Les Cahiers du GERAD ISSN: 0711–2440

On the completion of AI-based weather models

A. Paquette-Rufiange, J. Carreau

G–2026–13

March 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

On the completion of AI-based weather models

Antonin Paquette-Rufiange [a, b]

Julie Carreau [a, b]

[a] *Département de mathématiques et de génie industriel, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

antonin.paquette-rufiange@polymtl.ca
julie.carreau@polymtl.ca

March 2026
Les Cahiers du GERAD
G–2026–13

Abstract : Predicting the state of the weather, from tracking hurricanes and thunderstorms to assessing daily temperatures, is of crucial importance. Over the past decades, these predictions have relied on solving complex sets of partial differential equations and closure models spanning multiple temporal and spatial scales. Recently, advances in machine-learning technologies have given rise to data-driven weather emulators that are able to perform skillful predictions at a lower computational cost. However, these data-driven weather models predict only a limited subset of meteorological variables. The objective of this work is to complete AI-based weather models by predicting some secondary meteorological variables of interest. We adopt a data-driven approach in which the models considered consist of neural networks, with several simplifying hypotheses regarding spatial and temporal dependence. The capability of the model to generalize in time and space is assessed using two metrics: a weighted mean squared error and the Structural Similarity Index Measure. The model errors are further investigated by computing their spatial and temporal correlations. A spectral analysis of the model's predictions is also performed. Finally, we carry out a sensitivity analysis to identify the relevant parameters of the model.

Keywords: Weather simulations; AI-based weather emulators; sensitivity analysis

Résumé : Les prévisions météorologiques constituent des outils essentiels à la prise de décision. Au cours des dernières décennies, ces prévisions reposaient sur la résolution d'ensembles complexes d'équations différentielles partielles ainsi que sur des modèles de fermeture couvrant de multiples échelles temporelles et spatiales. Récemment, les avancées en matière de technologies d'apprentissage automatique ont donné naissance à des émulateurs météorologiques fondés sur l'IA, capables de produire des prévisions performantes à un coût de calcul réduit. Cependant, ces modèles météorologiques basés sur les données ne prédisent qu'un sous-ensemble limité de variables météorologiques. L'objectif de ce travail est de compléter les modèles météorologiques basés sur l'IA afin de prédire des variables météorologiques d'intérêt actuellement manquantes. Nous adoptons une approche fondée sur les données, dans laquelle des modèles basés sur l'IA sont développés à l'aide d'hypothèses concernant les dépendances spatiales et temporelles. Les capacités de généralisation du modèle dans le temps et dans l'espace sont évaluées à l'aide de deux métriques : l'erreur quadratique moyenne pondérée et l'indice de similarité structurelle des images. Les erreurs du modèle sont ensuite étudiées plus en détail en calculant leurs corrélations spatiales et temporelles. Une analyse spectrale des prédictions est également réalisée. Finalement, nous effectuons une analyse de sensibilité afin d'identifier les paramètres pertinents du modèle.

Mots clés : Simulations météorologiques; émulateurs météorologiques IA; analyse de sensibilité

Acknowledgements: APR is grateful for the financial support of Environment and Climate Change Canada. The authors also want to thank Stéphane Beauregard, Simon-Philippe Breton and Miguel Tremblay from Environment and Climate Change Canada for their insights and discussions during the development of the research question. Computations were made on the supercomputer Narval, managed by Calcul Québec and the Digital Research Alliance of Canada (Alliance). The operation of this supercomputer is funded by the Alliance, le ministère de l'Économie, de l'Innovation et de l'Énergie du Québec (MEIE) and le Fonds de recherche du Québec (FRQ).

1 Introduction

The ability to perform accurate weather forecasts is critical to the functioning of societies. From the prediction of thunderstorms and hurricanes to the long-term assessment of hydrological capacity of dams, meteorological simulations provide crucial information to stakeholders in their decision-making process. Over the past decades, these forecasts relied on intricate numerical weather programs (NWPs) that approximate the solution of complex partial differential equations and closure models. Various hypotheses and discretization approaches are employed in order to accurately capture the different meteorological phenomena occurring at various temporal and spatial scales. Nowadays, several NWPs, such as the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Global Environmental Multiscale model (GEM) of Environment and Climate Change Canada (ECCC), can perform skillful forecasts up to several days.

In recent years, the rapid emergence of artificial intelligence methods combined with the development of extensive climate reanalysis datasets have enabled the creation of machine learning-based weather programs (MLWPs). Technological companies have showcased the possibility to perform weather forecasts by training various AI architectures on the ERA5 reanalysis dataset [9]. For example, Google developed GraphCast [10], an autoregressive model comprised of a graph neural network over the entire Earth. Later, the same company proposed a probabilistic version of GraphCast, coined GenCast [16] that relies on a similar graph architecture embedded in a diffusion model. Despite their strengths and key advantages, MLWPs still exhibit a number of notable limitations. Beyond their lack of interpretability, these models are heavily dependent on high-quality training data, struggle to accurately quantify forecast uncertainty, and frequently underperform in extreme weather scenarios. Their evaluation also tends to be incomplete, often relying on a narrow set of statistical metrics that may not fully capture operational forecast skill from a meteorological perspective. Furthermore, the range of predicted meteorological variables remains restricted; critical variables such as precipitation are either unavailable or insufficiently accurate, which significantly limits the suitability of these models for operational weather service applications [24].

In this work, we focus on this last limitation, namely how to extend the output of MLWPs to a broader set of meteorological variables. One natural approach is to reconstruct missing variables from the outputs of existing MLWPs. Hardy and Finney [8] addressed this for a single variable, 100-meter wind speed, over a restricted domain (Europe), using a U-Net model. More recent MLWPs, often referred to as foundation models, such as Aurora [3], offer a more flexible framework. These models can be fine-tuned on new data and/or new training objectives, providing a principled mechanism to introduce additional variables not originally included in the emulator's output. Exploiting this property, Lehmann et al. [12] reconstruct missing hydrological variables directly from Aurora's latent space, on the grounds that it encodes meaningful physical relationships between atmospheric states. Their proposed lightweight add-on models, trained with Aurora's weights kept frozen, are shown to achieve strong predictive performance while requiring only a fraction of the computational cost of fine-tuning Aurora directly.

Our work sits at the intersection of Hardy and Finney [8] and Lehmann et al. [12]. Specifically, we propose a lightweight add-on model that reconstructs meteorological variables not natively produced by MLWPs, taking as input the set of variables that such emulators typically output. Unlike Lehmann et al. [12], and in line with Hardy and Finney [8], we operate directly on the emulator's output variables rather than its latent space. This design choice enhances the generalizability of our approach, as it remains agnostic to the underlying emulator architecture, making it readily applicable across different MLWPs. We further consider a broader and more diverse set of target meteorological variables, spanning a range of physical properties and reconstruction challenges, thereby extending the scope of prior work. Finally, unlike methods that rely on fixed spatial grids (e.g., the U-Net model), our add-on model can reconstruct any target variable at arbitrary space-time locations, enabling flexible deployment across different domains and resolutions.

1.1 Objectives

In light of this discussion, the objective of this paper is to model the *secondary meteorological variables* that are not predicted by MLWPs. More precisely, we aim to complete the meteorological information provided by MLWPs, in order to provide more comprehensive forecast. Thus, we will work within the following methodological constraints. First, we restrict the data available to model the secondary variables to:

- The meteorological variables predicted by MLWPs, which will be coined *primary meteorological variables*;
- Time-invariant geospatial information (e.g. orography, land-sea mask);
- Spatiotemporal coordinates (e.g. latitude/longitude and time of day)

Second, we want the proposed solution to be lightweight compared to MLWPs. Indeed, the objective is to complete these emulators, not to replace them. Moreover, the end objective of this paper is not to present a complete and operational solution for the prediction of the secondary variables. Rather, this paper is devoted to exploring the various issues and challenges that arise when tackling this task. More precisely, we will study:

- The impacts of the various modeling choices and assumptions;
- The generalization capacities of the model under several quantity of interest;
- The meteorological variables that drive the prediction of the secondary variables.

1.2 Outline

Section 2 precisely describes the problem of completing the secondary variables. A minimal subset of meteorological variables predicted by MLWPs is identified as the primary variables. Next, the secondary variables considered in this work are defined, along with the normalization procedure employed. Section 3 introduces the proposed methodology to model the secondary meteorological variables. A simple model architecture is presented as well as the loss function considered. Then, the various datasets employed for the training, validation and testing are described. Section 4 discusses the various metrics employed to assess the quality of the developed model. The Power Density Spectrum (PDS) and the Structural Similarity Index (SSIM) will be described. A sensitivity analysis method will be presented, namely the Active Subspaces method. Section 5 evaluates multiple aspects of the proposed solution. The impact of the architecture and of the training datasets will be investigated. Then, the various metrics will be computed and an error analysis will be performed. The results of the sensitivity analysis are also presented. Finally, some concluding remarks are made in Section 6.

2 Problem description

In this section, we introduce the notation employed throughout this paper. Importantly, we precisely define the problem of predicting the secondary variables. To this end, we first identify the minimal set of variables predicted by MLWPs. We also describe our treatment of spatiotemporal coordinates via Fourier Feature Mapping [23]. The minimal set of meteorological variables predicted by MLWPs, together with the spatiotemporal coordinates, will be referred to as *primary variables* α. Then, we identify the relevant secondary variables, denoted by β, that we wish to model.

The surface position will be described by $x = (\theta, \psi)$, where θ is the latitude and ψ is the longitude. The temporal coordinates will be denoted by t. In this work, the pressure levels are not considered as geographical coordinates, but rather assimilated in the definition of the variables. For example, if the primary variables were to consist of the 2m temperature (t2m) and the geopotential (z) at pressure levels 500 and 600, the primary variables at position x and time t will be denoted

$$\alpha(x,t) = (\alpha_{\text{t2m}}(x,t), \alpha_{\text{z-500}}(x,t), \alpha_{\text{z-600}}(x,t))$$

2.1 Primary variables

In order to define the meteorological variables included in the primary variables α, we survey six MLWPs. More precisely, we will consider AIFS [11] by the European Centre for Medium-Range Weather Forecasts (ECMWF), GraphCast [10] and GenCast [16] from Google, Pangu [2] from Huawei, Aurora [3] from Microsoft and FourCastNet [15] from Nvidia. The various meteorological variables, as well as their respective use, are displayed in Table 1. We note that among the aforementioned MLWPs, only GenCast is a probabilistic model, compared to the other, which are deterministic. However, it is important to mention that once trained, the deterministic MLWPs can produce large ensemble-members forecasts due to their low evaluation cost.

Table 1: Meteorological variables used by MLWPs. The terms I and O indicate that the variables are employed as input and as output, respectively. The meteorological variables included in the primary variables are highlighted in gray.

Variable	AIFS	GraphCast	GenCast	Pangu	Aurora	FourCastNet
Land-sea mask	I	I	I	-	I/O	-
Orography	I	I	I	-	I/O	-
Standard deviation of sub-grid orography,	I	-	-	-	-	-
Slope of sub-scale orography orography	I	-	-	-	-	-
Soil type	-	-	-	-	I/O	-
Insolation	I	-	-	-	-	-
Latitude	I	I	I	I	I	I
Longitude	I	I	I	I	I	I
Time of day	I	I	I	-	-	-
Day of year	I	I	I	-	-	-
Top of atmosphere incident solar radiation	-	I	-	-	-	-
Surface pressure	I/O	-	-	-	-	I/O
Mean sea-level pressure	I/O	I/O	I/O	I/O	I/O	I/O
Skin temperature	I/O	-	-	-	-	-
2m temperature	I/O	I/O	I/O	I/O	I/O	I/O
2m dewpoint temperature	I/O	-	-	-	-	-
10m U wind component	I/O	I/O	I/O	I/O	I/O	I/O
10m V wind component	I/O	I/O	I/O	I/O	I/O	I/O
Total column water	I/O	-	-	-	-	-
Total precipitation	O	I/O	O	-	-	I/O
Convective precipitation	O	-	-	-	-	-
Sea surface temperature	-	-	I/O	-	-	-
Geopotential	I/O	I/O	I/O	I/O	I/O	I/O
U wind component	I/O	I/O	I/O	I/O	I/O	I/O
V wind component	I/O	I/O	I/O	I/O	I/O	I/O
Vertical wind component	I/O	I/O	I/O	-	-	-
Specific humidity	I/O	I/O	I/O	I/O	I/O	I/O
Temperature	I/O	I/O	I/O	I/O	I/O	I/O

By analyzing Table 1, we can observe that several variables are predicted by the majority of MLWPs. From these shared variables, we extract a minimal set of variables that will be incorporated in the primary variables. This minimal set is highlighted in grey in Table 1. Several of these meteorological variables are atmospheric and are thus defined at specific pressure levels. The description of the pressure levels employed by the different MLWPs is provided in Table 2.

For the atmospheric variables, an analysis of Table 2 indicated that 4 out of 6 MLWPs employed the levels used by WeatherBench [18]. We will thus consider these levels for the primary atmospheric variables. There will be a total of 6 surface primary variables plus five atmospheric primary variables at 13 pressure levels.

Alongside these meteorological variables, we enhanced the primary variables with information regarding the spatiotemporal coordinates. More precisely, we will employ the Fourier Feature Mapping to represent these additional variables [23].

Table 2: Pressure levels used by the MLWPs

MLWP	Pressure Levels
AIFS	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000
GraphCast	1, 2, 3, 5, 7, 10, 20, 30, 50, 70, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000
GenCast	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000
Pangu	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000
Aurora	50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000
FourCastNet	50, 500, 850, 1000

The temporal primary variables α_{day} and α_{hour} are obtained using the following Fourier Features Mappings

$$\alpha_{\text{day}}(x,t) = (\sin(2\pi t_d/365), \cos(2\pi t_d/365))\,, \tag{1a}$$
$$\alpha_{\text{hour}}(x,t) = (\sin(2\pi t_h/24), \cos(2\pi t_h/24))\,, \tag{1b}$$

where t_d represent the day in the year and t_h the hour (UTC) of the day.

The geographical position is also represented with the Fourier Features Mapping approach. To this end, we consider the real spherical harmonics of degree $l \geq 0$ and order $|m| \leq l$

$$Y_{l,m}(\tilde{\theta},\tilde{\psi}) = \begin{cases} (-1)^m\sqrt{2}\sqrt{\frac{(2l+1)(l-|m|)!}{4\pi(l+|m|)!}}P_l^{|m|}(\cos(\tilde{\theta}))\sin(|m|\tilde{\psi}) & \text{if } m<0 \\ \sqrt{\frac{2l+1}{4\pi}}P_l^m(\cos(\tilde{\theta})) & \text{if } m=0 \\ (-1)^m\sqrt{2}\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^m(\cos(\tilde{\theta}))\cos(m\tilde{\psi}) & \text{if } m>0 \end{cases} \tag{2}$$

where P_l^m are the Legendre polynomials. An important point is that the spherical harmonics (2) are defined with respect to polar angle $\tilde{\theta}$ (0 at the North pole and π at the South pole) and longitude $\tilde{\psi}$ ranging between $[0, 2\pi]$. The primary variable pertaining to the surface position α_x can then be defined as:

$$\alpha_x = \left(Y_{1,-1}(\tilde{\theta},\tilde{\psi}), Y_{1,0}(\tilde{\theta},\tilde{\psi}), Y_{1,1}(\tilde{\theta},\tilde{\psi})\right). \tag{3}$$

The first spherical harmonic is not taken into account since it is constant. We mention that higher frequencies can be considered in the Fourier Feature Mappings, both in the spatial and temporal domain. However, as a first approximation, we consider only one frequency/degree in each of the Fourier Feature Mappings. Hence, the number of primary variables is 78 (6 surface variables, 5 atmospheric variables on 13 pressure levels, 4 Fourier features in time and 3 Fourier features in space).

2.2 Secondary variables

Multiple meteorological variables are important for the operational activities of weather forecast agencies. Several of these variables are already predicted by some or all of the MLWPs. However, numerous meteorological variables of interest are not predicted and/or are not considered as primary variables. Moreover, the atmospheric variables predicted by MLWPs are available only for a restricted set of pressure levels.

In this work, we will consider a set of five secondary variables. We list here the variables selected alongside the rationale for such choices.

2m dewpoint temperature: This variable is a priori depends on the local meteorological state.

Total cloud cover: This variable consists in an aggregation of information occurring at multiple pressure levels.

Geopotential at pressure level 900: The pressure level 900 is not present in the primary variables. Hence, we wish to investigate the possibility of interpolating the primary meteorological variables at missing pressure levels.

Vertical velocity at pressure level 925 and 700: The vertical velocities at pressure levels 925 and 700 are considered because of their potential importance in modeling precipitation. Moreover, analyzing the same meteorological variables at two different pressure levels will allow us to investigate the effect of the pressure levels.

This set of secondary variables is far from being exhaustive. We remind the reader that the objective of this work is to explore the possibility to reconstruct secondary variables from the output of MLWPs. We do not wish here to provide a complete and operational solution, but rather to identify the challenges of such a task.

2.3 Data normalization

Due to the difference in units, scales, and marginal distributions of the primary and secondary variables, suitable normalization of the data need to be performed. The objective of this normalization is twofold: 1) It prevents poor scaling of the parameters of the models to be developed; 2) It ensures that all secondary variables are given equal relative importance.

In this work, we will normalize the primary and secondary variables so that they are of zero mean and unit variance

$$\alpha_i(x,t) \leftarrow \frac{\alpha_i(x,t) - \bar{\alpha}_i}{s(\alpha_i)}, \tag{4a}$$

$$\beta_i(x,t) \leftarrow \frac{\beta_i(x,t) - \bar{\beta}_i}{s(\beta_i)}. \tag{4b}$$

The quantities $\bar{\alpha}_i$ and $\bar{\beta}_i$ represent the empirical/sample mean of the primary and secondary variables for the dataset corresponding to the Canadian region for year 2016 (see Table 3 and Section 3.2). The empirical/sample standard deviation $s(\alpha_i)$ and $s(\beta_i)$ are also computed with respect to this specific dataset. This transformation is applied to each primary and secondary variable, except for the land-sea mask and the total cloud cover. These two variables are mapped to the interval $[-1, 1]$ with an affine transformation.

2.4 Statement of the objective

We can now state the objective of reconstructing the secondary variables. We want to design models, represented by the functions f, that take as input the primary variables and output the secondary variables. The precise description of the modeling hypotheses and assumptions will be provided in Section 3.

3 Data-driven solution

This Section describes the various hypotheses and simplifications employed for the modeling of the secondary variables. Then, the loss function used to train the models is specified. We discuss the different datasets that will be considered during the training and testing phases. We conclude this Section with a description of the training procedure.

3.1 Modeling hypotheses

In order to model the secondary variables, we suppose that the secondary variables at geographical location x and time t are determined solely by the value of the primary variable $\alpha(x,t)$, i.e.,

$$\beta(x,t) = f(\alpha(x,t)). \tag{5}$$

Hence, the space of the primary variables is a subset of $\mathbb{R}^{78}$ and the space of secondary variables is a subset of $\mathbb{R}^5$. This hypothesis leads to a *nodal model*, which is a significant departure from the hypotheses of the MLWPs mentioned in Section 2.

Two functions f are considered in this work. The first model is a simple affine model

$$f_{\text{lin}}(\alpha(x,t); W_{\text{lin}}, b_{\text{lin}}) = W_{\text{lin}}\alpha(x,t) + b_{\text{lin}}, \tag{6}$$

where $W \in \mathbb{R}^{5\times 78}$ are the weights and $b_{\text{lin}} \in \mathbb{R}^5$ are the offsets. This model will serve as a base model for comparing the performance of the subsequent model. This second model is a fully-connected neural network of n hidden layers with a residual connection. More precisely, the secondary variables are predicted according to

$$f_{\text{nn}}(\alpha(x,t); W, b) = T_n \circ (T_{n-1} \circ \cdots \circ T_1 + T_{\text{res}})(\alpha(x,t)), \tag{7a}$$

$$\text{with } T_i(y) = \sigma(W_i y + b_i) \qquad \text{for } i = 1, \ldots, n-1, \tag{7b}$$

$$\text{and } T_n(y) = \gamma(W_n y + b_n), \tag{7c}$$

$$\text{and } T_{\text{res}}(y) = W_{\text{res}} y + b_{\text{res}} \tag{7d}$$

To simplify the notation, we will note $W = (W_1, \ldots, W_n, W_{\text{res}})$ and $b = (b_1, \ldots, b_n, b_{\text{res}})$. The activation function σ is the hyperbolic tangent (apply component-wise). The activation function γ is the identity, but the component pertaining to the total cloud cover is constrained to the interval $[-1, 1]$ (sometimes referred to as the *hard* hyperbolic tangent). The hidden layers are of width h, so that $W_1 \in \mathbb{R}^{h\times 78}$, $W_i \in \mathbb{R}^{h\times h}$ for $i = 2, \ldots, n-1$, $W_n \in \mathbb{R}^{5\times h}$, and $W_{\text{res}} \in \mathbb{R}^{5\times 78}$. The biases are defined accordingly. We treat the number of hidden layers n and their width h as hyperparameters to be tuned.

3.2 Training, validation and testing datasets

To train and analyze the models, we employ the reanalysis datasets ERA5 from the European Centre for Medium-Range Weather Forecasts [9]. This choice is justified by two reasons. First, this dataset provides detailed and high-quality weather states of the entire Earth from January 1940 to the present. Indeed, the reconstructed weather state are extensive, consisting of several surface, atmospheric and accumulated variables with a resolution of 0.25 degrees. Second, this dataset is already employed to train the various MLWPs.

In this work, we consider the single high-resolution reanalysis, rather than the ensemble reanalysis. An important point to mention is that we will consider the ERA5 datasets as the ground truth, even though it contains inevitable errors arising from the modeling and data assimilation.

The data are separated into three distinct datasets, each fulfilling a specific objective:

Training set S_{train}: Data employed to train the models (6) and (7);

Validation set S_{val}: Data employed to terminate the training of the model (early stopping) and to compare the trained models;

Test set S_{test}: Data employed to assess the accuracy of the final model.

Employing the whole ERA5 reanalysis dataset for the training of the model is very computationally intensive, since the entire dataset takes up several petabytes of memory. In accordance with the objective to design a lightweight solution for the prediction of the secondary variables, we decide to adopt a parsimonious approach regarding the selection of the training datasets. Firstly, we extract ERA5 data at time 00, 06, 12 and 18 UTC, which is the usual approach to train the MLWPs. Secondly, we take inspiration from the work of Subich [22], where the GraphCast architecture is trained only on data from the Global Deterministic Prediction System spanning July 2019 to December 2021. Hence, we will restrict the temporal domain to a few years. Finally, given the nodal nature of the models (5), we can further subset the ERA5 reanalysis by considering specific geographical regions for the definition of the datasets. Table 3 provides a description of the subsets of ERA5 considered.

Table 3: Definition of the subsets of ERA5 with their purpose

	Regions		
Years	Canada	Brazil	California
2016	Training Validation	Training Validation	Test
2017	Training Validation	Training Validation	-
2018	Test	-	Test

Several rationales explain the choices made to determine these subsets. The research question investigated in this manuscript was developed in partnership with Environment and Climate Change Canada, hence explaining the consideration of the geographical region of Canada in both training/validation and test datasets. The starting year of 2016 is chosen somewhat arbitrary. We note that since 2016, the Integrated Forecasting System (IFS) of ECMWF operated at a finer resolution of 0.1 degree [20]. We remind the reader that the ERA5 datasets utilized in this paper are on an equiangular grid of 0.25 degree. Hence, if it becomes necessary to consider a finer grid, as with Bodnar et al. [3, Section 6], suitable datasets are available from 2016 to the present.

We also consider the year 2017 and another geographical region, Brazil, as training data. The objective behind expanding the training data, both in time and space, is to evaluate the impact of the choice of the training dataset. This procedure will be described in greater detail in Section 5.2. To assess the generalization performance of the models, we first consider the same regions employed for the training/validation, but for year 2018. This way, we test the ability of the model to generalize in time. We also wish to test the capacity of the model to generalize in space by considering the region of California. The ability to generalize both in time and space will also be investigated.

Due to the hypothesis (5), each spatiotemporal node constitutes data on which the models may be trained or evaluated. By considering the hours 00, 06, 12 and 18 UTC, the yearly Canada and Brazil datasets are comprised of approximately 85 and 50 million spatiotemporal nodes, respectively. In order to train the model, we randomly assign 80% of these spatiotemporal nodes to the training datasets, while the remaining 20% are attributed to the validation datasets.

3.3 Training procedure

To compute the optimal value of the weights and biases of models (6) and (7), we solve the following empirical risk minimization problem

$$(W^*, b^*) \in \operatorname*{argmin}_{W,b} R_{S_{\text{train}}}(W, b; f), \tag{8}$$

where $R_{S_{\text{train}}}$ is the loss function evaluated over the training dataset S_{train}. For the loss function, we consider a weighted mean squared error (WMSE)

$$R_S(W, b; f) = \frac{1}{5|S|} \sum_{(\alpha,\beta)\in S} w(\alpha) \|f(\alpha; W, b) - \beta\|_2^2, \tag{9}$$

where $|S|$ is the size of the dataset S (the number of spatiotemporal nodes) and w is the weighting function. Due to the equiangular nature of the ERA5 dataset, concentration of nodes appears at the poles. The weighting function takes into account this phenomenon by pondering according to $w(\alpha) = \cos(\theta)$, where θ is the latitude associated with the primary variable α (this weight can be recovered from the corresponding spherical harmonic $Y_{1,0}$). The factor $1/5$ accounts for the fact that five secondary variables are considered. Due to the normalization process described in Section 2.3, each type of secondary variable is of the same order of magnitude. Hence, we do not need to weight each component of the error $f(\alpha; W, b) - \beta$.

To find a (local) minimum of problem (8), we employ the algorithm L-BFGS with a line-search based on the Strong Wolfe conditions [13]. A relative early stopping criterion is applied to the validation loss function to terminate the training process. More precisely, at the k-th iteration of L-BFGS, we check if the following criterion is met

$$\frac{R_{S_{\text{val}}}(W^k, b^k; f) - R_{S_{\text{val}}}(\tilde{W}, \tilde{b}; f)}{R_{S_{\text{val}}}(\tilde{W}, \tilde{b}; f)} \leq 0.001, \tag{10}$$

where (W^k, b^k) represent the weights and biases at iteration k and $(\tilde{W}, \tilde{b})$ represents the best solution up to this k-th iteration. If the criterion is met, then $(\tilde{W}, \tilde{b}) \leftarrow (W^k, b^k)$ and the iteration proceed. If no significant relative improvement is detected over eight epochs, then the training terminates. The training is performed on a single NVIDIA A100SXM4 GPU with allocated time in the range of a few hours up to 36 hours.

4 Errors and sensitivity analysis

We adhere to the point of view that any model is merely an approximation of the true underlying physical phenomenon. Indeed, the modeling act in itself implies simplifications and hypotheses to render reality comprehensible. Inevitably, there will be errors in the model predictions, and it is the whole objective of the modeling process to reduce these errors as much as possible. These errors stem from various sources and for the problem at hand, we mention the following ones:

1. The choice of the primary variables α described in Section 2.1 may be too restrictive. The secondary variables β may depend upon additional variables that are not taken into account.
2. The hypotheses regarding the structure of the models introduced in Section 3 may be too simplistic. Moreover, the choice of the hyperparameters pertaining to the width h and number of layers m of the neural network may not be optimal.
3. The training datasets employed to train the models described in Section 3.2 may not encompasses all the possible meteorological states.
4. The weights and biases obtained with the optimization algorithm L-BFGS may not represent the global solution of Problem (8).

The first and second types of errors are often coined approximation errors, the third type of errors are the generalization errors and the fourth type are the optimization errors [1, 17, 21]. The approximation errors will be investigated to a limited extent. Indeed, the choice of primary variables α is considered fixed, as described in Section 2.1. However, we will investigate, in Section 5.1, the influence of the hyperparameters pertaining to the neural network model. The impact of the choice of training datasets will be investigated in Section 5.2, providing some insights into the generalization errors. Finally, the optimization errors will not be investigated in this work. However, the aggregation of these errors will be analyzed on the aforementioned test datasets.

4.1 Additional metrics of interest

The previous discussion pertains to the error analysis of the loss function R. However, the Earth weather system is so complex that it will be highly reductive to limit our study to one type of metric. To expand the analysis of our model predicting the secondary variables, we consider two additional quantities to assess the qualities and limitations of the predictions performed.

4.1.1 Structural similarity index measure

The first metric investigates the ability of nodal models to recover spatial structures. To this end, we will evaluate the Structural Similarity Index Measure (SSIM) [28]. The SSIM measures the similarity

between two images, which in our study will represent the model's predictions and the ERA5 ground truth at a given timestep. The SSIM is composed of three parts, namely:

- **Luminance:** Comparison between the pixel's mean value of the two images
- **Contrast:** Comparison between the pixel's variance of the two images
- **Structure:** Correlation coefficient between the two images

We refer the interested readers to [5, 28] for a more in-depth discussion of this index. To simplify the notation, we will note by $B_{k,t} = \beta_k(\cdot, t)$ and by $B_{f,k,t} = f_k(\alpha(\cdot, t))$ the ERA5 and model's prediction images respectively. These images are obtained at time t and for the k-th secondary variable. In an Euclidean setting, the SSIM is generally computed as

$$\mathrm{SSIM}(B_{k,t}, B_{f,k,t}) = \frac{2\overline{B_{k,t}}\,\overline{B_{f,k,t}} + c_1}{\overline{B_{k,t}}^2 + \overline{B_{f,k,t}}^2 + c_1} \frac{s_{B_{k,t},B_{f,k,t}} + c_2}{s^2_{B_{k,t}} + s^2_{B_{f,k,t}} + c_2} \tag{11}$$

where $\overline{B_{k,t}}, \overline{B_{f,k,t}}$ respectively represent the mean of the ERA5 ground truth and of the model's prediction, $s_{B_{k,t},B_{f,k,t}}$ is the covariance, $s^2_{B_{f,k,t}}, s^2_{B_{k,t}}$ are the variance, and c_1 and c_2 are small constant. An SSIM value of one indicates that the images are identical, whereas an SSIM value near zero indicates that the images are very different. Usually, the SSIM is not computed on the whole image at once. Each pixel is attributed an SSIM score by computing (11) on a local patch, where the various means, variances and covariances are computed with some weighting function. A global SSIM value is then computed by averaging all the local SSIM values. An additional averaging is performed along the time dimension t.

Given that the model's prediction and the ERA5 datasets are located on the sphere on a regular latitude/longitude, the aforementioned procedure to compute the global SSIM value needs to be modified. We adopt the procedure taken by Chen et al. [19]. The local SSIM values are computed in the Euclidean setting, because locally, the planar approximation is fairly valid. However, we aggregate the local SSIM by pondering their contribution with respect to their latitude (identical to the weighting of the loss function (9)).

4.1.2 Power density spectrum

A notable challenge encountered in weather modeling resides in the presence of several physical phenomena occurring at various spatial and temporal scales. In order to assess the performance of the developed models across the different spatial scales, we will compute the power density spectrum (PDS) of the model predictions as well as that of the ERA5 ground truth.

Given that the data reside on the globe and span relatively large regions, we need to consider the spectral decomposition on the sphere. Indeed, if we were to perform the spectral decomposition on the plane (on the latitude/longitude grid), significant errors will be introduced. We assume that all the secondary variables considered in this work are square-integrable on the sphere, hence the spherical harmonic decomposition reads

$$\beta(x, t) = \sum_{l=0}^{320} \sum_{m=-l}^{m=l} \hat{\beta}_{l,m}(t) Y_{l,m}(x), \tag{12}$$

where $\hat{\beta}_{l,m}$ are the coefficients of the spherical decomposition for the spherical harmonics $Y_{l,m}$ defined at (2). The truncation $l = 320$ is due to the angular resolution of 0.25 degree of the latitude/longitude grid. Once the spectral decomposition is computed, the power P associated to each degree l at time t is

$$P(l, t) = \sum_{m=-l}^{m=l} \hat{\beta}^2_{l,m}(t). \tag{13}$$

The regions considered for testing the model do not span the whole globe, hence a localized spectral analysis must be performed. In order to do so, we employ a multitaper analysis [6, 26, 27]. In a nutshell, the multitaper method consists in multiplying the signal of interest by tapers (window functions) that are bandlimited and whose energy resides primary in the region of interest. The spectral decomposition of the product of the tapers and the signal is performed and then averaged. We refer the interested readers to the aforementioned articles for in-depth discussion of the multitaper approach in the spherical domain. In this work, we employ the library SHTools to perform such analysis [25].

4.2 Sensitivity analysis

The numerous primary variables α may not all be relevant for predicting the secondary variables β. To identify the important primary variables with respect to predicting the secondary variables, we will employ a sensitivity analysis method. More precisely, we will employ the Active Subspaces method [4].

The Active Subspace computes an *influence matrix* $M_{f,k}$ for the k-th secondary variable of model f according to

$$M_{f,k} = \mathbb{E}(\nabla f_k(\alpha) \nabla f_k(\alpha)^T) = \mathrm{Cov}(\nabla f_k(\alpha)) + \mathbb{E}(\nabla f_k(\alpha))\ \mathbb{E}(\nabla f_k(\alpha))^T \tag{14}$$

The gradients are computed with respect to the primary variables α. The expectation is performed according to the distribution of the primary variables. This influence matrix $M_{f,k}$ encodes how the primary parameters α influence the surface of response of the k-th secondary variable of model f. Indeed, the term $\mathrm{Cov}(\nabla f_k(\alpha))$ will capture if the derivatives vary considerably, while the term $\mathbb{E}(\nabla f_k(\alpha))$ indicates, in average, the magnitude of the derivatives.

By performing an eigendecomposition of $M_{f,k}$ we can extract the relevant linear structures

$$M_{f,k} = V \Lambda V^T. \tag{15}$$

The eigenvector associated with the largest eigenvalue indicates which directions in the space of primary variables influence the model f the most. A way to aggregate this information for each primary variables is to compute sensitivity indices

$$SI_i = \sum_j \lambda_j V_{i,j}^2. \tag{16}$$

Normalized sensitivity indices can be computed such that their sum equals one. By ranking these sensitivity indices, we can identify which primary variables influence the response of the model as well as the primary variable that do not influence the prediction of the secondary variables.

5 Results

In the following, we first investigate the impact of the neural network architecture as well as the impact of the training datasets. From these results, a final model predicting the secondary variables is selected. The performance of the chosen model will be assessed under the metrics presented in Section 4.1 evaluated for the various test sets of Table 3. An analysis of the model's errors will also be presented. Finally, the sensitivity analysis will be conducted on the selected model to identify the relevant primary variables employed for predicting the secondary variables.

5.1 Impact of the model architecture

We first investigate the impact of the network architecture upon the performance of the model. The objective is to seek the architecture that allows the minimization of the approximation errors, while

not overfitting the training dataset. This problem is often referred the bias-variance trade-off. To avoid considering a model that overfit the training data, we evaluate the empirical risk (9) on the validation dataset. Table 4 presents, for several width h and number of layers n, the values of the empirical risk on the validation dataset of Canada 2016.

Table 4: Loss function on the validation dataset Canada 2016 for different neural networks architectures

	Layers				
Width	2	4	6	8	10
50	0.153	0.131	0.129	0.130	0.134
100	0.145	0.121	0.119	0.121	0.124
150	0.141	0.115	0.114	0.116	0.120
200	0.138	0.112	0.110	0.113	0.115
250	0.136	0.108	0.107	0.111	0.114
300	0.135	0.106	0.106	0.109	0.114*
350	0.134	0.103	0.104	0.112*	0.114*
400	0.133	0.105	0.105	0.107	-

From Table 4, we can observe that validation losses vary significantly with respect to the neural network hyperparameters. This variation is rather smooth, demonstrating a stable behavior with respect to these hyperparameters. We can identify a (local) minimum at width $h = 350$ and number of layers $n = 4$. We can also notice directly the bias-variance trade-off, where too many parameters in the neural network may lead to overfitting the training dataset. The values with an asterisk indicate that with the allocated time, the early-stopping criteria was not met.

The production of the results presented in Table 4 are computationally possible given the simplicity of the model as well as the relative small scale of the training dataset. Given these results, we think that the impact of the neural network architecture ought to be further investigated for the MLWPs presented in Section 6. We are aware that the size of these models renders this type of analysis rather complex, but it may be that more parsimonious models present similar or better performance.

We mention that one can envision employing more sophisticated neural network architectures or consider more complex modeling assumptions compared to (5) to improve the validation loss. However, we reiterate that the objective of this paper is to investigate the completion of secondary variables via relatively simple and lightweight solutions. This is the reason why we limit our investigation to the effect of the width and number of layers of model (7).

5.2 Impact of the training dataset

We now investigate the impact of the training dataset on the performance of the model. To this end, we will train the model using several training datasets. The idea is to incorporate additional information, both relative to the spatial and temporal domain, in the training process. This process is akin to the one employed by Oden and Prudhomme [14] in the context of model calibration. On one hand, if the quality of the empirical risk on the validation dataset varies significantly by considering expanding training datasets, this indicates the presence of non-negligible generalization error. On the other hand, if the quality of the models does not vary while trained on expanding training datasets, this may suggest small generalization error.

Table 5 illustrates the various loss values according to the training datasets employed (as described in Table 3) for both linear and neural network models introduced in Section 3.1. For example, the neural network model trained with the Canadian and Brazilian region on year 2016 possesses a validation loss of 0.149. We mention that the same normalization process, described in Section 2.3, is adopted whatever the training dataset considered.

We remind the reader that the linear model is employed as a baseline. Given the results displayed in Table 5, we can observe that the neural network models outperformed the baseline models for all four

Table 5: Normalized training and validation loss for the linear and neural network models (NN) trained on different datasets

	Canada 2016		Canada 2016-2017		Canada/Brazil 2016		Canada/Brazil 2016-2017	
	Linear	**NN**	**Linear**	**NN**	**Linear**	**NN**	**Linear**	**NN**
Training Loss	0.205	0.100	0.206	0.103	0.277	0.147	0.281	0.150
Validation Loss	0.205	0.103	0.208	0.105	0.277	0.149	0.282	0.152

training datasets. Moreover, we note that the training and validation loss increase as we expand the training datasets. This result is expected, since the models need to fit/account for more data. However, this increase is relatively large when the training dataset is expanded in space. This indicates that the behavior of the given secondary variables β vary significantly according to the geographical extent. On the other hand, the increase in the training and validation loss is relatively small when the datasets are expanded in time. Although this analysis is limited in terms of the number of training datasets, it suggests that future models should be trained on very large geographical extents (up to the whole globe), rather than for very large temporal extents.

From these results, we retain for further analysis the neural network model trained on the Canadian and Brazilian regions for the years 2016-2017. It is important to mention that the retained architecture of four layers of size 350 may not be optimal when considering this particular training dataset. We can repeat the same analysis presented in Section 5.1 to identify the (new) optimal architecture. However, the objective of this work is not to present the best operational solution, so we will keep the trained model without seeking to re-optimize the architecture hyperparameters.

5.3 Error analysis

We now analyze in greater detail the errors made by the adopted model. First, we expand the loss function (9) according to the loss associated with each secondary variable as displayed in Table 6. The total loss is the mean of the loss associated to each secondary variable (the factor five in (9)). We also compute the SSIM values associated to each secondary variable.

Table 6: Normalized loss function (WMSE) and SSIM on the testing datasets

	Canada 2018		California 2016		California 2018	
	WMSE	**SSIM**	**WMSE**	**SSIM**	**WMSE**	**SSIM**
Dew-point Temperature	1.41e-3	0.9833	1.48e-2	0.9373	1.46e-2	0.9338
Total Cloud Cover	6.08e-2	0.5246	0.144	0.4493	0.150	0.4502
Geopotential Level 900	5.44e-6	0.9999	1.61e-5	0.9995	1.75e-5	0.9994
Vertical Velocity Level 925	0.248	0.7269	0.816	0.5563	0.779	0.5416
Vertical Velocity Level 700	0.290	0.8245	1.36	0.6304	1.29	0.5723
Total	0.120	-	0.467	-	0.447	-

From the results displayed in Table 6, we clearly observe that some secondary variables, such as the geopotential at level 900 and the dew-point temperature, are predicted relatively well by the model. Indeed, the loss values are relatively small and the SSIM values are close to one. The ability of the model to predict the dew-point temperature is expected, since it is essentially a function of the local pressure, temperature and humidity. For the geopotential at level 900, we can speculate that the model predictions consist in local interpolation of the geopotential at the levels included in the primary variables. The loss and SSIM values are more degraded when the generalization is performed in space, that is, for the Californian region. However, the loss values remain small and the SSIM values stay close to one, indicating good generalization capabilities of the model for these secondary variables.

For the total cloud cover, we observe a clear degradation in the SSIM values. Because SSIM computes the similarity between the model predictions and that of ERA5 ground truth at each timestep, we can assess that there are significant mismatches in the spatial structures. Interestingly, for the secondary variables of vertical velocities, the loss function is greater, while the SSIM is slightly better compared to the total cloud cover. Finally, we note that the prediction quality degrades markedly when the generalization is performed in space.

The results of Table 6 represent the aggregation of the loss values and the SSIM values. In order to better visualize the errors committed, we plot in Figure 1 the predictions and the errors produced by the model for the Californian region at 2018-10-28 00 UTC. The comparison between the ERA5 ground truth and the model predictions at this particular time step sheds some light on the previous discussion. Overall, the predictions of the dew-point temperature generate small errors, but we note patches where the errors can become quite large (10 degrees Kelvin). The model appears to introduce significant smoothing in the prediction of the total cloud cover, somewhat struggling to capture the quasi-binary nature of this secondary variable. This explains the lower SSIM values observed for the total cloud cover, since the sharp boundaries of clouds are not predicted.

The geopotential at level 900 closely resembles that of the ground truth, as hinted by the loss and SSIM values. For the vertical velocities, we also observe significant smoothing in the model predictions, as the high-frequencies of the ground truth are absent from the model predictions. Interestingly, for all secondary variables, we observe spatial structure in the errors, hinting that the modeling assumption (5) is not adequate. To perform a more detailed analysis of these spatial correlations, we compute the Spearman correlation between a reference time series at a given geographical location and the time series of another geographical point. Figure 2 presents the maps of these pairwise Spearman correlations for two choices of reference geographical points.

By analyzing Figure 2, we observe the presence of strong Spearman correlation for all secondary variables in the vicinity of the first reference point. Indeed, this reference point $x = (-120, 37)$ is located in California's Central Valley, and we can observe high correlation along the valley for the dew-point temperature, the geopotential, and vertical velocities of interest. Moreover, we note oscillations in the Spearman correlation for the geopotential and the vertical velocities that spread far from this point of interest. For the total cloud cover, the correlations are more isotropic around the point of interest. With the second reference point $x = (-116, 40)$, we can also observe strong spatial correlation in the neighbouring regions.

We now perform a similar analysis to highlight the correlation in time. More precisely, we choose a reference map at a specific time and we compute the Spearman correlation with another map representing a different time step. Figure 3 presents the time series of these pairwise Spearman correlations for two choices of reference times.

The results presented in Figure 3 do not show strong Spearman correlation with respect to the time coordinates. Contrary to the spatial correlation, there are no strong correlations in the vicinity of the reference times. A possible explanation may be that the time step of 6 hours is rather large. An analysis of these time-series may reveal some important frequencies, but since the overall Spearman correlations are relatively low, we do not perform such analysis.

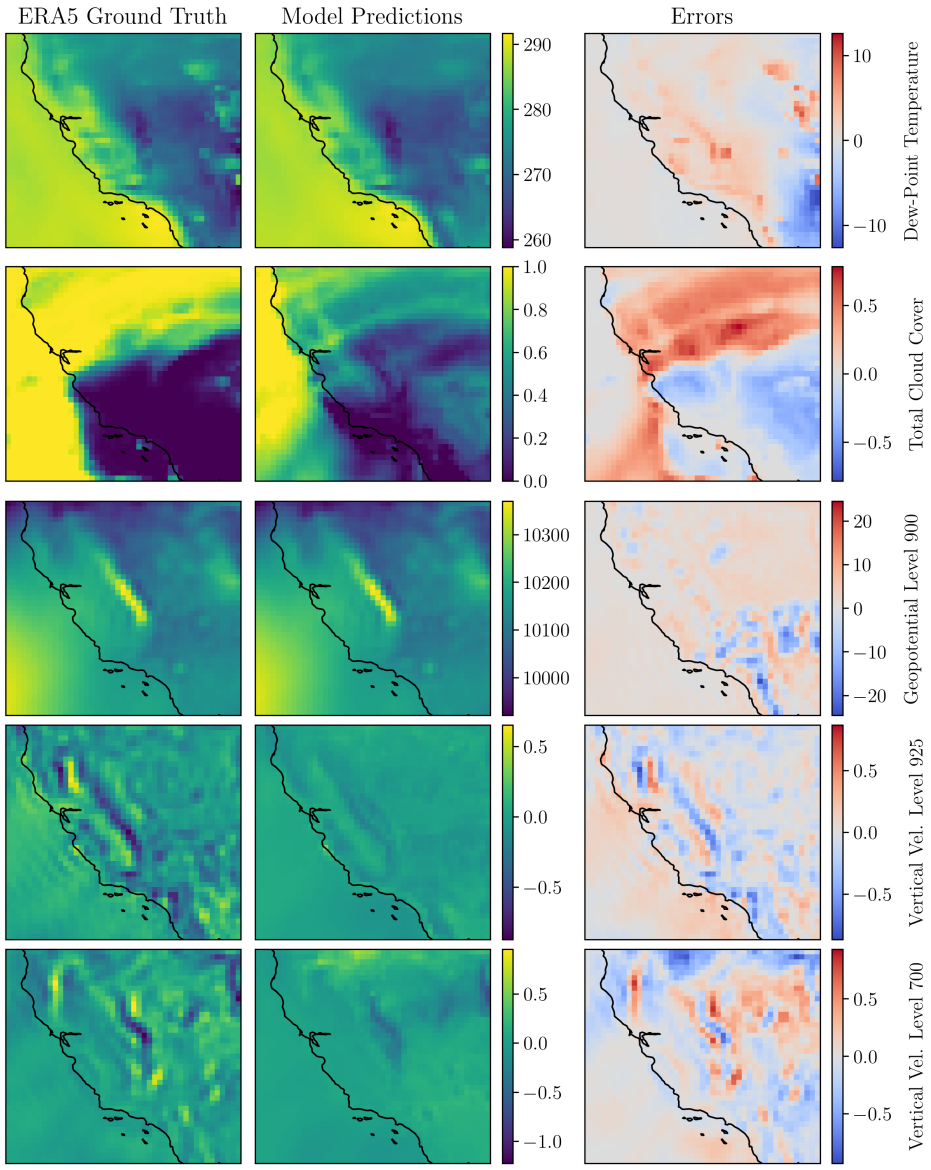

Figure 1: ERA5 ground truth (left), model predictions (center) and the corresponding errors (right) for the Californian region at 2018-10-28 00 UTC. From the top row to the bottom row, the secondary variables are the 2m dewpoint temperature, the total cloud cover, the geopotential at level 900, the vertical velocity at level 925, and the vertical velocity at pressure level 700.

5.4 Power density spectrum

We now analyze how the model performs when its power density spectrum is compared to that of the ERA5 ground truth. The Californian region is of relatively small size compared to the whole globe, so window functions of high bandwidth should be considered to ensure that their energy remains localized. Because of that, we decide to compute the power density spectrum for the Canadian region for the year 2018 only. To perform this analysis, we consider bandlimited window functions of maximal degree of 29. With a concentration factor of 0.99, we obtain 15 tapers for the Canadian region (whereas no taper possesses 99% of its energy for the Californian region). We compute the power density spectra for four different time steps throughout the year 2018. The results are displayed in Figure 4.

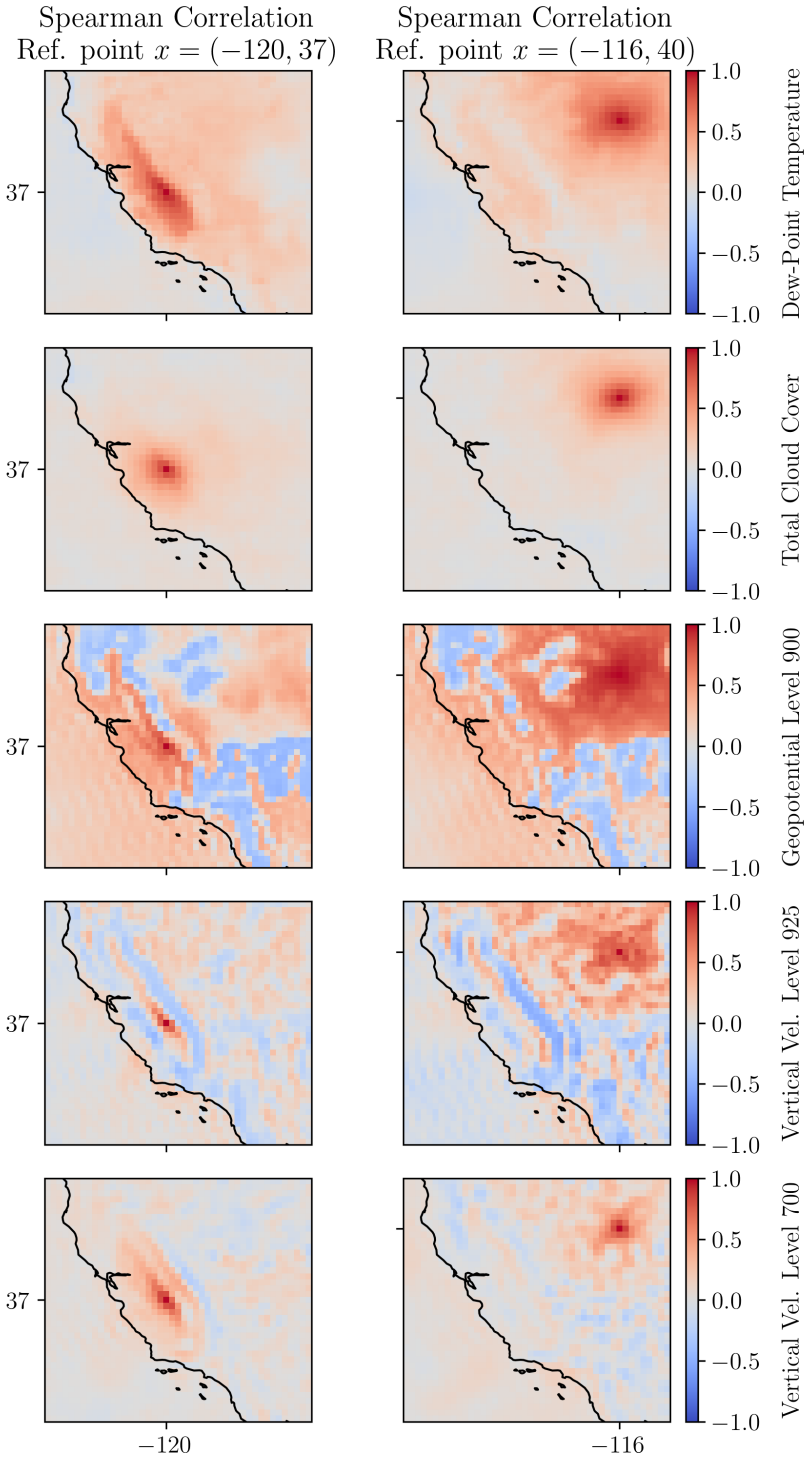

Figure 2: Spearman correlation map of the secondary variables for the reference point $x = (-120, 37)$ (left column) and for the reference point $x = (-116, 40)$ (right column).

First, we can clearly observe the energy injected by the window functions around degree 29. It is important to mention that the same taper functions are employed for the ERA5 ground truth and for the model predictions. Hence, the two types of curves can be directly compared to each other, since the influence of the taper functions is identical. For the dew-point temperature and the geopotential at level

900, we note that the power density spectra are nearly superposed. These results echo the previous findings in that these two secondary variables possess relatively low errors and good SSIM values. Regarding the total cloud cover, we can observe the smoothing in the model predictions because the high frequencies are dampened compared to the ERA5 ground truth. Moreover, this damping increases with the degree of the spherical harmonics. For the vertical velocities, we observe again the smoothing of the predictions compared to ERA5. However, we also note non-negligible mismatch in the power density spectra at low degree, indicating that even the large scale patterns of the vertical velocities are not exactly captured by the model. Again, these results are coherent with the previous ones indicating the poor performance of the model for these types of secondary variables.

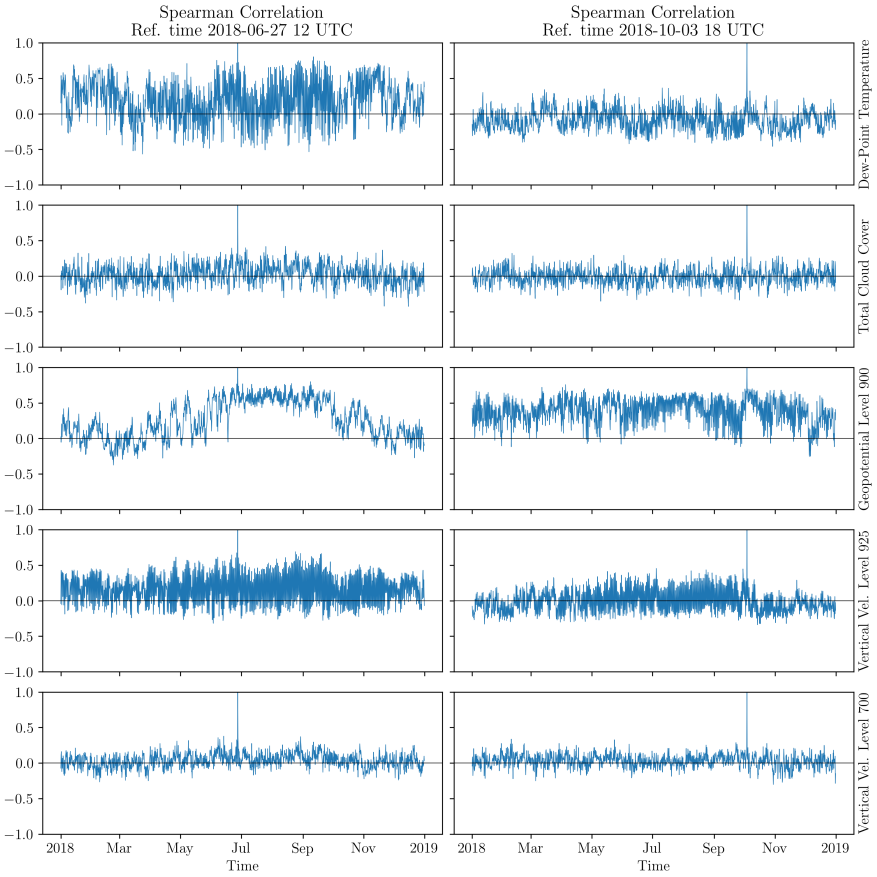

Figure 3: Spearman correlation time series of the secondary variables for the reference time 2018-06-27 12 UTC (left column) and for the reference time 2018-10-03 18 UTC (right column).

5.5 Relevant primary variables

To assess which primary variables influence the prediction of the secondary variables, we compute the normalized sensitivity indices described in Section 4.2. Figure 5 presents the relative contribution of each primary variable whose individual normalized SI is greater than 4%. The remaining primary variables are aggregated in the category *Others*.

For the dew-point temperature, the most influential primary variables are the temperature at 2 meters (t2m) and the specific humidity at level 1000 (q 1000). This result is coherent with the underlying physics implemented in ERA5 to simulate the dew-point temperature. We also note that the mean sea-level pressure (msl) and the geopotential at level 1000 (z 1000) play a non-negligible role in predicting the dew-point temperature, which is again consistent with the physics [7].

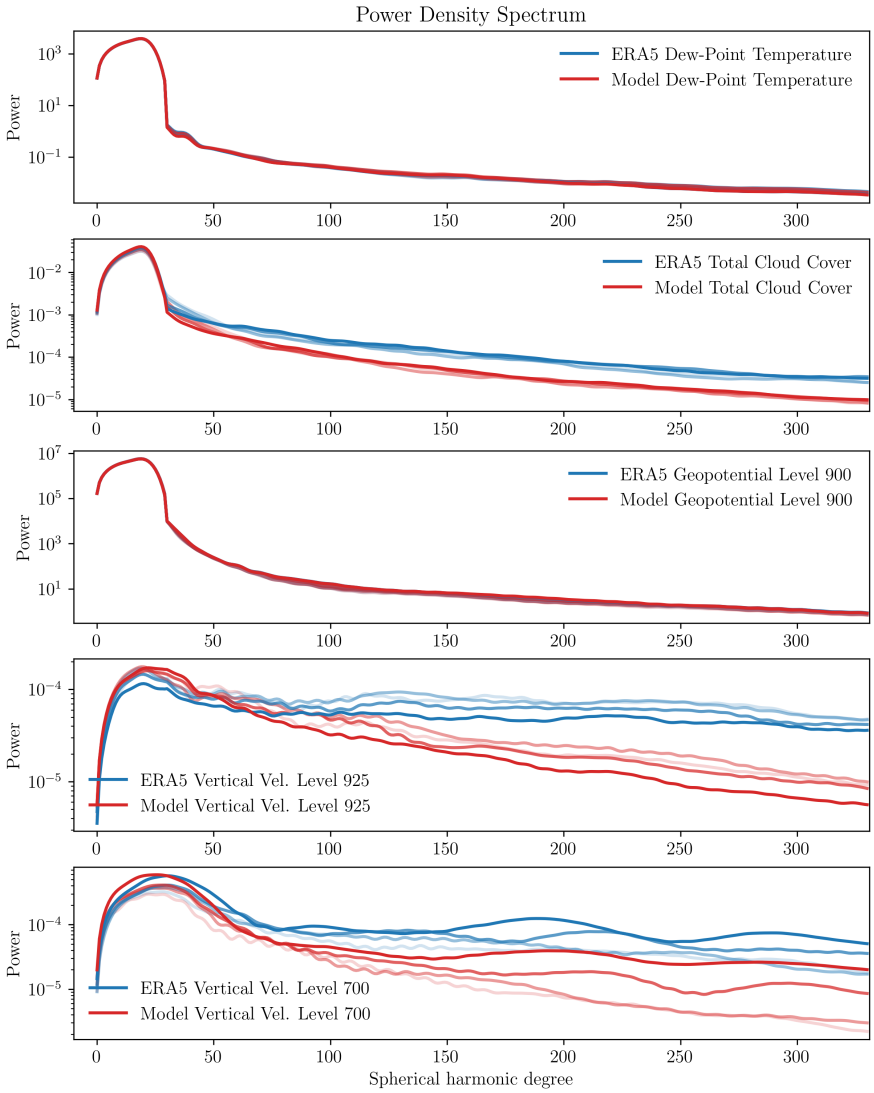

Figure 4: Power density spectra for the Canadian region at four different time steps for year 2018. The power density spectra of the ERA5 ground truth are in blue, whereas the power density spectra of the model predictions are in red.

The sensitivity analysis of the total cloud cover reveals that the mean sea-level pressure in addition to the geopotential at various levels are important primary variables. This observation is to be expected, since the total cloud cover represents the aggregation of the clouds at various levels. We also note that the cumulative contribution of the *Others* primary variables accounts for nearly one third of the SI. This result is coherent with the fact that the underlying physics describing the clouds involves multiple phenomena [7].

The repartition of the SI for the geopotential at level 900 indicates that the prediction of this particular secondary variable is an interpolation of the geopotential included in the primary variables. Indeed, the geopotential at level 925 and 850 (z 925 and z 850 respectively) contribute to almost 99% of the total contribution.

For the vertical velocities, we note that the important primary variables are the mean-sea level pressure and the geopotential at various levels. This result is again coherent with the underlying physics encoded in the momentum equations [7]. We also note that the geographical location, encoded

in the Spherical Fourier coefficients $Y_{1,-1}$ and $Y_{1,1}$, plays a non-negligible role in predicting the vertical velocities.

From this sensitivity analysis, we can observe that only an handful of the 78 primary variables influence significantly the secondary variables. What is noteworthy is that the primary variables representing the u/v components of the wind at all levels influence very marginally the secondary variables. Moreover, the atmospheric primary variables located at low pressure levels (high altitude) do not influence much the secondary variables. It may be that these primary variables could be discarded from the model without affecting its performance. In turn, it may indicate that the MLWPs predict several primary variables that are either redundant and/or not relevant for the prediction of secondary variables.

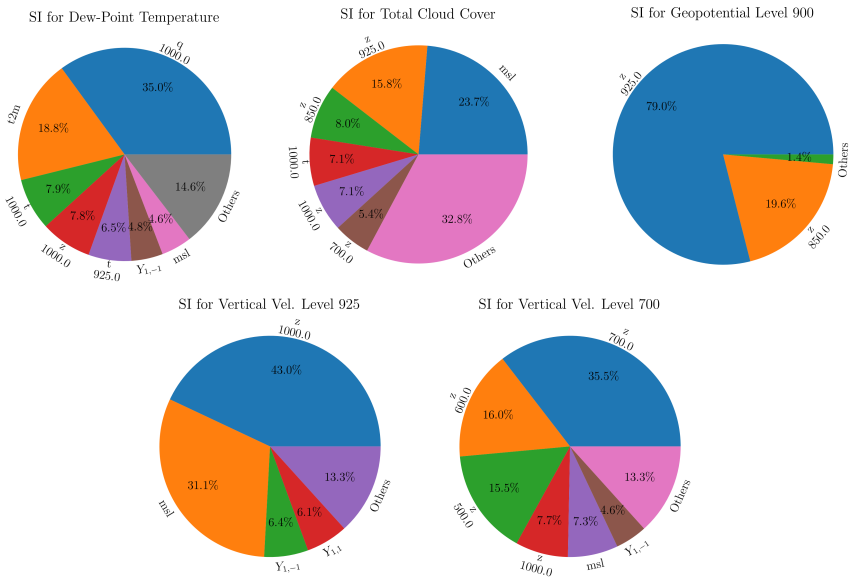

Figure 5: Sensitivity indices for the secondary variables. To improve readability, all the primary variables whose sensitivity indices contribute for less than 4% are aggregated in Others.

6 Conclusion

In this work, we have explored the possibility of recovering secondary meteorological variables not predicted by MLWPs. First, we have identified a restricted subset of meteorological variables predicted by various MLWPs that serves as the primary variables. Then we have posed a central hypothesis: The models are nodal, i.e. the secondary variables located at a specific spatiotemporal node are predicted using solely the primary variables at the same spatiotemporal node. These models have been trained on the ERA5 reanalysis datasets using a weighted mean-squared loss function.

The influence of the neural architecture and of the training dataset has been investigated. According to the training and validation loss functionals, the neural network architecture outperformed the simple linear model. Moreover, we have illustrated that the choice of the training datasets may have a significant impact upon the models. Indeed, expanding the training dataset in the spatial dimension leads to a greater increase of the loss function compared to expanding the dataset in the temporal dimension. From these results, a final nodal model has been selected for testing. The generalization capabilities of this model have been investigated with respect to the loss function and the SSIM. Some secondary variables, such as the dew-point temperature and the geopotential at level 900 can be relatively well recovered by the proposed model. However, the total cloud cover and the vertical velocities present large errors. A closer analysis reveals that the nodal model produces errors that are

correlated in space. Moreover, the comparison of the PDS of the nodal model and of the ground truth exposes significant smoothing for the prediction of the total cloud cover and the vertical velocities. Finally, a sensitivity analysis of the nodal model has been performed. For each secondary variable, the primary variables identified as relevant are in accordance with the expected physics.

These results shed insightful light on the limitations of the nodal model. Apart from the geopotential at level 900, the nodal hypothesis appears too restrictive. Indeed, the variations of training and validation losses when expanding training datasets in space, the strong spatial correlation of the errors, and the mismatch of the PDS all point toward the necessity of considering global models in space. A future improvement could be to consider a model in the spectral domain. Since the data lie on the sphere, an idea could be to employ spherical harmonics to represent both the primary and secondary variables. It should also be interesting to investigate the application of *Physics-Informed Neural Networks* (PINNs) to add physical constraints to the secondary variables. As a final note, we think that performing a sensitivity analysis on the MLWPs should provide some valuable information regarding their structures. Indeed, it may be that the MLWPs invest too much in predicting meteorological variables that are either redundant/highly correlated or not relevant for predicting other missing variables.

References

[1] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen. The Modern Mathematics of Deep Learning. In P. Grohs and G. Kutyniok, editors, Mathematical Aspects of Deep Learning, pages 1–111. Cambridge University Press, Cambridge, 2022. doi: 10.1017/9781009025096.002. URL https://www.cambridge.org/core/product/7C3874F83A5D934E5FDC984B8457D553.

[2] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619(7970):533–538, July 2023. doi: 10.1038/s41586-023-06185-3. URL https://doi.org/10.1038/s41586-023-06185-3.

[3] C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, P. Garvan, M. Riechert, J. A. Weyn, H. Dong, J. K. Gupta, K. Thambiratnam, A. T. Archibald, C.-C. Wu, E. Heider, M. Welling, R. E. Turner, and P. Perdikaris. A foundation model for the earth system. Nature, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

[4] P. G. Constantine. Active subspaces: emerging ideas for dimension reduction in parameter studies. Number 2 in SIAM spotlights. Society for Industrial and Applied Mathematics, Philadelphia, 2015.

[5] D. Brunet, E. R. Vrscay, and Z. Wang. On the Mathematical Properties of the Structural Similarity Index. IEEE Transactions on Image Processing, 21(4):1488–1499, Apr. 2012. doi: 10.1109/TIP.2011.2173206.

[6] F. A. Dahlen and F. J. Simons. Spectral estimation on a sphere in geophysics and cosmology. Geophysical Journal International, 174(3):774–807, Sept. 2008. doi: 10.1111/j.1365-246X.2008.03854.x. URL https://doi.org/10.1111/j.1365-246X.2008.03854.x.

[7] ECMWF. IFS Documentation CY41R2 - Part III: Dynamics and Numerical Procedures. In IFS Documentation CY41R2, IFS Documentation. ECMWF, 2016. doi: 10.21957/83wouv80. URL https://www.ecmwf.int/node/16647.

[8] L. Hardy and I. Finney. Leveraging state-of-the-art ai models to forecast wind power generation using deep learning. Meteorological Applications, 32(2):e70038, 2025. doi: https://doi.org/10.1002/met.70038. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.70038.

[9] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730): 1999–2049, July 2020. doi: 10.1002/qj.3803. URL https://doi.org/10.1002/qj.3803. Publisher: John Wiley & Sons, Ltd.

[10] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. Science, 382(6677):

1416–1421, Dec. 2023. doi: 10.1126/science.adi2336. URL https://doi.org/10.1126/science.adi2336. Publisher: American Association for the Advancement of Science.

[11] S. Lang, M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M. C. A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. B. Bouallègue, A. P. Nemesio, P. D. Dueben, A. Brown, F. Pappenberger, and F. Rabier. AIFS – ECMWF's data-driven forecasting system, Aug. 2024. URL http://arxiv.org/abs/2406.01465. arXiv:2406.01465 [physics].

[12] F. Lehmann, F. Ozdemir, B. Soja, T. Hoefler, S. Mishra, and S. Schemm. Finetuning a Weather Foundation Model with Lightweight Decoders for Unseen Physical Processes, June 2025. URL http://arxiv.org/abs/2506.19088. arXiv:2506.19088 [cs].

[13] J. Nocedal and S. J. Wright. Numerical optimization. Springer series in operations research. Springer, New York, 2nd ed edition, 2006.

[14] J. T. Oden and S. Prudhomme. Control of modeling error in calibration and validation processes for predictive stochastic models. International Journal for Numerical Methods in Engineering, 87(1-5):262–272, July 2011. doi: 10.1002/nme.3038. URL https://doi.org/10.1002/nme.3038. Publisher: John Wiley & Sons, Ltd.

[15] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, Feb. 2022. URL http://arxiv.org/abs/2202.11214. arXiv:2202.11214 [physics].

[16] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson. Probabilistic weather forecasting with machine learning. Nature, Dec. 2024. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

[17] A. Quarteroni, P. Gervasio, and F. Regazzoni. Combining physics-based and data-driven models: advancing the frontiers of research with Scientific Machine Learning, Jan. 2025. URL http://arxiv.org/abs/2501.18708. arXiv:2501.18708 [math].

[18] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. Journal of Advances in Modeling Earth Systems, 12(11):e2020MS002203, Nov. 2020. doi: 10.1029/2020MS002203. URL https://doi.org/10.1029/2020MS002203.

[19] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang. Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment. In 2018 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, July 2018. doi: 10.1109/ICME.2018.8486584. Journal Abbreviation: 2018 IEEE International Conference on Multimedia and Expo (ICME).

[20] S. Malardel, Nils Wedi, Willem Deconinck, Michail Diamantakis, Christian Kuehnlein, G. Mozdzynski, M. Hamrud, and Piotr Smolarkiewicz. A new grid for the IFS. ECMWF Newsletter, (146):23–28, 2016. doi: 10.21957/zwdu9u5i. URL https://www.ecmwf.int/node/17262. Publisher: ECMWF Section: Meteorology.

[21] S. Shalev-Shwartz and S. Ben-David. Understanding machine learning: from theory to algorithms. Cambridge university press, New York, 2014.

[22] C. Subich. Efficient Fine-Tuning of 37-Level GraphCast with the Canadian Global Deterministic Analysis. Artificial Intelligence for the Earth Systems, 4(3):e240101, July 2025. doi: 10.1175/AIES-D-24-0101.1. URL https://journals.ametsoc.org/view/journals/aies/4/3/AIES-D-24-0101.1.xml.

[23] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains, June 2020. URL http://arxiv.org/abs/2006.10739. arXiv:2006.10739 [cs].

[24] M. Waqas et al. Artificial Intelligence and Numerical Weather Prediction Models: A Technical Survey. Natural Hazards Research, 2024. doi: https://doi.org/10.1016/j.nhres.2024.11.004. URL https://www.sciencedirect.com/science/article/pii/S266659212400091X.

[25] M. A. Wieczorek and M. Meschede. SHTools: Tools for Working with Spherical Harmonics. Geochemistry, Geophysics, Geosystems, 19(8):2574–2592, Aug. 2018. doi: 10.1029/2018GC007529. URL https://doi.org/10.1029/2018GC007529. Publisher: John Wiley & Sons, Ltd.

[26] M. A. Wieczorek and F. J. Simons. Localized spectral analysis on the sphere. Geophysical Journal International, 162(3):655–675, Sept. 2005. doi: 10.1111/j.1365-246X.2005.02687.x. URL https://doi.org/10.1111/j.1365-246X.2005.02687.x.

[27] M. A. Wieczorek and F. J. Simons. Minimum-Variance Multitaper Spectral Estimation on the Sphere. Journal of Fourier Analysis and Applications, 13(6):665–692, Dec. 2007. doi: 10.1007/s00041-006-6904-1. URL https://doi.org/10.1007/s00041-006-6904-1.

[28] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, Apr. 2004. doi: 10.1109/TIP.2003.819861.