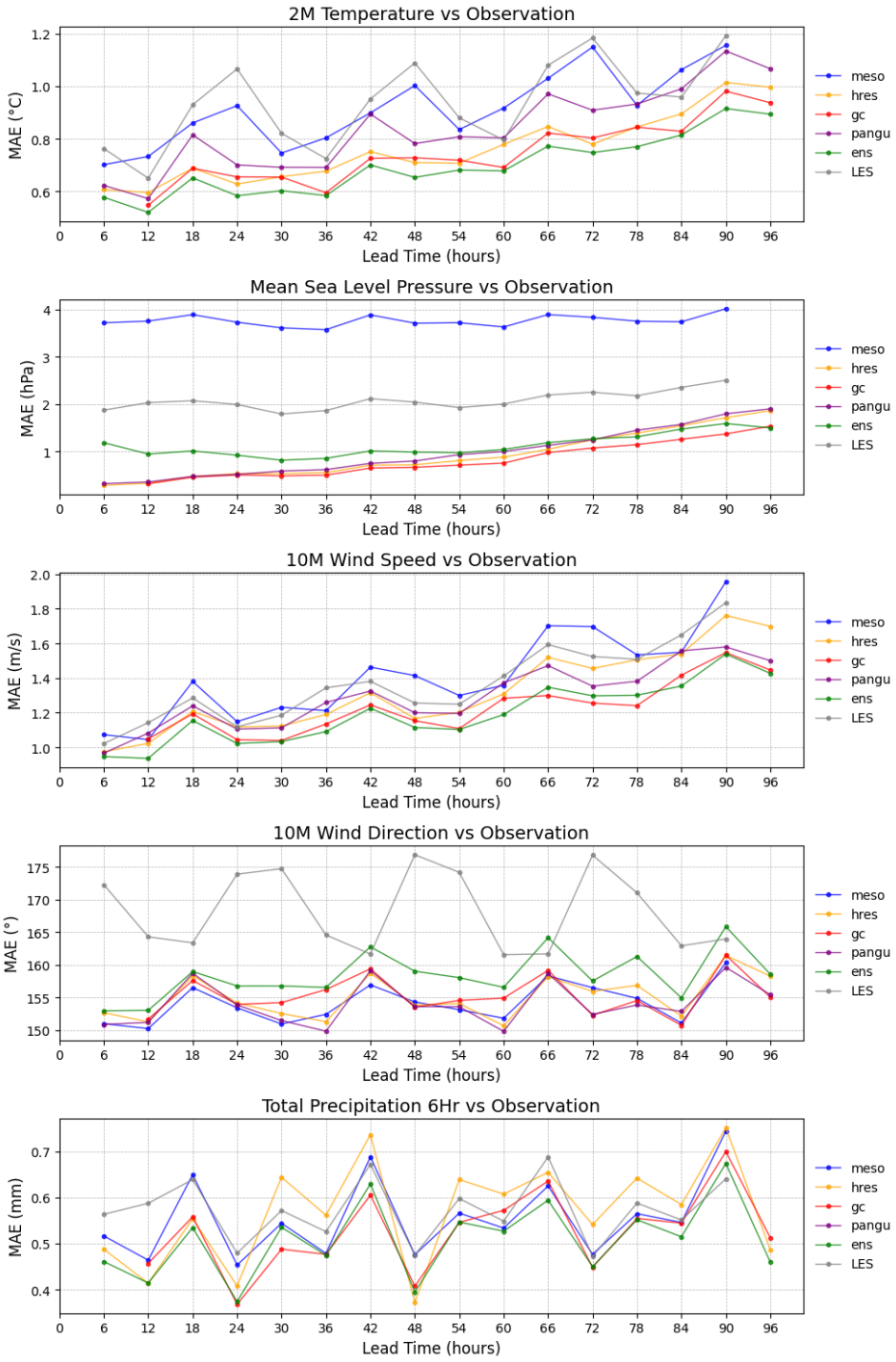


Figure B.2: Lead time MAE as a function of lead time for various weather variables and models at station 06203 (P11-B) and station 06348 (Cabauw). This figure provides a measure of average forecast errors over time, offering a comparison of accuracy across lead times in both offshore and onshore conditions.

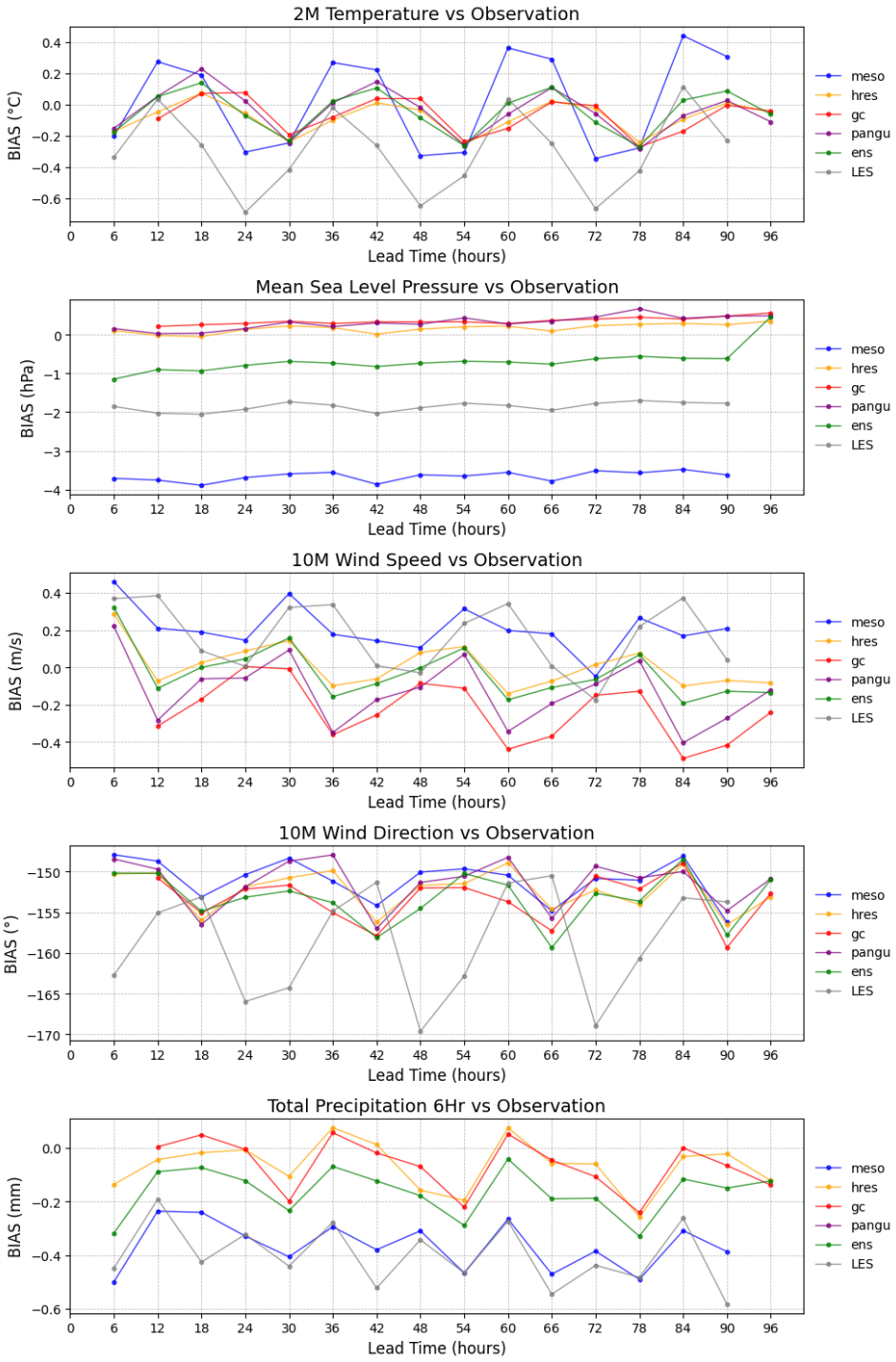


Figure B.3: Lead time Bias as a function of lead time for various weather variables and models at station 06203 (P11-B) and station 06348 (Cabauw). The bias plots indicate systematic over- or under-prediction tendencies, revealing critical insights into model behavior and calibration needs at both locations.

C

Imbalance Cost Calculation for Example Wind Farm

C.1. Introduction

This appendix provides a detailed calculation of the imbalance cost due to weather forecasting errors for the Gemini offshore wind farm in the North Sea. We use five different weather forecasting models, including an ensemble of all models, to estimate their impact on the financial performance of the wind farm. The forecast errors are evaluated at a 24-hour lead time, which is crucial for the day-ahead electricity market, where accurate predictions are necessary for scheduling power delivery and minimizing imbalance costs.

C.2. Assumptions

The following assumptions and model errors were made for the imbalance cost calculation:

- **Installed Capacity** ($P_{\text{installed}}$): 600 MW, representing the maximum potential output of the wind farm.
- **Capacity Factor** (CF): 40%, indicating the average utilization of the wind farm over time, accounting for varying wind conditions.
- **Annual Hours** (H_{year}): 8760 hours, representing the total number of hours in a year.
- **Wind Speed to Power Error Conversion Factor**: A 1% error in wind speed forecasts results in a 1.5% error in wind power generation.
- **Imbalance Cost** ($C_{\text{imbalance}}$): 6 euros per megawatt-hour (€/MWh), reflecting the average financial penalty due to deviations from scheduled power delivery (PBL, 2022).
- **Baseline Model (HRES)**: The average wind speed forecasting error at a 24-hour lead time is 1.153 m/s.
- **GraphCast Model**: This model shows an improved average error of 1.134 m/s, representing a reduction in forecast error relative to the baseline.
- **Pangu-Weather Model**: This model has a higher average error of 1.251 m/s, representing a deterioration in forecast error relative to the baseline.
- **Aspire Meso Model**: This model has a higher average error of 1.288 m/s, indicating a deterioration in forecast accuracy compared to the baseline.
- **Four Model Ensemble**: The ensemble of all four models shows an improved average error of 1.097 m/s, representing a reduction in forecast error relative to the baseline.

C.3. Annual Energy Production

The expected annual energy production of the wind farm is calculated using the formula:

$$E_{\text{annual}} = P_{\text{installed}} \times CF \times H_{\text{year}} \tag{C.1}$$

Substituting the values:

$$E_{\text{annual}} = 600\,\text{MW} \times 0.4 \times 8760\,\text{hours} = 2,102,400\,\text{MWh} \tag{C.2}$$

This represents the total energy production for the year under average wind conditions.

C.4. Baseline Imbalance Cost Calculation

The total imbalance cost for the baseline (HRES) model is calculated as:

$$\text{Total Imbalance Cost}_{\text{baseline}} = E_{\text{annual}} \times C_{\text{imbalance}} \tag{C.3}$$

$$\text{Total Imbalance Cost}_{\text{baseline}} = 2,102,400\,\text{MWh} \times 6\,€/\text{MWh} = 12,614,400\,\text{euros} \tag{C.4}$$

C.5. Adjusted Imbalance Costs for Model Performance Changes

To evaluate the impact of forecast model differences, the total imbalance cost is adjusted based on the relative forecast errors of each model.

C.5.1. GraphCast Model

The relative error reduction factor for the GraphCast model is given by:

$$\text{Relative Error Reduction}_{\text{GraphCast}} = \frac{1.134}{1.153} \tag{C.5}$$

The adjusted imbalance cost is calculated as:

$$\text{Total Imbalance Cost}_{\text{GraphCast}} = \text{Total Imbalance Cost}_{\text{baseline}} \times \text{Relative Error Reduction}_{\text{GraphCast}} \tag{C.6}$$

$$\text{Total Imbalance Cost}_{\text{GraphCast}} = 12,614,400\,\text{euros} \times \frac{1.134}{1.153} = 12,400,996\,\text{euros} \tag{C.7}$$

C.5.2. Pangu-Weather Model

The relative error increase factor for the Pangu-Weather model is given by:

$$\text{Relative Error Increase}_{\text{Pangu}} = \frac{1.251}{1.153} \tag{C.8}$$

The adjusted imbalance cost is calculated as:

$$\text{Total Imbalance Cost}_{\text{Pangu}} = \text{Total Imbalance Cost}_{\text{baseline}} \times \text{Relative Error Increase}_{\text{Pangu}} \tag{C.9}$$

$$\text{Total Imbalance Cost}_{\text{Pangu}} = 12,614,400\,\text{euros} \times \frac{1.251}{1.153} = 13,682,563\,\text{euros} \tag{C.10}$$

C.5.3. Aspire Meso Model

The relative error increase factor for the Aspire Meso model is given by:

$$\text{Relative Error Increase}_{\text{Aspire Meso}} = \frac{1.288}{1.153} \tag{C.11}$$

The adjusted imbalance cost is calculated as:

$$\text{Total Imbalance Cost}_{\text{Aspire Meso}} = \text{Total Imbalance Cost}_{\text{baseline}} \times \text{Relative Error Increase}_{\text{Aspire Meso}} \tag{C.12}$$

$$\text{Total Imbalance Cost}_{\text{Aspire Meso}} = 12,614,400\,\text{euros} \times \frac{1.288}{1.153} = 14,084,579\,\text{euros} \tag{C.13}$$

C.5.4. Four Model Ensemble

The relative error reduction factor for the four model ensemble is given by:

$$\text{Relative Error Reduction}_{\text{Ensemble}} = \frac{1.097}{1.153} \tag{C.14}$$

The adjusted imbalance cost is calculated as:

$$\text{Total Imbalance Cost}_{\text{Ensemble}} = \text{Total Imbalance Cost}_{\text{baseline}} \times \text{Relative Error Reduction}_{\text{Ensemble}} \tag{C.15}$$

$$\text{Total Imbalance Cost}_{\text{Ensemble}} = 12,614,400\,\text{euros} \times \frac{1.097}{1.153} = 12,004,730\,\text{euros} \tag{C.16}$$

C.6. Cost Impact Assessment

To evaluate the cost impact of each model compared to the baseline, we calculate the savings or additional costs.

C.6.1. GraphCast Model Cost Savings

$$\text{Cost Savings}_{\text{GraphCast}} = \text{Total Imbalance Cost}_{\text{baseline}} - \text{Total Imbalance Cost}_{\text{GraphCast}} \tag{C.17}$$

$$\text{Cost Savings}_{\text{GraphCast}} = 12,614,400\,\text{euros} - 12,400,996\,\text{euros} = 213,404\,\text{euros} \tag{C.18}$$

C.6.2. Pangu-Weather Model Additional Cost

$$\text{Additional Cost}_{\text{Pangu}} = \text{Total Imbalance Cost}_{\text{Pangu}} - \text{Total Imbalance Cost}_{\text{baseline}} \tag{C.19}$$

$$\text{Additional Cost}_{\text{Pangu}} = 13,682,563\,\text{euros} - 12,614,400\,\text{euros} = 1,068,163\,\text{euros} \tag{C.20}$$

C.6.3. Aspire Meso Model Additional Cost

$$\text{Additional Cost}_{\text{Aspire Meso}} = \text{Total Imbalance Cost}_{\text{Aspire Meso}} - \text{Total Imbalance Cost}_{\text{baseline}} \tag{C.21}$$

$$\text{Additional Cost}_{\text{Aspire Meso}} = 14,084,579\,\text{euros} - 12,614,400\,\text{euros} = 1,470,179\,\text{euros} \tag{C.22}$$

C.6.4. Four Model Ensemble Cost Savings

$$\text{Cost Savings}_{\text{Ensemble}} = \text{Total Imbalance Cost}_{\text{baseline}} - \text{Total Imbalance Cost}_{\text{Ensemble}} \tag{C.23}$$

$$\text{Cost Savings}_{\text{Ensemble}} = 12,614,400\,\text{euros} - 12,004,730\,\text{euros} = 609,670\,\text{euros} \tag{C.24}$$

This expanded analysis demonstrates the varying financial impacts of different forecast models and an ensemble approach on the day-ahead market, emphasizing the potential cost savings and additional costs associated with different levels of forecast accuracy.