Karlsruhe Institute of Technology

Singular vectors in machine learning-based weather prediction models

Master's Thesis by

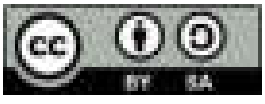
Kevin Gramlich

At the KIT Department of Physics
Institute of Meteorology and Climate Research

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

01. December 2024 – 27. November 2025

Karlsruher Institut für Technologie
Fakultät für Physik
76131 Karlsruhe



Singular vectors in machine learning-based weather prediction models (Master's Thesis)

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

Karlsruhe, 27. November 2025

. . . . . .

(Kevin Gramlich)

Abstract

The past years have been characterized by the rapid progress of Machine Learning based Weather Prediction (MLWP) models. MLWP models are trained on reanalysis data sets to predict future atmospheric states. Modern MLWP models achieve forecast accuracy similar to that of state-of-the-art Numerical Weather Prediction (NWP) models. At the same time, the forecast inference times[1] of MLWP models, such as *Pangu-Weather*, are orders of magnitude faster than the forecast inference times of NWP models. The low computational costs of MLWP models allow for the creation of large ensembles. However, it is unclear how to optimally perturb the initial conditions to generate these large ensembles. Here, we introduce the concept of the Arnoldi method for Singular Vector (SV) calculation, which is an adjoint-free approach. The Arnoldi method for SV calculation allows the use of the full non-linear MLWP model to predict perturbation growth. Thus, there is no need to maintain tangent linear and adjoint models parallel to the full non-linear forecast model.

To assess the validity and potential value of Arnoldi-SVs (A-SVs) in Pangu-Weather, this thesis first explores the general structure of A-SVs in Pangu-Weather and compares A-SVs to the general structure of SVs in physics-based NWP models. A case study on Cyclone Xynthia outlines the atmospheric sensitivities triggered by A-SVs. Third and finally, A-SVs are used to initialize a medium-range ensemble forecast with Pangu-Weather for the year 2022. The ensemble initialization follows the general set-up of the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) ensemble initialization scheme. The forecast skill of the SV-based Pangu-Weather ensemble is evaluated both with respect to the benchmark physics-based IFS ensemble and other Pangu-Weather-based ensemble approaches.

The general structure of A-SVs in Pangu-Weather is very similar to the structure of SVs in physics-based NWP models. A-SV perturbations show cross-variable consistency. A-SV-based sensitivities in Pangu-Weather are very similar to previous studies of Cyclone Xynthia. A-SVs also indicate the potential for further intensification of Cyclone Xynthia under more optimal conditions. Across the European domain, the A-SV-based Pangu-Weather ensemble is generally well-calibrated, with a tendency towards under-dispersion at lead times of 8 to 10 d. From the forecast variables tested, the 2 m Temperature (T2M) ensemble exhibits significant forecast skill over the ECMWF IFS ensemble. The other forecast variables have an 8 to 10 % higher Continuous Ranked Probability Score (CRPS) than the IFS ensemble. Compared to other Pangu-Weather-based ensemble approaches, the A-SV-based ensemble presented in this thesis is highly competitive. Overall, this thesis reveals the potential of

[1]The duration it takes for a MLWP model to make a prediction after being trained.

MLWP models to generate initial condition perturbations at low computational costs that could be used to initialize MLWP and NWP ensemble forecasts.

Zusammenfassung

In den letzten Jahren wurde die Numerische Wettervorhersage (NWP) durch den rapiden Fortschritt Künstlicher Intelligenz revolutioniert. Neue durch Maschinelles Lernen (ML) getriebene Wettervorhersage (MLWP) Modelle erreichen eine ähnliche Vorhersagegüte wie die besten physikalischen NWP Modelle. MLWP Modelle, wie *Pangu-Weather*, zeichnen sich durch eine vielfach geringere Rechenzeit einer Vorhersage aus. Dies eröffnet die Möglichkeit große Ensembles zu generieren. Es ist jedoch unklar wie man optimale Anfangswertstörungen für große Ensembles erzeugt. Für MLWP Modelle lässt sich die Arnoldi Methode deutlich effizienter zur Berechnung von Singulären Vektoren (SVs) verwenden. Im Gegensatz zu der traditionellen Methode für die SVs Berechnung ist die Arnoldi Methode nicht auf ein lineares Tangentenmodell angewiesen, um Anfangswertstörungen zu generieren. Es kann vielmehr mit dem nicht-linearen Vorhersagemodell für den Algorithmus gearbeitet werden, wodurch kein lineares und kein adjungiertes Modell parallel zum nicht-linearen Vorhersagemodell entwickelt werden muss.

In dieser Arbeit werden die physikalische Konsistenz und der potentielle Nutzen von SVs, berechnet mit der Arnoldi Methode (A-SVs), in Pangu-Weather analysiert. Hierfür werden zuerst die Strukturen von A-SVs in Pangu-Weather beleuchtet und mit den bekannten Strukturen von SVs in physikalischen NWP Modellen verglichen. Eine Fallstudie zu Zyklon Xynthia betrachtet die atmosphärischen Sensitivitäten, die von A-SVs in Pangu-Weather angeregt werden. Als letztes wird eine Ensemblevorhersage mit Pangu-Weather für das Jahr 2022 mit Hilfe der A-SVs berechnet. Die Ensemble Initialisierung folgt dabei dem Schema des ECMWF IFS Ensembles. Die Vorhersagegüte der A-SV-basierten Ensemblevorhersage wird im Vergleich zu der IFS Ensemblevorhersage, dem Maßstab für physikalische NWP Modelle, und weiteren Pangu-Weather-basierten Ensemblevorhersagen verifiziert.

Die Struktur von A-SVs in Pangu-Weather ist ähnlich wie die bekannten Strukturen von SVs in physikalischen NWP Modellen. Zudem sind die A-SV Störungen physikalisch konsistent. Die Fallstudie zu Zyklon Xynthia zeigt, dass A-SVs in Pangu-Weather sehr ähnliche atmosphärische Sensitivitäten im Vergleich zu früheren Studien anregen. Zudem zeigen A-SVs, dass Zyklon Xynthia unter optimaleren Umgebungsbedingungen intensiver ausgefallen wäre. Über Europa ist die A-SV-basierte Ensemblevorhersage von Pangu-Weather gut kalibriert, mit einer Tendenz zu unterdispersiven Ensemblevorhersagen für lange Vorhersagezeiten. Aus der Auswahl an getesteten Vorhersagegrößen ist die 2 m-Temperatur (T2M) die beste Vorhersagegröße und weist signifikanten Skill gegenüber dem ECMWF IFS Ensemble auf. Die weiteren Vorhersagegrößen haben eine geringere Vorhersagegüte im Vergleich zum IFS Ensemble. Der Vergleich zu anderen Pangu-Weather-basierten Ensemblevorhersagen zeigt, dass die A-SV-basierte Ensemblevorhersage konkurrenzfähig ist.

Contents

List of Figures

1 Introduction

A plethora of processes govern the dynamics of Earth's atmosphere. In the last decades, substantial amounts of research have been dedicated to describing atmospheric physics as accurately as possible. To predict the future state of the atmosphere, NWP models integrate equations, derived from the laws of physics, forward in time. More and more sophisticated NWP models have been developed since the second half of the 20^{th} century. These led to a continuous, small-step increase in forecast accuracy in the last decades (Bauer et al., 2015). A major step in this "quiet revolution" was the introduction of ensemble forecasts that provide probabilistic predictions of the future atmospheric state. NWP-based ensembles introduce two types of errors in the prediction of future atmospheric states. First, the initial condition error arises from a limited knowledge of the atmospheric state at initial time. The limited and non-homogeneous coverage of observation systems prevents an accurate determination of the atmospheric state at initial time (Leutbecher and Palmer, 2008). Secondly, model errors are induced into NWP models by, e. g., truncations or parameterizations, which are required to describe atmospheric processes with the available computational resources (Buizza et al., 1999).

Although model error can be reduced by improving the NWP model, the inherent chaotic nature of the atmosphere fundamentally limits predictability in the system (Lorenz, 1969). In this chaotic system, small changes in the initial state of the atmosphere can rapidly amplify during forward integration (e. g., Zhang et al., 2007; Baumgart et al., 2019). This effect is commonly known as the *butterfly effect*. Hence, even with perfect models, there is an intrinsic predictability limit that restricts the forecast horizon due to the rapid growth of even the smallest initial condition perturbations (Selz et al., 2022). Thus, every forecast of a NWP model has a certain, but unknown, uncertainty attached to it. Therefore, the forecast of the forecast uncertainty is equally as important as the forecast itself (Leutbecher and Palmer, 2008). A forecast of forecast uncertainty is achieved by calculating an ensemble of forecasts. Ideally, every state variable of the state vector of the atmospheric system would be perturbed individually. The state vector contains all values of every variable at every grid point in a NWP model. With modern NWP models, the dimensional size of the state vector easily exceeds 10^8 (Winkler and Denhard, 2025). As such, it is not feasible to perturb every state variable of the system (Leutbecher and Palmer, 2008; Diaconescu and Laprise, 2012). To still obtain an accurate sample of the unknown forecast uncertainty, a large number of methods have been developed to estimate the uncertainty in the system (e. g., Leutbecher and Palmer, 2008; Cohn, 1997; Buizza et al., 1999; Plant and Craig, 2008). The world-leading ECMWF IFS ensemble forecast uses SVs to identify the fastest growing perturbations in the initial conditions. It is assumed that most of the forecast error is projected into the models by the fastest growing perturbations (Buizza, 1998; Leutbecher and Palmer, 2008).

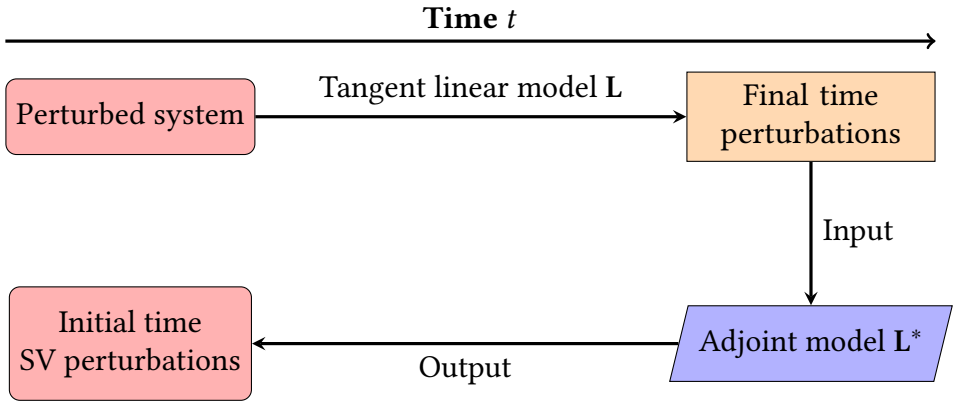


Figure 1.1: Schematic overview of SV computation with the Lanczos algorithm

Lanczos-SV (L-SV) calculation starts with an initial perturbation of the system. The perturbation is then propagated forward in time by the tangent linear model, as shown in Figure 1.1. At final time, the largest perturbations are easily identified. These perturbations are provided to the adjoint model as input. As output, the adjoint model evaluates which initial time perturbations are most sensitive to the final time perturbations. Hence, these initial time perturbations lead to the largest perturbation growth at final time.

The third decade of the 21st century saw a revolution of NWP. MLWP models have made rapid progress in terms of forecast accuracy on a global scale (Ben Bouallègue et al., 2024). Today, MLWP models achieve similar forecast accuracy compared to state-of-the-art physics-based NWP models (e. g. Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Kochkov et al., 2024; Lang et al., 2024; Price et al., 2024). After the training period, these MLWP models have rapid forecast inference times that are orders of magnitude faster than the inference times of state-of-the-art physics-based NWP models (Bi et al., 2023). This opens the opportunity for a novel approach to SV calculation. The Arnoldi method for SV calculation was first introduced by Winkler et al. (2020). It is adjoint-free and can be used with the full non-linear model for the forward integration of the initial time perturbations (Winkler and Denhard, 2025). Figure 1.2 shows a schematic overview of the Arnoldi method for SV calculation. The Arnoldi method develops a lower-dimensional representation of the perturbation growth from which the largest perturbations are derived.

The following work uses the MLWP model *Pangu-Weather* to study A-SVs. Pangu-Weather is one of the first MLWP models to report competitive forecast accuracy with the benchmark physics-based ECMWF IFS model (Bi et al., 2023). In the following analysis, this work aims to answer three key research questions:

1. Do A-SVs in Pangu-Weather have similar structures to L-SVs in physics-based NWP models in terms of their vertical profiles, energy partitions, and horizontal scales at initial and final time?
2. Can A-SVs in Pangu-Weather identify sensitivity patterns similar to those previously analyzed for extra-tropical cyclones in the example of Cyclone Xynthia (2010)?

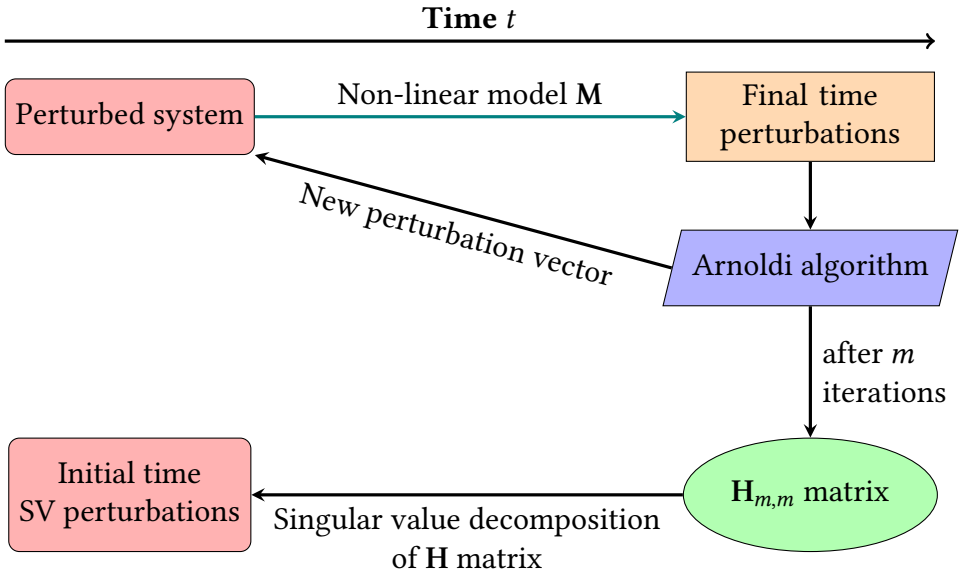


Figure 1.2: Schematic overview of SV computation with the Arnoldi algorithm

3. Can A-SVs be used to generate a reliable ensemble forecast with Pangu-Weather that mimics the L-SV-based approach of the ECMWF IFS ensemble forecast? Does the A-SV-based ensemble have forecast skill over the ECMWF IFS ensemble or other Pangu-Weather-based ensemble approaches?

The theoretical background of SVs, their calculation using the Lanczos and Arnoldi algorithms, as well as their characteristic structures in physics-based NWP models, are introduced in Chapter 2. Chapter 3 presents ECMWF Reanalysis v5 (ERA5), the reference data set used in this study, Pangu-Weather, and the implementation of A-SVs for ensemble generation with Pangu-Weather. The structure of A-SVs in Pangu-Weather is discussed in Chapter 4. The sensitivity analysis for Cyclone Xynthia is presented in Chapter 5. In Chapter 6, the forecast skill of the A-SV-based Pangu-Weather ensemble is evaluated. This work concludes in Chapter 7 with a summary of the key findings of this study and an outlook for future research on this topic.

2 Theoretical background

The following chapter outlines the theoretical concepts of the SV calculation. First, we introduce the concept of SVs. In a second step, the SV computation with the Lanczos algorithm (Lanczos, 1950) in physics-based NWP models is discussed. Then, we present the structure of L-SVs in physics-based NWP models. The fourth section introduces the Arnoldi method for SV calculation. The chapter concludes with an overview of the initialization of the L-SV-based ECMWF IFS ensemble forecast.

2.1 Singular vectors

SVs, first introduced by Lorenz (1965), are one method to estimate the fastest growing perturbations of a system. The system is a non-linear model $\mathbf{M}$ that propagates the state vector x_0 from initial time t_0 to final time t. The result is the final time state vector x_t given by the following relation

$$\mathrm{x}_t = \mathbf{M}\left(\mathrm{x}_0\right). \tag{2.1}$$

Similarly, a small perturbation x_0' can be added to x_0 and propagated to time t. The final time perturbation x_t' is then given by

$$\mathrm{x}_t' = \mathbf{M}\left(\mathrm{x}_0'\right). \tag{2.2}$$

In the context of atmospheric sciences, $\mathbf{M}$ represents a forecast with a NWP model. The assumption can be made that, on a sufficiently short time scale and over sufficiently large horizontal scales, the processes in the atmosphere, and therefore $\mathbf{M}$ itself, can be linearized by a tangent linear operator $\mathbf{L}$ (e. g., Gilmour et al., 2001). $\mathbf{L}$ propagates x_0' to time t by

$$\mathrm{x}_t' = \mathbf{L}\left(t_0, t\right)\mathrm{x}_0'. \tag{2.3}$$

The largest elements of $\mathbf{L}$ are derived from a Singular Value Decomposition (SVD) of the tangent linear operator $\mathbf{L}$. $\mathbf{L}$ can be decomposed as

$$\mathbf{L} = \mathbf{E}^{-1/2}\mathbf{U}\mathbf{D}\mathbf{V}\mathbf{E}^{1/2}, \tag{2.4}$$

with $\mathbf{D}$ being a diagonal matrix containing the singular values σ_i of $\mathbf{L}$ such that $\mathbf{D} = \mathrm{diag}\left(\sigma_1, \sigma_2, \sigma_3, ..., \sigma_n\right)$ for $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ... \geq \sigma_n$. The columns of $\mathbf{V}$ contain the initial time SVs of $\mathbf{L}$, while the columns of $\mathbf{U}$ contain the final time SVs of $\mathbf{L}$. $\mathbf{E}$ is a metric for measuring

perturbation growth. The leading singular value of **L** maximizes the ratio of initial to final time perturbation energy, given by the following relation

$$\sigma^2 = \frac{||\mathrm{x}(t)||^2_{E_t}}{||\mathrm{x}(t_0)||^2_{E_0}} = \frac{\langle \mathbf{L}\mathrm{x}'_0; E_t \mathbf{L}\mathrm{x}'_0 \rangle}{\langle \mathrm{x}'_0; E_t \mathrm{x}'_0 \rangle}, \tag{2.5}$$

with E_0 and E_t being the respective norms at initial and final time.

2.2 Traditional NWP SV computation

To compute the leading SVs of matrix **L**, the adjoint of matrix **L**, $\mathbf{L}^*$, is used. In this context, $\mathbf{L}^*$ is the adjoint tangent linear operator of **L**, which derives the initial time sensitivities of the final time perturbations (Errico, 1997). NWP models use the following relation

$$(\mathbf{L}^*\mathbf{P}^* E_t \mathbf{P}\mathbf{L})\, \mathrm{v}_i(t_0) = \lambda_i^2 E_0 \mathrm{v}_i(t_0), \tag{2.6}$$

to iteratively compute the leading SVs of **L** (Leutbecher and Palmer, 2008). v_i is the i-th initial time SV of **L**. **P** is a local projection operator that limits the domain in which perturbation growth is calculated, with $\mathbf{P}^*$ being the respective adjoint of **P** (Hartmann et al., 1995).

In the context of atmospheric sciences, Equation 2.6 can be interpreted by the following process. The system, i. e., the atmosphere, is perturbed by some random perturbation at initial time. This perturbation is then propagated by the tangent linear model **L** to final time to compute which perturbations have grown largest at final time. In addition, a local projection operator **P** can be applied to the atmospheric state to limit the domain in which the perturbation growth is calculated. A common usage for **P** is e. g., to limit the domain to the vicinity of a tropical cyclone when computing the leading L-SVs of that tropical cyclone (e. g., Peng and Reynolds, 2006; Kim and Jung, 2009a,b; Reynolds et al., 2009). Then, the final time perturbations are provided to the adjoint model $\mathbf{L}^*$ as input. As output, $\mathbf{L}^*$ derives the initial time perturbations responsible for the largest perturbation growth at final time.

The iterative computation of the leading SVs of **L** with Equation 2.6 forms an E-orthonormal basis. Hence, the second SV of **L** maximizes perturbation growth under the constraint of being orthogonal to the leading SV. The third SV of **L** maximizes perturbation growth under the constraint of being orthogonal to the first two SVs of **L**, and so on.

2.2.1 Constraints on Lanczos-SV computation

By construction, SVs do not identify the absolute largest perturbations in the system. Rather, the construction method implies constraints under which L-SVs maximize perturbation growth. Following Diaconescu and Laprise (2012), these constraints are presented in the following paragraphs.

Norms

First, the norms E_0 and E_t applied to L-SVs in Equation 2.6 impact the perturbations that have grown the largest by final time. A common choice is the dry total energy norm (Diaconescu and Laprise, 2012), which can be computed using the following relation

$$\|x'_t\|^2 = \frac{1}{2}\int_{p_0}^{p_1}\int_A \left(u'^2 + v'^2 + \frac{c_\mathrm{p}}{T_\mathrm{r}}T'^2\right) dpdA + \frac{1}{2}R_\mathrm{d}T_\mathrm{r}p_\mathrm{r}\int_A (\ln p'_\mathrm{sfc})^2 dA. \tag{2.7}$$

p_0 and p_1 are the lower and upper pressure thresholds between which the dry total energy norm is calculated. u' and v' are the zonal and meridional wind perturbations, and T' is the temperature perturbation. c_p is the specific heat constant for dry air at constant pressure, and T_r is a reference temperature. R_d is the gas constant for dry air. p_r is a reference pressure and p'_sfc is the surface pressure perturbation. Equation 2.7 includes three types of perturbation energies, the kinetic perturbation energy ($u'^2 + v'^2$), the potential perturbation energy (T'^2), and the surface pressure perturbation energy $\left((\ln p'_\mathrm{sfc})^2\right)$.

Optimization time interval

Implicitly included in Equation 2.6 is the Optimization Time Interval (OTI), the time interval between initial t_0 and final time t. The L-SVs depend on the chosen OTI (Buizza, 1994). Common choices for the OTI range between 24 and 48 h (e. g., Buizza, 1994; Peng and Reynolds, 2006; Leutbecher and Palmer, 2008; Kim and Jung, 2009a; Diaconescu et al., 2012). On longer timescales, the linear assumption of the tangent linear model is violated (Gilmour et al., 2001). In the tropics, shorter OTIs are discussed due to larger non-linear perturbation growth compared to the extratropics (e. g., Kim and Jung, 2009a; Diaconescu and Laprise, 2012).

Tangent linear and adjoint model

Third and finally, L-SVs depend on the tangent linear and adjoint models. Both the resolution of the model (e. g., Buizza, 1994; Buizza et al., 1997; Coutinho et al., 2004) and the parameterizations included in these models (Zadra et al., 2004) affect the derived L-SVs. Typically, the tangent linear and adjoint models have a much coarser resolution than the full non-linear NWP model (e. g., Buizza, 1998; Coutinho et al., 2004; Leutbecher and Palmer, 2008). Furthermore, the choice of parameterizations significantly influences the computed L-SVs (e. g., Palmer et al., 1998; Zadra et al., 2004; Kim and Jung, 2009a). Zadra et al. (2004) analyzed the effects of different parameterizations in the tangent linear and adjoint models on the computed L-SVs. Vertical diffusion parameterizations must be included in the tangent linear and adjoint models to prevent spurious near-surface perturbation growth (Zadra et al., 2004). Parameterization of sub-grid-scale orographic drag reduces perturbation growth by inducing a damping term in the models (Zadra et al., 2004). Large-scale precipitation is distinguished into two types of parameterizations. The large-scale stratiform precipitation parameterization increases perturbation growth by L-SVs (Zadra et al., 2004). Interestingly,

the inclusion of a parameterization for large-scale convective precipitation into the tangent linear and adjoint models has only limited effects on perturbation growth (Zadra et al., 2004).

2.2.2 Limitation of Lanczos-SV computation

In the context of operational NWP models, there are significant limitations to L-SV computation (Diaconescu and Laprise, 2012). Operational NWP models are continuously improved (Bauer et al., 2015). This implies the need for continuous updates to the tangent linear and adjoint models to maintain the linearization assumption (Winkler and Denhard, 2025). As mentioned, changes in the tangent linear and adjoint models also lead to changes in the calculated L-SVs (e. g., Palmer et al., 1998; Buizza, 1998; Zadra et al., 2004; Coutinho et al., 2004; Kim and Jung, 2009a). In addition, there are only limited computational resources for L-SV computation (Leutbecher and Palmer, 2008). Therefore, only a low number of L-SVs can be calculated with the coarser-resolved tangent linear and adjoint models with Equation 2.6 (Leutbecher and Palmer, 2008). However, only a limited number of L-SVs have rapidly growing perturbations (Diaconescu and Laprise, 2012). Hence, a sample of the fastest-growing L-SVs will account for a large fraction of the forecast error (Buizza, 1998; Diaconescu et al., 2012).

Additionally, the linear assumption must be maintained during the forward integration of the perturbations (Kim and Jung, 2009a; Diaconescu and Laprise, 2012). On synoptic scales in the extratropics, the linear assumption holds for a maximum of 2 days (Buizza, 1994; Gilmour et al., 2001). In the tropics, larger contributions of non-linear perturbation growth lead to an even shorter validity of the linear approximation (Diaconescu and Laprise, 2012).

2.3 Structure of Lanczos-SVs in physics-based NWP models

This section presents key characteristics of L-SVs in physics-based NWP models. First, the vertical structure of L-SVs in physics-based NWP models is presented. In a second step, we discuss the energy partition of L-SV perturbations between kinetic and potential perturbation energy at initial and final time. The third key feature that is highlighted is the horizontal scales of the perturbation energy of L-SVs. The purpose of this discussion is to provide a reference framework for the evaluation of A-SVs in Pangu-Weather.

2.3.1 Vertical structure of Lanczos-SVs in physics-based NWP models

A first key characteristic of L-SVs in physics-based NWP models is the vertical structure of L-SV perturbations. Figure 2.1 shows the vertical profile of the L-SV-based temperature perturbations over the North American continent. At initial time, the opposing temperature perturbations have a westward tilt with height. This westward tilt indicates baroclinically

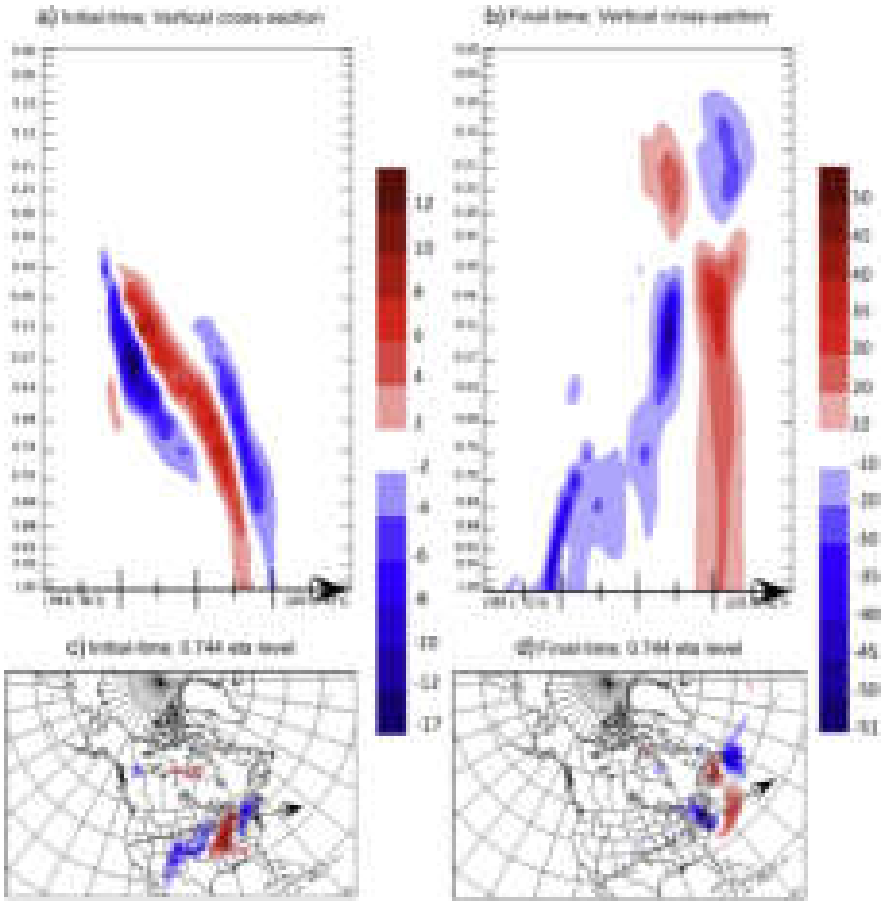


Figure 2.1: Vertical profile of the temperature perturbations of the leading extratropical L-SV over North America at initial time (a) and final time (b) for forecasts initialized on 05 December 1992, 00 UTC and L-SVs optimized for an OTI of 36 h with the dry total energy norm. The vertical axis of the cross-sections corresponds to the η levels of the NWP model. The vertical cross-section trajectories in the upper plots are indicated by the arrows in the respective lower plots. The lower plots show the temperature perturbations at vertical level $\eta = 0.744$. The norm of the final time cross-section changes to account for the perturbation growth. (Diaconescu and Laprise, 2012, Fig. 1)

unstable perturbations in the mid-latitudes (e. g., Buizza, 1994; Coutinho et al., 2004; Diaconescu et al., 2012). Figure 2.1 shows that the strongest temperature perturbations at initial time are located in the mid-troposphere.

At final time, the temperature perturbations are aligned vertically. In addition, the magnitude of the temperature perturbations has grown significantly during forward integration. Note that the location of the cross-section trajectory in Figure 2.1 has changed compared to initial time. The upright vertical axis of the temperature perturbations at final time is consistent with baroclinic disturbances that are typically triggered by L-SVs (Hoskins et al., 2000; Diaconescu et al., 2012). A detailed analysis of the final time perturbations reveals that the temperature perturbation maximum has shifted upward.

Figure 2.2 shows the vertical profile of the perturbation energy at both initial and final time. These profiles highlight the distinct upward propagation of the perturbation energy

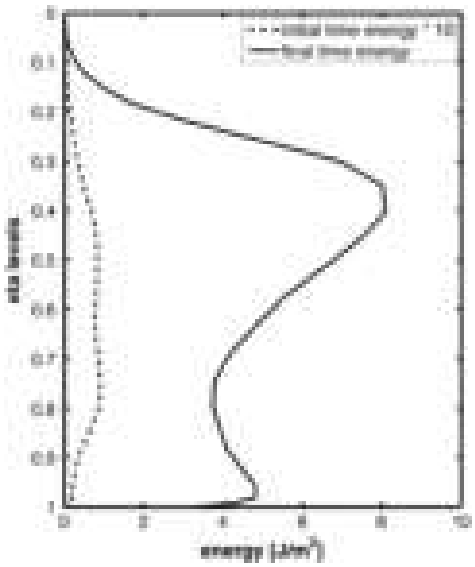


Figure 2.2: Vertical distribution of the total perturbation energy at initial (dashed line) and final time (solid line) averaged across the leading ten extratropical L-SVs for forecasts initialized on 05 December 1992, 00 UTC and L-SVs optimized for an OTI of 36 h with the dry total energy norm. The initial time total perturbation energy profile is multiplied by a factor of ten to improve visibility. (Diaconescu and Laprise, 2012, Fig. 3)

maximum from initial to final time. From a broad mid-tropospheric maximum at initial time, the perturbation energy maximum propagates upward to the jet level. This upward perturbation growth is consistent with the general dynamics of error growth in NWP models described by Baumgart et al. (2019). Baumgart et al. (2019) show that small-scale perturbations are initialized by mesoscale convective systems. The divergent outflow of these mesoscale convective systems projects the perturbations to the tropopause. Here, perturbation growth is driven by balanced dynamics on synoptic scales at the jet level. The perturbations are then propagated across the hemisphere by the mid-latitude jet (Baumgart et al., 2019). In addition, the final time vertical profile in Figure 2.2 shows a secondary perturbation energy maximum at lower levels.

2.3.2 Energy partition of Lanczos-SVs in physics-based NWP models

The energy partition of L-SV perturbations is a second key characteristic of L-SVs that is presented in the following. Figure 2.3 depicts the energy partition of the total perturbation energy at initial and final time. At initial time, the total perturbation energy is dominated by the potential perturbation energy (Zadra et al., 2004; Diaconescu et al., 2012). Equation 2.7 shows that the potential perturbation energy is directly proportional to the temperature perturbation within the atmosphere. At final time, the total perturbation energy is dominated by the kinetic perturbation energy (Zadra et al., 2004; Diaconescu et al., 2012). The surface pressure perturbation energy contributes only marginally to the total perturbation energy at both initial and final time (Zadra et al., 2004).

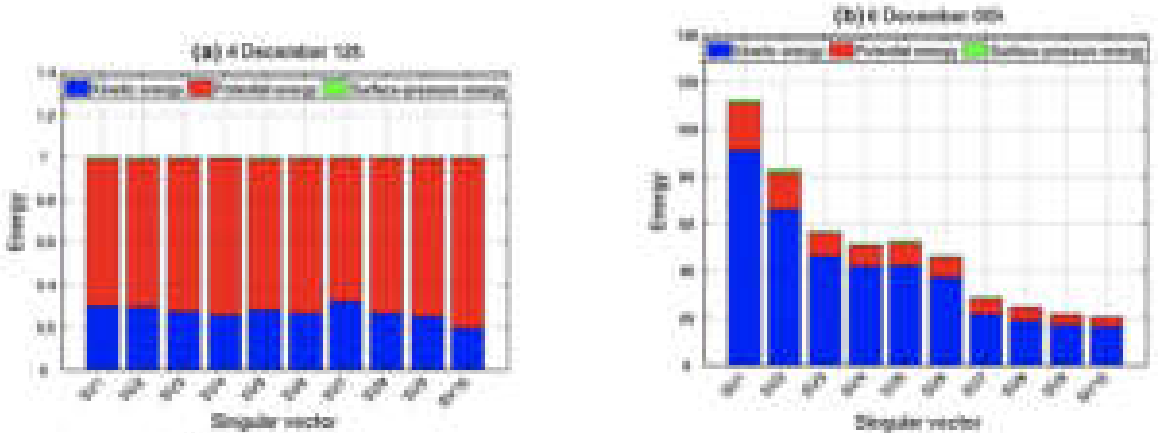


Figure 2.3: Energy partition of the leading ten extratropical L-SVs at initial time (a) and at final time (b) for forecasts initialized on 05 December 1992, 00 UTC and L-SVs optimized for an OTI of 36 h with the dry total energy norm. The kinetic perturbation energy is shown in blue, the potential perturbation energy in red, and the surface-pressure perturbation energy in green. The initial time perturbation energy is normalized to one. (Diaconescu et al., 2012, adapted from Fig. 6)

Figure 2.3 shows that at initial time the ratio of potential to kinetic perturbation energy is 3:1. By final time, this ratio shifts to 1:3. These characteristic energy partitions are independent of the configuration of the L-SV computation in physics-based NWP models (Zadra et al., 2004). The shift from potential to kinetic perturbation energy is already present after the first integration step of the forecast model (Diaconescu et al., 2012).

From a mathematical perspective, the initial change in the energy partition of the L-SV perturbations is attributed to the one-time rotation of the initial time L-SVs onto the attractor of the system (Szunyogh et al., 1997). By construction, the initial time L-SVs are located off the attractor of the system (Kalnay, 2002). During forward integration, the L-SVs quickly turn towards the attractor in the direction of fastest perturbation growth (Szunyogh et al., 1997).

2.3.3 Horizontal scale of Lanczos-SVs in physics-based NWP models

A third key characteristic of L-SVs is the horizontal scale of the L-SV perturbations. The horizontal scales of L-SV perturbations can be analyzed using power spectra of the total perturbation energy at initial and final time, as shown in Figure 2.4. At initial time, the maximum spectral power of the total perturbation energy is located at wavenumbers 16 to 18. The maximum spectral power of the total perturbation energy shifts to smaller horizontal wavenumbers of around 12 to 14 at final time (Diaconescu et al., 2012). A smaller horizontal wavenumber implies that the horizontal scale of the perturbations has grown. Baroclinic disturbances exhibit a similar spectral maximum of the perturbation energy at synoptic scales (Buizza and Palmer, 1995).

Comparison of the respective spectral powers of the kinetic and potential perturbation energies at initial and final time in Figure 2.4 shows that the upscale energy growth for

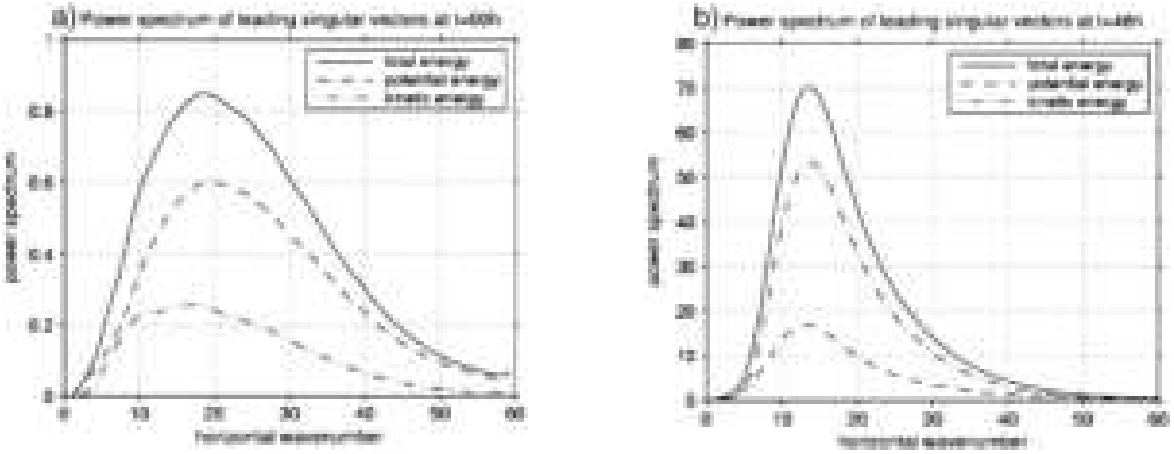


Figure 2.4: Power spectra at initial (a) and final time (b) averaged across the leading 45 L-SVs. Forecasts are initialized from the analysis on 16 February 2002, 00 UTC. L-SVs are optimized for an OTI of 48 h with the dry total energy norm. At initial time, the L-SVs are optimized globally, while, at final time, the norm is limited to latitudes north of 30 °N. The spectrum of the total perturbation energy is given by the continuous, the potential perturbation energy by the dashed, and the kinetic perturbation energy by the dot-dashed line. (Diaconescu and Laprise, 2012, adapted from Fig. 4)

the kinetic perturbation energy is smaller compared to that of the potential perturbation energy (Zadra et al., 2004). For the potential perturbation energy, the spectral maximum is initially located at larger horizontal wavenumbers (i. e., smaller horizontal scales) than the spectral maximum of the total perturbation energy. At final time, the spectral maximum of the total perturbation energy is located at horizontal wavenumbers 12 to 14. This implies that the maximum of potential perturbation energy is located at a larger horizontal scale than the total perturbation energy maximum. For the kinetic perturbation energy, the shift in spectral power is much less pronounced. Different behaviors of the final time kinetic perturbation energy maximum have been found in previous studies (e. g., Zadra et al., 2004; Coutinho et al., 2004).

2.4 Arnoldi method

The routine described in section 2.2 derives the leading SVs iteratively from the matrix product $\mathbf{L}^*\mathbf{L}$ (Leutbecher and Palmer, 2008). The Arnoldi method allows for the adjoint-free computation of the leading SVs of a matrix $\mathbf{A}$ with the help of the Arnoldi algorithm (Arnoldi, 1951; Winkler et al., 2020). In the Arnoldi method, $\mathbf{A}$ is never computed explicitly and is only required for the theoretical basis of the method. Hence, there is no need to compute or store matrix $\mathbf{A}$, which would not be feasible in the context of NWP (Winkler et al., 2020; Winkler and Denhard, 2025). To derive the leading SVs of matrix $\mathbf{A}$, a lower-dimensional representation $\mathbf{H}$ of matrix $\mathbf{A}$ is needed (Winkler and Denhard, 2025). As with the Lanczos algorithm, the Arnoldi algorithm uses the mathematical concept of *Krylov subspaces* (Winkler and Denhard, 2025). The assumption is made that the dynamically most

relevant modes of $\mathbf{A}$ are covered after the computation of a few SVs of $\mathbf{H}$ (Winkler and Denhard, 2025). The transformation of matrix $\mathbf{A}$ is given by the following relation

$$\mathbf{A}_{n,n}\mathbf{Q}_{n,m} = \mathbf{Q}_{n,m}\mathbf{H}_{m,m} + RES, \tag{2.8}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{H} \in \mathbb{R}^{m\times m}$. $\mathbf{Q} \in \mathbb{R}^{n\times m}$ is the orthonormal Krylov basis that links the subspace and matrix $\mathbf{H}$ to the $\mathbb{R}^{n\times n}$ and the matrix $\mathbf{A}$ (Winkler and Denhard, 2025).

Once again, the full non-linear model $\mathbf{M}$ propagates the state vector x_0 at initial time t_0 to the state vector x_t at final time t. In the context of the Arnoldi method, the evolved increments I are defined as the difference between the perturbed and unperturbed state vector x_0 at time t as

$$I(\mathrm{v}) \equiv \mathbf{M}(\mathrm{x}_0 + h\mathrm{v}) - \mathbf{M}(\mathrm{x}_0), \tag{2.9}$$

with v being a normalized perturbation vector and h being a perturbation amplitude. With the evolved increments $I(\mathrm{v})$, the Evolved Increment Matrix (EIM) can be defined by

$$\mathbf{A} \equiv \big(I(e_1), ..., I(e_n)\big), \tag{2.10}$$

collecting evolved increments. e_i is the i-th unitary vector of $\mathbb{R}^n$. The EIM is a linear map that approximates the perturbation growth between initial time t_0 and final time t. It is important to note that although the EIM itself is a linear map, the full non-linear model $\mathbf{M}$ is used to compute the error growth of each evolved increment (Winkler and Denhard, 2025). The EIM is a matrix consisting of non-linear evolved states.

In the context of a NWP forecast, the EIM can be interpreted as follows. Each column of $\mathbf{A}$ consists of the difference between the model forecast of the perturbed state vector $\tilde{\mathrm{x}}_0$, where one state variable is perturbed, and the forecast of the unperturbed state vector x_0. For each column of $\mathbf{A}$, a single state variable of the state vector is perturbed (Winkler and Denhard, 2025).

For the Arnoldi method, the linearization assumption can be justified in that the non-linear evolved increments of arbitrary vectors can be approximated by a linear combination of the non-linear evolved unitary vectors (Winkler and Denhard, 2025). With the EIM, an evolved increment $I(\mathrm{v})$ can be approximated as follows:

$$I(\mathrm{v}) \approx \sum_{i=1}^{n} \big(\mathbf{M}(\mathrm{x}_0 + he_i) - \mathbf{M}(\mathrm{x}_0)\big)v_i = \big(I(e_1), ..., I(e_n)\big)\mathrm{v}, \tag{2.11}$$

where $\mathrm{v} = (v_1, ..., v_n)$. A SVD of the EIM provides the largest elements of $\mathbf{A}$. In the context of NWP forecasts, these point in the direction of the largest perturbation growth by final time (Winkler and Denhard, 2025).

2.4.1 SV construction with the Arnoldi method

Figure 2.5 provides a schematic overview of the SV calculation using the Arnoldi method in the context of MLWP. The Arnoldi method requires at least one initial perturbation

vector to start the iteration process. Initially, the state vector x_0 can be perturbed by a perturbation vector $\mathrm{v}_{(1)}$ consisting of pure white noise to start the algorithm (Winkler and Denhard, 2025). The perturbed state is propagated by the full non-linear model $\mathbf{M}$ to final time, which gives the evolved increment $I\left(\mathrm{q}_{(1)}\right)$. $\mathrm{q}_{(i)}$ are the basis vectors of the Krylov subspace $\mathbf{Q}_{n,m}$. The Arnoldi algorithm provides the entries of the first iteration of the $\mathbf{H}$ matrix, forming $\mathbf{H}_{(1)}$ after one iteration loop. The residual between the evolved increment $I\left(\mathrm{q}_{(1)}\right)$ and the representation of the evolved increment in the Krylov subspace $(\mathrm{q}_{(1)} \cdot \mathbf{H}_{(1)})$ forms the new perturbation vector $\mathrm{q}_{(2)}$. After using Gram-Schmidt orthonormalization (Winkler et al., 2020; Winkler and Denhard, 2025), the initial state vector x_0 at time t_0 is perturbed by the new perturbation vector $\mathrm{q}_{(2)}$. Hence, the algorithm *collects* a set of evolved increments. In addition, $\mathrm{q}_{(2)}$ is the second column of $\mathbf{Q}_{n,m}$, which is the basis of the Krylov subspace. Forward integration provides the new evolved increment $I\left(\mathrm{q}_{(2)}\right)$ from which the new iteration of the $\mathbf{H}$ matrix and the next perturbation vector $\mathrm{q}_{(3)}$ are calculated.

After $m \ll n$ in-place model forecasts and Arnoldi algorithm iterations, the matrix $\mathbf{H}_{m,m}$ is formed. The right SVs of $\mathbf{H}_{m,m}$ are multiplied with $\mathbf{Q}_{n,m}$ in order to transform them back into the original space $\mathbf{A}$. These SVs are the final outcome of the Arnoldi iteration. The SVs can subsequently be used to perturb the state vector x_0 at initial time. The computation time for the $\mathbf{H}_{m,m}$ matrix is reduced due to the Hessenberg shape of the matrix (Winkler et al., 2020; Winkler and Denhard, 2025).

2.4.2 Differences to traditional SV computation

By construction, the two methods presented for SV computation for traditional NWP models (section 2.2) and the Arnoldi method (section 2.4) differ from one another. The following section highlights key differences between the two methods.

To compute SVs with traditional NWP models, a tangent linear and adjoint operator must be developed to generate a linearized forecast model that together with the adjoint operator $\mathbf{L}^*$ can identify initial time sensitivities. In traditional NWP, SVs are derived from the matrix product $\mathbf{L}^*\mathbf{L}$ (see Equation 2.6). In contrast, the Arnoldi method allows for an adjoint-free computation of the leading SVs of the matrix $\mathbf{A}$. Every forward integration is usually performed using the full non-linear model $\mathbf{M}$. In addition, the Arnoldi method uses a different linearization assumption given by Equation 2.11. The Arnoldi method assumes that the non-linear development of arbitrary perturbations can be approximated by a linear combination of the non-linear development of unitary vector perturbations. For the right SVs, the SVD of the lower-dimensional matrix $\mathbf{H}$ is equivalent to solving the eigenvalue problem $\mathbf{H}^T\mathbf{H}$.

These differing linearization assumptions also imply that the two methods approximate the initial model trajectory differently. For the tangent linear operator used in traditional SV calculation in NWP models, the initial trajectory of the model forecast is approximated by a tangent. In contrast, the $\mathbf{H}$ matrix in the Arnoldi method approximates the initial trajectory of the model forecast using secants (Winkler et al., 2020; Winkler and Denhard, 2025). By construction, both methods set up different Krylov subspaces. Hence, the derived SVs will

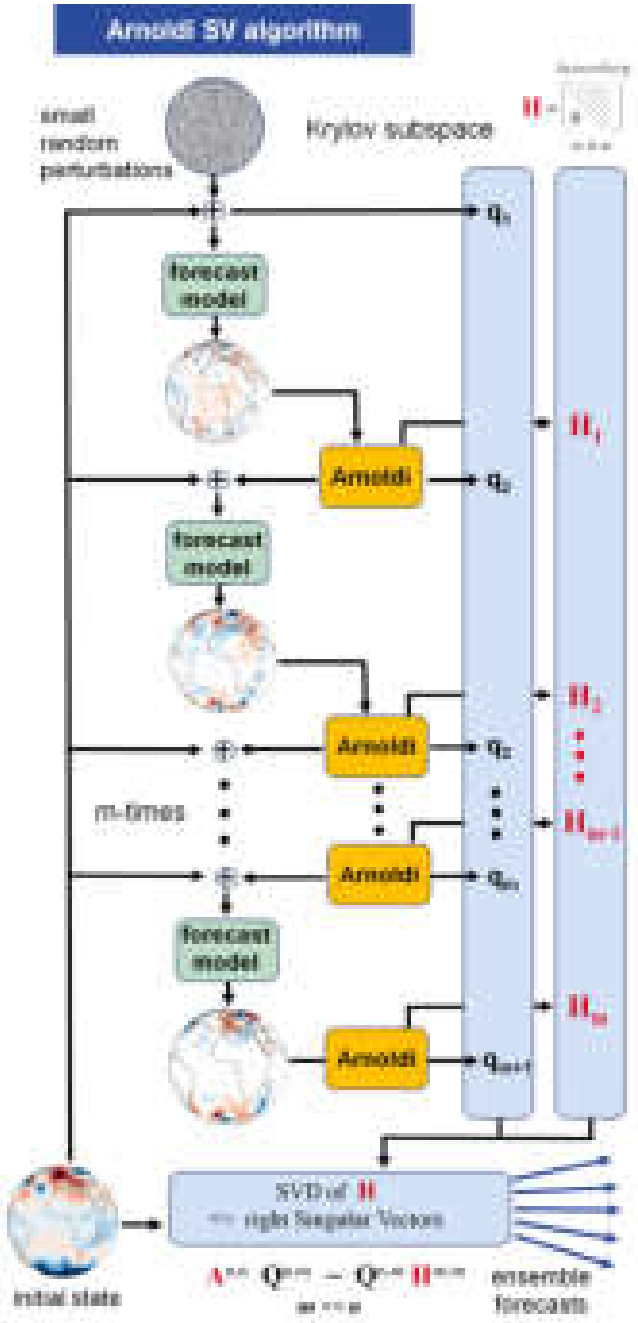


Figure 2.5: Schematic overview of the Arnoldi method for SV calculation. The forecast model is iteratively applied to generate new perturbation vectors q_i. The Arnoldi algorithm, in the yellow boxes, forms the Krylov subspace by building the lower-dimensional **H** matrix ($m \ll n$). In addition, the Arnoldi algorithm uses the occurring perturbations at initial time to perturb the reference state. A SVD of **H** gives the right SVs of **H**. The Krylov basis **Q** projects the right SVs of **H** into the original space, providing the final SVs (of **A**), which can be used as perturbations at initial time. (Winkler and Denhard, 2025, Fig. 1)

most certainly differ depending on the chosen construction method (Winkler and Denhard, 2025).

2.4.3 Block Arnoldi method

The computational efficiency of the Arnoldi method can be further increased through parallelization. Blocks of initial perturbation vectors are used to generate batch-wise forecasts of final time perturbation vectors with the full non-linear model. The Arnoldi algorithm returns blocks of new perturbation vectors and, in a batch, extends the **H** matrix (Winkler and Denhard, 2025). For a parallelizable forecast model **M**, the Block Arnoldi method reduces the number of forecast inferences required to compute the **H** matrix. Thus, the computational resources required for the Arnoldi method are greatly reduced.

2.5 Ensemble initialization at ECMWF

With the help of L-SV-based perturbations, the initial conditions of a NWP model can be perturbed to generate an ensemble of forecasts (Leutbecher and Palmer, 2008). The need for such ensemble forecasts is given by the inherent chaotic nature of the atmosphere (Lorenz, 1969). As mentioned above, the butterfly effect leads to the rapid amplification of small perturbations in the initial conditions (e. g., Zhang et al., 2007; Baumgart et al., 2019). Every forecast has an inherent forecast uncertainty tied to it (Selz et al., 2022). This makes the forecast of the forecast uncertainty equally as important as the forecast itself (Leutbecher and Palmer, 2008). With ensemble forecasting, the inherent but unknown forecast uncertainty is sampled (Leutbecher and Palmer, 2008). In the following section, the L-SV-based ECMWF IFS ensemble initialization is presented.

The ECMWF IFS and its ensemble forecast are commonly used to benchmark new MLWP models and their ensemble forecast skill (e. g., Bi et al., 2023; Ben Bouallègue et al., 2024; Baño-Medina et al., 2025; Bülte et al., 2025). The IFS uses L-SVs to initialize the ensemble prediction system (Leutbecher and Palmer, 2008). The IFS ensemble initialization system uses tangent linear and adjoint models to construct the leading L-SVs as described in section 2.2 (Leutbecher and Palmer, 2008). L-SVs are used to perturb the initial conditions after data assimilation (Leutbecher and Palmer, 2008). The IFS uses Four-Dimensional Variational Analysis (4D-Var) as the data assimilation routine. 4D-Var accounts for observational data across a certain time frame instead of a single time step to adjust the initial model trajectory to the observational data (Le Dimet and Talagrand, 1986). The IFS performs an Ensemble Data Assimilation (EDA) for which a separate 4D-Var iteration is run for every ensemble member. Hence, the EDA provides an ensemble of initial conditions (Leutbecher and Palmer, 2008). This ensemble of initial conditions is subsequently perturbed by the L-SVs.

The L-SVs are optimized for an OTI of 48 h (Leutbecher and Palmer, 2008). The initial conditions are perturbed by a set of random perturbations generated by a linear combination of the leading L-SVs with the following relation

$$\sum_{i=1}^{m} \alpha_i \mathrm{v}_i. \tag{2.12}$$

v_i is i-th initial time L-SV, and the coefficients α_i are sampled from a Gaussian distribution with a mean $\mu = 0$ and a standard deviation of $\sigma = 1$. The Gaussian distribution is truncated at $[-3\sigma; 3\sigma]$ to obtain a compact support sampling distribution and avoid numerical instabilities (Leutbecher and Palmer, 2008). Both at initial and final time, the dry total energy norm (Equation 2.7) is used to measure perturbation growth (Leutbecher and Palmer, 2008).

For each hemisphere, separate sets of leading L-SVs are computed (Leutbecher and Palmer, 2008). Due to the seasonal cycle, the winter hemisphere is the more dynamically active hemisphere (Buizza and Palmer, 1995). If the leading L-SVs were optimized globally, this would confine the L-SV perturbations mainly to the winter hemisphere (Buizza and Palmer, 1995; Leutbecher and Palmer, 2008). Hence, the leading 50 extratropical L-SVs are computed separately for each hemisphere. Perturbation growth is optimized for the extratropical domain extending from 30° to 90° for the respective hemisphere. In addition, for up to six tropical cyclones, the sets of the leading five L-SVs are calculated (Leutbecher and Palmer, 2008). Therefore, the local projection operator **P** limits the perturbation growth to the vicinity of each tropical cyclone. The tropical cyclone-related L-SVs are calculated in the subspace orthogonal to the perturbation space spanned by the leading extratropical L-SVs of the respective hemisphere (Leutbecher and Palmer, 2008).

In the extratropics, the initial condition perturbations of the ECMWF IFS ensemble contain linear combinations of the current initial time L-SV perturbations and the evolved L-SV perturbations of the ensemble initialization 48 h prior (Leutbecher and Palmer, 2008). Due to the limited sample of the zero-mean Gaussian distribution, the mean of the coefficients α_i in the linear combination most often differs from zero (Leutbecher and Palmer, 2008). This would lead to an initial deterioration of the ensemble forecast. Perturbing the ensemble in both directions (positive/negative) enforces a zero-mean distribution of the ensemble perturbations (Leutbecher and Palmer, 2008). For a positively perturbed forecast, the ensemble perturbation is added to the analysis, and for the negatively perturbed forecast, the ensemble perturbation is subtracted from the analysis. From a mathematical point of view, the direction of a SV perturbation is arbitrary and is therefore mathematically equivalent. In addition, the L-SV perturbations are scaled to the typical size of the analysis error (Leutbecher and Palmer, 2008).

L-SV perturbations only cover the fraction of error of NWP models resulting from an imperfect analysis of the initial conditions (Leutbecher and Palmer, 2008). The model error induced by, e. g., parameterizations and truncations can be represented in the ensemble forecast by perturbations of parameterizations of subgrid-scale processes (e. g., Buizza et al., 1999; Plant and Craig, 2008; Leutbecher and Palmer, 2008).

3 Methods and data

Chapter 2 gives an overview of L-SVs in NWP models. In addition, the Arnoldi method for SV calculation is presented in Chapter 2. The following chapter introduces important methods and data sets that are used for the analysis of A-SVs in MLWP models. First, the reference data set ERA5 is presented. Then, we introduce Pangu-Weather, a state-of-the-art MLWP model. In addition, the implementation of the Block Arnoldi method in Pangu-Weather is discussed. The Block Arnoldi method provides the A-SVs to generate an ensemble forecast. Thus, the following section presents a proposed framework for A-SV-based ensemble generation with Pangu-Weather. Finally, this chapter concludes with the introduction of common ensemble evaluation metrics. These are used in the subsequent analysis to evaluate the forecast skill of the A-SV-based Pangu-Weather ensemble.

3.1 ERA5

ERA5 is the fifth generation reanalysis data set of the ECMWF (Hersbach et al., 2020). The data set provides a detailed analysis of past atmospheric states from 1940 onward. For this, a previous version of the ECMWF IFS is used for the computation. The data is available on a 0.25° equiangular grid with a horizontal resolution of 31 km. ERA5 has 137 vertical levels that stretch from the surface to 1 hPa, corresponding to an altitude of 80 km. The temporal resolution of ERA5 is hourly. 4D-Var (Le Dimet and Talagrand, 1986) is used to assimilate observational data into the ERA5 system. The assimilation window of the 00 UTC forecast extends from 21 UTC on the previous day to 09 UTC. For the 12 UTC forecast, the data assimilation window stretches between 09 and 21 UTC.

MLWP models commonly use ERA5 as their training data set (e. g., Bi et al., 2023; Lang et al., 2024). Often, these MLWP models omit parts of the ERA5 data set with valid times before 1979 during training (e. g., Keisler, 2022; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Kochkov et al., 2024; Lang et al., 2024; Price et al., 2024). In addition, ERA5 is used to validate the ensemble forecasts of MLWP models (e. g., Bülte et al., 2025; Baño-Medina et al., 2025). ERA5 is not a direct assessment of the atmospheric state. It is rather the best estimate of the past atmospheric state using the available observational data. Hence, ERA5 cannot perfectly describe past atmospheric states and, therefore, cannot be considered the ultimate truth. The quality of ERA5 data depends on the availability and quality of the observational data.

3.2 Pangu-Weather

After decades of continuous minor improvements in the forecast quality of NWP models (Bauer et al., 2015), the beginning of the third decade of the 21st century saw rapid progress in MLWP models (Ben Bouallègue et al., 2024). Pangu-Weather was among the first models to demonstrate competitive forecast skill compared to the leading physics-based NWP models (Bi et al., 2023).

Pangu-Weather is a deep neural network based on a Vision-transformer structure that is commonly used in image recognition (Liu et al., 2021; Dosovitskiy et al., 2021). Pangu-Weather is trained on the ERA5 data set (Hersbach et al., 2020), with the training period spanning from 1979 to 2017. The model is tested on data from 2018 and validated on data from 2019 (Bi et al., 2023). Four neural networks were trained to make 1, 3, 6 and 24 h forecasts. The Pangu-Weather temporal aggregation algorithm combines the different forecast models to generate hourly forecasts with minimum forecast model inferences required (Bi et al., 2023). The forecasts retain the native 0.25° equiangular grid of ERA5. Pangu-Weather can generate forecasts for five upper atmosphere variables, zonal U and meridional V wind, temperature T, specific humidity q, and geopotential Z, at 13 pressure levels ranging from 1,000 to 50 hPa. In addition, Pangu-Weather provides prognostic forecasts for four surface variables. These include the 10 m zonal Wind (U10M), 10 m meridional Wind (V10M), T2M, and surface pressure p_{sfc} (Bi et al., 2023).

To make use of the Vision transformer, Pangu-Weather assumes that the spatial patterns of the state vector of an atmospheric variable at an arbitrary pressure level resemble a picture (Bi et al., 2023). From this large picture of the atmosphere, small patches are taken via patch embedding. These are down-sampled and fed to the Vision transformer (Bi et al., 2023). The deep neural network then connects these initial patches with the predicted patches. The predicted patches are then reinflated to the native 0.25° equiangular grid with a patch recovery operation (Bi et al., 2023). This down-sampling and reinflating process of the deep neural network leads to a significant denoising of Pangu-Weather forecasts (Selz and Craig, 2023).

The deep neural network structure allows for rapid inference times of Pangu-Weather forecasts on Graphical Processing Units (GPUs). Bi et al. (2023) indicate that the inference time of a Pangu-Weather forecast is more than 10,000 times faster compared to ECMWF IFS forecasts.

3.2.1 Arnoldi method in Pangu-Weather

L-SV computation in NWP models uses a coarser resolved tangent linear model, which approximates the initial model trajectory (see Equation 2.6), and also an adjoint model. This leads to manageable computational costs of the L-SV computation, making it feasible in an operational setting, but on the other hand it needs the development and maintenance of these additional models. In contrast, the Arnoldi method for SV computation does not need

a linear or adjoint model, only forecasts. Hence, in the case of NWP models, the full non-linear model is usually needed. Repeated forward integrations with the full non-linear NWP models are expensive and challenging in an operational setting. Although parallelization and running a low resolution model for SV generation itself might be a solution therefor, the situation becomes much more comfortable for MLWP models, since full non-linear forecasts become very cheap. Hence, the Arnoldi method for SV calculation is much more advantageous if used with MLWP models. In addition, the usage of the full non-linear model in the Arnoldi method allows a wider choice of the OTI because Arnoldi iterations can take non-linear developments into account since the **H** matrix is built on the basis of full forecasts (see Equation 2.11).

Parameter choices before computation

With the Arnoldi method, the following parameters must be selected before the computation. First, the OTI must be chosen. The choice is not limited by any linearization assumption of the forecast model. However, the secant approximation, i. e., the direct connection of initial to final time perturbations, can lead to significant deviations for very long OTIs (Winkler and Denhard, 2025). Non-trivial is the choice of the perturbation amplitude h. h scales the initial perturbation vector v. Winkler and Denhard (2025) denote that for real forecast models h should be in the magnitude of 100 to 10,000. Due to the large size of model states, the norm of an increment of two nearby states will result in a rather large numerical value, typically located in the named range. For the analysis in this work, $h = 4{,}000$ is selected for global SV optimization. In addition, the cut-off number of SVs derived from the **H** matrix must be chosen. To ensure mathematical consistency of A-SVs, the final SVs of **H** should not be used for further computation (Winkler and Denhard, 2025). m should be at least 4/3 times larger than the desired number of SVs used for further computation as an estimate of the surplus size of the **H** matrix. Winkler and Denhard (2025) show that less than one third of A-SVs has a singular value $\sigma_i > 1$. More details on the Arnoldi method for SV calculation can be found in Winkler et al. (2020) and Winkler and Denhard (2025).

Block Arnoldi method in Pangu-Weather

To increase the computational efficiency of A-SV computation, the following analysis utilizes the Block Arnoldi method. Perturbation vectors are initialized with pure white noise. U, V, and T are perturbed at all pressure levels. The resulting perturbations and their growth are measured by the dry total energy norm given by Equation 2.7. The Arnoldi method omits the perturbation of the Mean Sea Level Pressure (MSLP) field due to the negligible general contribution of the surface pressure term p'_{sfc} to the dry total energy norm (see Figure 2.3). The initial conditions of the Deutscher Wetterdienst (DWD) operational ICOsahedral Nonhydrostatic (ICON) NWP model (Zängl et al., 2015) can be given to Pangu-Weather as input. This requires re-gridding the ICON initial conditions from their native icosahedral grid to the 0.25° equiangular grid of Pangu-Weather. As an extension of the work of Winkler and Denhard (2025), this work has implemented the option for ERA5 data

to be used as initial conditions. This allows A-SV computation for any valid time in the ERA5 data set. The following analysis focuses exclusively on A-SVs computed from ERA5 initial conditions.

3.2.2 Ensemble generation with Pangu-Weather

The computationally inexpensive forecast inference of Pangu-Weather forecasts opens the opportunity for low-cost ensemble generation with Pangu-Weather and A-SVs. The proposed framework aims to mimic the basic characteristics of the operational ECMWF IFS ensemble initialization based on L-SV perturbations (see Section 2.5).

Similarly to the ECMWF IFS ensemble forecast, the proposed A-SV-based Pangu-Weather ensemble consists of 50 members and an unperturbed deterministic forecast. Section 3.2.1 details the necessary parameter choices for the Block Arnoldi method. The OTI is chosen to be 48 h as in the ECMWF IFS ensemble forecast. The A-SVs are globally optimized with an amplitude $h = 4{,}000$. To generate a 50-member ensemble, at least 25 A-SVs must be computed to create 25 independent linear combinations. Thus, the cutoff value of the $\mathbf{H}$ matrix is set to 25. As Section 3.2.1 discusses, the size of the $\mathbf{H}$ matrix must be at least 4/3 times greater than the desired amount of SVs. To accommodate this constraint, the block size of the perturbation vectors $b = 5$ and the loops of the Arnoldi method $l = 7$. This results in a Krylov subspace / $\mathbf{H}$ matrix with a dimensional size of $m = l \cdot b = 35$.

The 25 A-SVs are linearly combined to generate 25 ensemble perturbations. The A-SVs v_i are multiplied by random number coefficients α_i, as given by Equation 2.12. The coefficients α_i of the linear combination are sampled from a truncated Gaussian distribution ($[-3\sigma; 3\sigma]$) with a mean $\mu = 0$ and a standard deviation $\sigma = 1$, as in the ECMWF IFS ensemble forecast (Leutbecher and Palmer, 2008). As mentioned, the limited sample of the Gaussian distribution most often leads to a non-zero sample mean of the coefficients α_i. The subsequent bias in the initial time ensemble perturbations would initially deteriorate the ensemble forecast (Leutbecher and Palmer, 2008). To enforce a zero-mean of the ensemble perturbations at initial time, the initial conditions are positively/negatively perturbed by each ensemble perturbation. However, the magnitude of the ensemble perturbations after the linear combination is too large and would lead to unphysical perturbations of the initial conditions. Hence, the ensemble perturbations are scaled with an a posteriori amplitude scaling factor $\gamma = 0.2$ to adjust the magnitude of the ensemble perturbations to the typical analysis error ($\Delta U, \Delta V = 2\,\mathrm{m\,s^{-1}}$, $\Delta T = 2\,\mathrm{K}$, Pereira and Berre, 2006; Leutbecher and Palmer, 2008). The choice of the amplitude scaling factor is analyzed in more detail in Section 6.1.

As Section 2.5 outlines, the L-SV-based ensemble perturbations only aim to cover the initial condition error induced by NWP models (Leutbecher and Palmer, 2008). Thus, this method does not account for model error. For MLWP models, there are certainly different sources of model error present than in physics-based NWP models (Selz and Craig, 2023). It is beyond the scope of this work to finalize a strategy for incorporating model error into the proposed ensemble initialization scheme. In the case of Pangu-Weather, an obvious choice would be to compute separate sets of A-SVs using both Pangu-Weather's 6 h forecast

model (Pangu6) and Pangu-Weather's 24 h forecast model (Pangu24) as the full non-linear model **M** in Equation 2.11. Additionally, the OTI of A-SVs could be varied. Section 4.2 outlines that different Pangu-Weather forecast models show systematic differences in the A-SV perturbations. Hence, this could be a convenient approach to induce model-related errors in the ensemble perturbations.

The A-SV-based Pangu-Weather ensemble is implemented with the aim of being as flexible as possible. An operational use on, e. g., the ICON output is possible. The quality of the Pangu-Weather ensemble is evaluated in Chapter 6, including a comparison to other Pangu-Weather-based ensemble approaches by Bülte et al. (2025).

3.3 Ensemble evaluation metrics

The following section introduces ensemble evaluation methods to assess the forecast skill of the A-SV-based Pangu-Weather ensemble. The mean $\bar{\chi}$ of forecast variable χ of an ensemble with N ensemble members is given by the following relation

$$\bar{\chi} = \frac{1}{N} \sum_{n=1}^{N} \chi_{i,j,n,t}. \tag{3.1}$$

With $\bar{\chi}$, the spread of an ensemble can be derived as follows

$$\mathrm{Var}(\chi) = \frac{1}{N-1} \sum_{n=1}^{N} (\chi_{i,j,n,t} - \bar{\chi})^2. \tag{3.2}$$

Similarly, the Root Mean Squared Error (RMSE) of an ensemble can be derived using the following relation

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\bar{\chi} - \psi)^2}, \tag{3.3}$$

where ψ is an observation. In this context, ψ is the reference data set that is assumed to be the ground truth. For a well calibrated ensemble, RMSE and spread should match (Fortin et al., 2014). In this case, high spread values of the ensemble would indicate the potential for large errors, i. e., large RMSE values, and vise versa. For a forecaster, a sharp forecast, i. e., small spread values, is desirable. However, if the forecast is neither reliable (i. e., the forecast probabilities match the observed occurrence frequencies) nor has good resolution (i. e., the ability of the forecast to correctly discriminate from the climatological occurrence frequency), it is of no great benefit to the forecaster. In ensemble forecasting, a commonly used metric to evaluate both the reliability and resolution of an ensemble forecast is the CRPS (Hersbach, 2000). The CRPS can be defined as

$$CRPS(F, y) = \int \left(F(y) - \theta(y - x)^2 \right) \tag{3.4}$$

following Matheson and Winkler (1976). x is an observation, and $F(y)$ is the cumulative distribution function of an empirical probabilistic forecast. $\theta(x)$ is the Heaviside step function that is zero for negative arguments and $\theta(x) = 1$ for $x \geq 0$. The Continuous Ranked Probability Skill Score (CRPSS) is a measure used to analyze the forecast skill of an ensemble forecast $CRPS_{\mathrm{f}}$ against a reference forecast $CRPS_{\mathrm{ref}}$. The CRPSS of an ensemble is given by the following relation

$$CRPSS = 1 - \frac{CRPS_{\mathrm{f}}}{CRPS_{\mathrm{ref}}}. \tag{3.5}$$

Positive CRPSS values imply that the ensemble forecast is beneficial and demonstrates forecast skill over the reference forecast. $CRPSS = 0$ signals that the ensemble forecast performs equally well/poorly as the reference forecast. If $CRPSS < 0$, the ensemble forecast is worse than the reference.

4 Structure of Arnoldi-SVs in Pangu-Weather

The theoretical background of L-SV computation in NWP models and SV computation with the Arnoldi method is outlined in Chapter 2. Section 2.3 provides the reference framework for comparing A-SVs in Pangu-Weather with L-SVs in physics-based NWP models. The discussion is limited to three key features of L-SVs. In physics-based NWP models, L-SV perturbations exhibit a westerly tilt with height at initial time. At final time, these shift to a slight easterly tilt with height. In addition, there is a distinct upward propagation of the perturbation energy maximum from initial to final time. Secondly, L-SV perturbations in physics-based NWP models feature a characteristic shift from predominantly potential perturbation energy at initial time to predominantly kinetic perturbation energy at final time. Third and finally, L-SV perturbations show upscale perturbation growth in physics-based NWP models during forward integration.

Due to the fundamentally different inference method of A-SVs (see Section 2.4.2), the Arnoldi method sets up a different Krylov subspace (Winkler and Denhard, 2025). Thus, different SVs are derived using the Arnoldi method. In the same order as in Section 2.3, first, the vertical structure of A-SVs in Pangu-Weather is analyzed in the following section. In addition, we assess the physical consistency of A-SV perturbations in Pangu-Weather. Secondly, the energy partition of A-SVs in Pangu-Weather is presented. And third, the horizontal scale of A-SV perturbations in Pangu-Weather is evaluated. The analysis aims to highlight the similarities and differences of SV perturbations in both different model architectures and different inference methods of the SVs itself.

For all A-SV-perturbed forecasts in this analysis, the respective parameter choices are denoted in Section 3.2.2. The following analysis is restricted to the ten leading A-SVs computed from the ERA5 initial conditions of 01 January 2022, 00 UTC. The unique structure of Pangu-Weather, with its four trained models for different inference steps (Bi et al., 2023), allows, in combination with the A-SV algorithm, to use different forecast models and vary OTIs. The following analysis includes four different combinations of Pangu-Weather forecast models and OTIs. This also provides an indication of both the model dependence and OTI dependence of A-SVs in Pangu-Weather. The analysis below contains the combination of Pangu24 with OTIs of 24 and 48 h, and the combinations of Pangu6 with OTIs of 6 and 24 h. For this analysis, the Pangu-Weather temporal aggregation algorithm is omitted. Hence, the chosen forecast model $\mathbf{M}$ in the Arnoldi method (see Equation 2.11) is also responsible for all forecasts considered in the analysis. This implies that Pangu24 forecast combinations have a temporal resolution of 24 h and Pangu6 forecast combinations have a temporal resolution of 6 h.

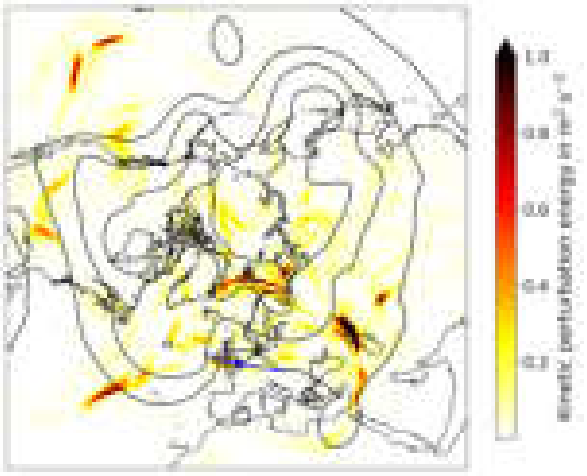
4.1 Vertical structure of Arnoldi-SVs in Pangu-Weather

Section 2.3.1 details the vertical structure of L-SV perturbations in physics-based NWP models. In the following section, first, the cross-sections of initial time A-SV perturbations and, secondly, the final time perturbation cross-sections are analyzed. These are also compared to the respective cross-sections of L-SV perturbations in physics-based NWP models. For the analysis below, Pangu24 is used as the forecast model in Equation 2.11, and the OTI is set to 48 h. The forecasts are initialized from the ERA5 data valid for 01 January 2022, 00 UTC. The A-SV perturbations are scaled by a factor $\gamma = 0.2$.

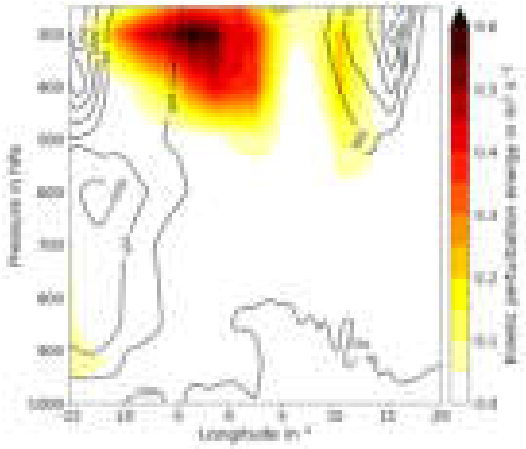
4.1.1 Initial time Arnoldi-SV perturbations

The kinetic perturbation energy is used to identify dynamically active regions of the atmosphere. The kinetic perturbation energy is determined by the difference in kinetic energy between the unperturbed and perturbed model forecasts. Regions with high kinetic perturbation energy values highlight areas sensitive to A-SV perturbations. In Figure 4.1(a), the 300 hPa kinetic perturbation energy is displayed. The 300 hPa pressure level is chosen because it is located near the jet level in the mid-latitudes. In addition, the 500 hPa Geopotential (Z500) is shown to provide an indication of the general synoptic set-up at initial time. The vertical cross-section is taken in an area with high kinetic perturbation energy values along a slanted line spanning southeastward from the British Isles into Eastern Central Europe. This area is located in the transition region between a large trough to the west over the Northern Atlantic and a ridge to the east that expands across Western Europe. The vertical cross-section extends mainly into the influence region of the ridge, as seen in Figure 4.1.

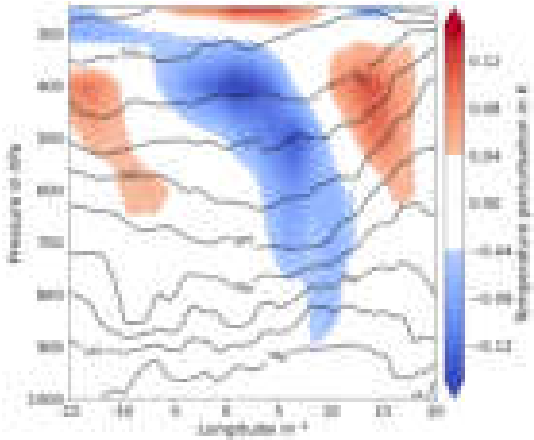
In the vertical cross-section of the kinetic perturbation energy, the highest values are located at 300 hPa west of the Greenwich meridian. Hence, the largest kinetic perturbation energy is found in the transition zone between the trough and the ridge. In the eastern part and at lower elevations, the kinetic perturbation energy is much smaller. Similar results are found for L-SVs in physics-based NWP models (Pereira and Berre, 2006). In addition to the kinetic perturbation energy, the potential perturbation energy is the second contributor to the total perturbation energy (see Equation 2.7). The vertical cross-section of the temperature perturbation in Figure 4.1(c) shows vertical perturbation patterns with a clear westward tilt with height. A similar westward tilt with height is also observed for initial time L-SV perturbations in physics-based NWP models (see Figure 2.1). Figure 4.1 reveals that the temperature perturbations are physically consistent with the patterns of the kinetic perturbation energy. A negative temperature perturbation is located beneath the kinetic perturbation energy maximum. A positive temperature perturbation is situated above. In the upper troposphere, there seems to be a correlation between the gradient of the kinetic perturbation energy and the temperature perturbation. Vise versa, high gradients of the temperature perturbation tend to correlate with large kinetic perturbation energy values.



(a) 300 hPa kinetic perturbation energy. Z500 is indicated by the black contour lines at 50,400, 52,800, 55,200 and 57,600 $m^2\ s^{-2}$.



(b) Vertical cross-section of the kinetic perturbation energy. Black contour lines show the kinetic energy of the unperturbed forecast at 100, 500, 750, 1,000, 1,250 and 1,500 $m^2\ s^{-2}$.



(c) Vertical cross-section of the temperature perturbation. Black lines indicate the potential temperature of the reference forecast with iso-lines every 4 K between 284 and 324 K.

Figure 4.1: Hemispheric overview of the kinetic perturbation energy along with vertical cross-sections of the kinetic perturbation energy and the temperature perturbation of the leading A-SV at initial time (01 January 2022, 00 UTC). A-SVs are optimized globally for an OTI of 48 h with the dry total energy norm. The vertical cross-sections are taken along the blue line in Figure 4.1(a). The arrow indicates the orientation of the cross-sections.

The kinetic perturbation energy is a quadratic function of the difference between the kinetic energies of the unperturbed reference forecast and the perturbed forecast. Therefore, the kinetic perturbation energy does not provide any information on which of the two is larger. In contrast, the temperature perturbation is a linear function. Thus, it indicates the direction of the perturbation. The potential temperature is included in the analysis to assess whether the temperature perturbation and the kinetic perturbation energy are in thermal wind balance. In the upper troposphere, where the largest perturbations are located,

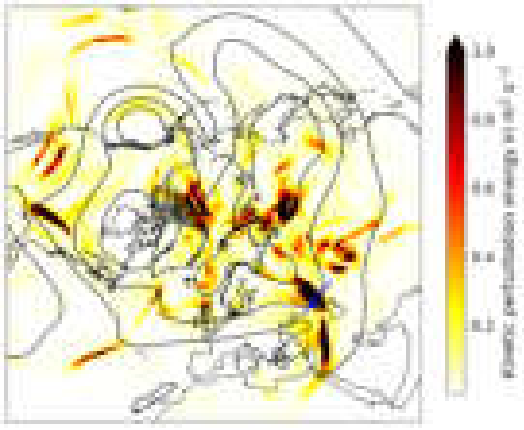
the temperature perturbations decrease the horizontal temperature gradients of the A-SV-perturbed forecast. The thermal wind balance implies that, due to the weaker horizontal temperature gradients, the thermal wind in the A-SV-perturbed forecast is weaker than in the unperturbed reference forecast. Indeed, the kinetic energy of the A-SV-perturbed forecast is reduced. The negative kinetic energy perturbation is co-located with high kinetic perturbation energy values in the upper troposphere (not shown). Hence, the A-SV perturbations are consistent with the thermal wind balance. This highlights that there is a clear consistency between the initial time kinetic perturbation energy and the temperature perturbations of A-SVs in Pangu-Weather.

In the lower troposphere, the scales of both the kinetic perturbation energy and the temperature perturbation are much smaller. Below 600 hPa, the westward tilt of the temperature perturbation vanishes. In contrast, temperature perturbations of L-SVs in physics-based NWP models have a maximum around 500 hPa (see Figure 2.1). Thus, the maximum temperature perturbation in Pangu-Weather is located at a higher vertical level. Near the surface, L-SVs in physics-based NWP models also tend to exhibit a reduced westward tilt with height. In Figure 4.1(c), the westward tilt with height of the temperature perturbation completely vanishes near the surface. However, the key features of L-SV perturbations at initial time, known from physics-based NWP models, are also found for A-SVs in Pangu-Weather.

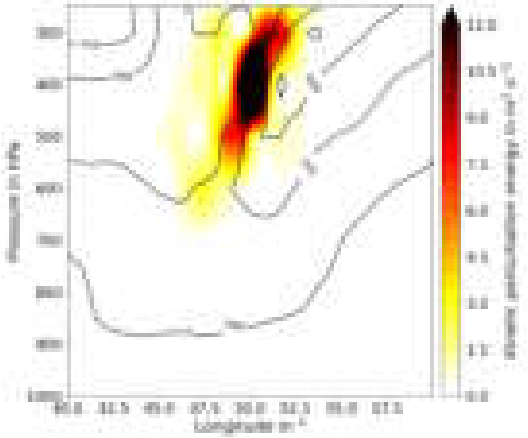
4.1.2 Final time Arnoldi-SV perturbations

Similarly to the initial time A-SVs in Pangu-Weather, the final time A-SV perturbations are also compared to final time L-SV perturbations in physics-based NWP models. For synoptic consistency, the final time vertical cross-section is taken through the same ridge as the initial time vertical cross-section. During the OTI, the ridge propagated eastward across Europe due to a strong westerly flow (see Z500 lines in Figure 4.2(a)). In addition, the axis of the ridge has tilted to the south. For this reason, the cross-latitudinal component of the final time vertical cross-section is much larger. As indicated, the vertical cross-section is taken in an area of high kinetic perturbation energy in the transition region between the ridge over Turkey and a large trough over Central Asia. The vertical cross-section stretches from Central Eastern Europe to the Middle East. The different scalings of the color-bar of the kinetic perturbation energy cross-sections show that the A-SV perturbations have grown significantly from initial time. Similarly, the amplitudes of the temperature perturbations have also increased significantly from initial to final time.

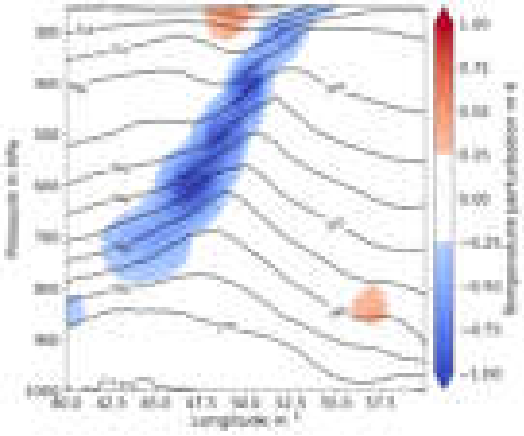
The vertical cross-section of the kinetic perturbation energy shows a large signal east of 36 °E. The coherent structure is tilted eastward with height. A similarly significant temperature perturbation is apparent in Figure 4.2(c). In contrast to initial time, the temperature perturbation and kinetic perturbation energy patterns are co-located. For L-SVs in physics-based NWP models, the kinetic and potential perturbation energies are not co-located and, thus, are unbalanced at initial-time. At final-time, the kinetic and potential perturbation energies are co-located and thus balanced (Montani and Thorpe, 2002). Figure 4.2 indicates that similar balancing processes of A-SV perturbations are also present in Pangu-Weather. In



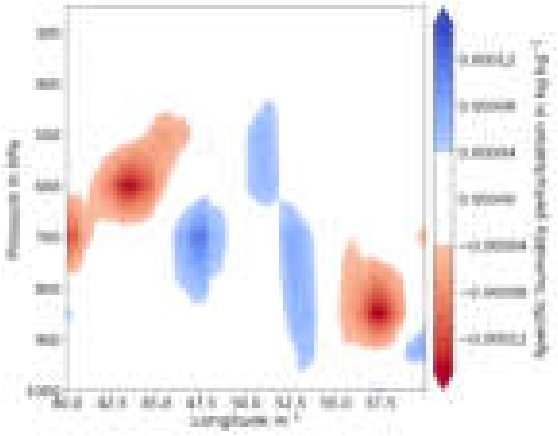
(a) 300 hPa kinetic perturbation energy. Z500 is indicated by the black contour lines at 50,400, 52,800, 55,200 and 57,600 $m^2\ s^{-2}$.



(b) Vertical cross-section of the kinetic perturbation energy. Black contour lines show the kinetic energy of the unperturbed forecast at 100, 250, 500 and 750 $m^2\ s^{-2}$.



(c) Vertical cross-section of the temperature perturbation. Black contour lines indicate the potential temperature of the unperturbed forecast every 4 K between 272 and 324 K.



(d) Vertical cross-section of the specific humidity perturbation

Figure 4.2: Hemispheric overview of the kinetic perturbation energy along with vertical cross-sections of the kinetic perturbation energy, temperature perturbation, and specific humidity perturbation of the leading A-SV at final time (03 January 2022, 00 UTC). A-SVs are optimized globally for an OTI of 48 h with the dry total energy norm. The vertical cross-sections are taken along the blue line in Figure 4.2(a). The arrow indicates the orientation of the cross-sections.

the perturbed forecast, the negative temperature perturbation leads to a westward displacement of the frontal boundary. The associated jet feature, east of the frontal region, shifts in the perturbed forecast with the frontal region to the west. Hence, the kinetic perturbation energy is induced by a horizontal shift of the jet feature in the upper troposphere.

The shift to an eastward tilt with height of the temperature perturbation at final time aligns, at least in part, with the known characteristics of L-SVs in physics-based NWP models (Diaconescu et al., 2012). At final time, Figure 2.1 shows upright vertical temperature

perturbations. In Pangu-Weather, the easterly tilt of the temperature perturbation is more pronounced at final time. In addition, there is no upward propagation of the vertical temperature perturbation maximum. Figure 4.2(c) indicates only a minor downward shift of the temperature perturbation maximum. For L-SVs in physics-based NWP models, the vertical maximum of the perturbation energy shifts upward (see Figure 2.2).

Until now, the analysis focuses exclusively on the variables that were initially perturbed by A-SVs. The specific humidity is not perturbed at initial time (see Equation 2.7). Therefore, it freely adjusts to the perturbations of the other forecast variables. Previous studies have shown that Pangu-Weather generally adjusts well to initial condition perturbations, with some caveats (Hakim and Masanam, 2024). The broad positive temperature perturbation situated around 50 °E exhibits an overlap with positive specific humidity perturbations. Following the Clausius-Clapeyron equation, warmer air can have a higher specific humidity before reaching saturation. In the upper troposphere, the amplitude of the specific humidity perturbations decreases, which is expected because of the generally low humidity content in the upper troposphere. East of the dominant negative temperature perturbation, Figure 4.2(d) shows a positive specific humidity perturbation that extends upward. Negative specific humidity perturbations are located west of the negative temperature perturbation, creating a dipole structure. Hence, the A-SVs likely induce a shift in the moisture advection pattern ahead of the surface cold front related to the trough over Central Asia. Although not all specific humidity perturbations can be directly associated with temperature perturbations, Figure 4.2 shows significant cross-variable consistency between specific humidity and temperature perturbations.

The previous section highlights that A-SVs in the ML-based Pangu-Weather model generate vertical perturbation structures consistent with the known vertical structures of L-SV perturbations in physics-based NWP models. Initial time A-SV perturbations tilt westward with height, consistent with baroclinic disturbances. At final time, the tilt of the A-SV perturbations shifts to an eastward vertical tilt, which is more pronounced than in physics-based NWP models. Simple theoretical assumptions, such as the thermal wind balance, prove that A-SV perturbations in Pangu-Weather exhibit cross-variable consistency. In addition, A-SV perturbations are situated in parts of the atmosphere that are favored for significant perturbation growth.

4.2 Perturbation energy of Arnoldi-SVs in Pangu-Weather

Section 4.1 shows that A-SVs in Pangu-Weather have vertical structures similar to L-SVs in physics-based NWP models. The following section analyzes the perturbation energy of A-SVs in Pangu-Weather. Therefore, we evaluate both the composition and the vertical profiles of the perturbation energy. In addition, the latitudinal dependencies of both are highlighted. For the analysis, the diagnostic of the Difference Kinetic Energy (DKE) is used (Selz et al., 2022; Selz and Craig, 2023). DKE is defined as follows

$$DKE = \frac{\mathrm{Var}(u) + \mathrm{Var}(v)}{2}, \tag{4.1}$$

where $\mathrm{Var}(u), \mathrm{Var}(v)$ are the respective variances of the zonal wind u and the meridional wind v in the ensemble dimension. For this analysis, the ensemble dimension is defined as the respective A-SV-perturbed forecast of the leading ten A-SVs.

Extending upon the definition of the DKE, the Difference Potential Energy (DPE) is determined by the following relation

$$DPE = \frac{\mathrm{c_p}}{T_{\mathrm{ref}}} \cdot \mathrm{Var}(T), \tag{4.2}$$

where $\mathrm{Var}(T)$ is the temperature variance in the ensemble dimension. $\mathrm{c_p}$ is the isobaric specific heat capacity of air, and T_{ref} is the reference temperature given by the unperturbed forecast. The sum of these perturbation energies is defined as the Difference Total Energy (DTE),

$$DTE = DKE + DPE, \tag{4.3}$$

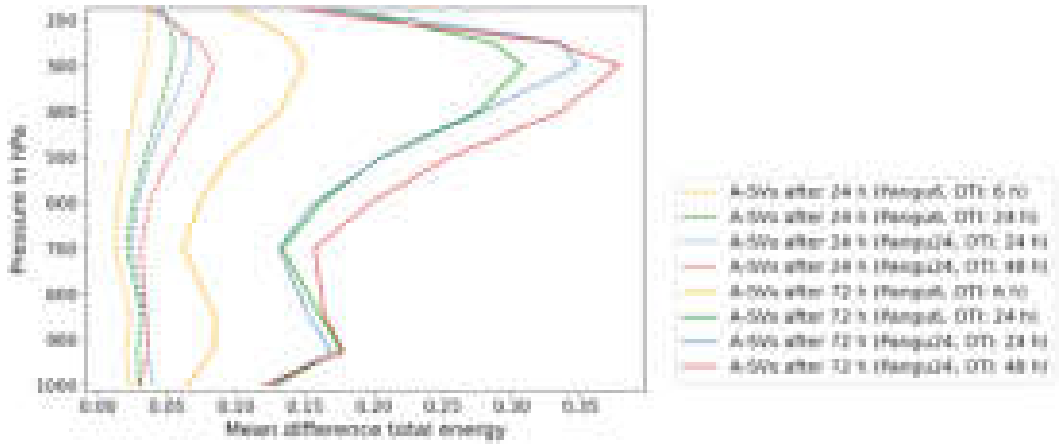
of the forecast. With these diagnostics, the energy partition in terms of the Fraction of Difference Kinetic Energy (FrDKE) of A-SVs in Pangu-Weather is specified as follows

$$FrDKE = \frac{DKE}{DTE}. \tag{4.4}$$

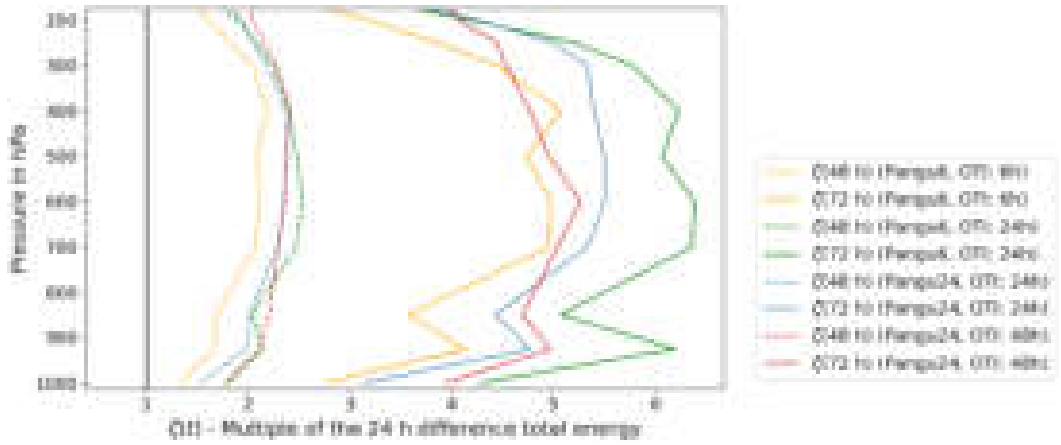
Figure 2.3 shows that in physics-based NWP models, L-SVs exhibit a characteristic shift in the energy partition. The predominant potential perturbation energy at initial time is converted into predominantly kinetic energy at final time. The respective ratios shift from a 3:1 ratio of potential to kinetic perturbation energy at initial time to a 1:3 ratio at final time. From a mathematical point of view, this change can be interpreted as the shift of the off-attractor initial time L-SVs onto the attractor of the system (Szunyogh et al., 1997; Kalnay, 2002). The following analysis will include both Pangu6 and Pangu24, as well as different OTIs detailed above, to investigate potential influences of the OTI and the two tested Pangu-Weather forecast models on the A-SVs.

4.2.1 Difference total energy

Figure 4.3 includes all the combinations of OTIs and Pangu-Weather forecast models tested. In all cases, A-SV perturbations grow rapidly, as indicated by the strong increase of DTE with lead time. However, significant differences between the combinations of OTIs and Pangu-Weather forecast models are apparent in Figure 4.3. At a lead time of 72 h, the combination of an OTI of 48 h and Pangu24 gives the largest perturbation growth. For all combinations of OTIs and Pangu-Weather forecast models, the maximum DTE is located near the jet level (around 300 hPa). Figure 4.3(a) shows that the DTE vertical profiles are similar to the total perturbation energy of L-SVs in physics-based NWP models. Baumgart et al. (2019) show that perturbation growth at early lead times is dominated by the propagation of mesoscale disturbances along the mid-latitude jet stream. The high DTE values around 300 hPa suggest that similar mechanisms may also be active in Pangu-Weather. Figure 4.3(a) shows significantly lower global-mean DTE values for the combination of Pangu6 and an OTI of 6 h than for all other forecast combinations. The DTE of this combination at a lead



(a) DTE



(b) Multiple of the DTE $\zeta(t)$. The vertical black line indicates $\zeta = 1$.

Figure 4.3: Vertical profiles of the DTE of the leading ten A-SVs for different combinations of Pangu-Weather forecast models and OTIs. Vertical profiles for the combination of Pangu24 and an OTI of 24 h are shown in blue, for Pangu24 and an OTI of 48 h in red, for Pangu6 and an OTI of 6 h in yellow, and, for Pangu6 and an OTI of 24 h, the vertical profiles are given by the green lines. All forecasts are initialized on 01 January 2022, 00 UTC with globally optimized A-SVs with respect to the dry total energy norm. Vertical profiles of the DTE at a lead time of 24 h and vertical profiles of $\zeta(48\,\mathrm{h})$ are indicated by dashed lines, vertical profiles of both at a lead time of 72 h are shown by the continuous lines.

time of 72 h is similar to the DTE of the other forecast combinations at a lead time of 48 h (not shown).

In the lower troposphere, the DTE develops a secondary maximum located at 925 hPa. A similar secondary maximum near the top of the boundary layer is also observed for L-SVs in physics-based NWP models (see Figure 2.2). The similarity of both vertical profiles supports the hypothesis that, both in Pangu-Weather and in physics-based NWP models, SVs trigger similar perturbation growth mechanisms. For an OTI of 24 h, the DTE growth of Pangu6 and Pangu24 can be compared. Both have very similar vertical profiles (see Figure 4.3(a)). Compared to the forecast combination of Pangu24 and an OTI of 48 h, the

relative magnitudes of the primary and secondary DTE maxima differ between the forecast models. Figure 4.3(a) shows that for Pangu24, the primary DTE maximum at jet level is more pronounced compared to Pangu6. In addition, the maximum is shifted upwards. Contrasting observations can be made regarding the secondary DTE maximum near the top of the boundary layer. Here, the secondary DTE maximum is more pronounced for Pangu6. At their more prominent maxima, the DTE values reach a magnitude similar to the DTE of the forecast combination of Pangu24 and an OTI of 48 h.

The increase in global-mean DTE at lead time t, $DTE(t)$, relative to the global-mean DTE at a lead time of 24 h, $DTE(24\,\mathrm{h})$, is defined by the following relation

$$\zeta(t) = \frac{DTE(t)}{DTE(24\,\mathrm{h})}, \tag{4.5}$$

to quantify the perturbation growth of A-SVs in Pangu-Weather. For the various forecast combinations, significant variances of $\zeta(48\,\mathrm{h})$ and $\zeta(72\,\mathrm{h})$ are apparent in Figure 4.3(b). For all forecast combinations, the relative increase ζ is slightly above $2\,\mathrm{d}^{-1}$, which corresponds to a doubling time of around 24 h. In physics-based NWP models, the doubling time of L-SVs is less than 10 h, which shows that L-SV perturbations grow more rapidly in these models (Zadra et al., 2004). All vertical profiles of ζ reveal a broad mid-tropospheric maximum around 500 hPa. For $\zeta(72\,\mathrm{h})$, a secondary maximum develops around 925 hPa. A comparison between the different forecast combinations shows that the A-SV perturbations of the combination Pangu6 and an OTI of 6 h not only have a significantly lower DTE, but their relative increase ζ is also much smaller. In contrast, the DTE growth for the combination of Pangu6 and an OTI of 24 h is the largest. In addition, this forecast combination also has the most pronounced secondary maximum at 925 hPa.

For the Pangu6 forecast combinations, there is an opportunity to study the perturbation growth patterns for lead times of less than 24 h. For these, the relative increase ζ for a lead time t is defined as

$$\zeta(t) = \frac{DTE(t)}{DTE(6\,\mathrm{h})}. \tag{4.6}$$

For the combination of Pangu6 and an OTI of 24 h, the relative increase $\zeta(t) > 1$ at almost every pressure level for all lead times (not shown). This shows that A-SV perturbations grow for every forecast inference following the first one, similar to L-SVs in physics-based NWP models (Diaconescu and Laprise, 2012). The combination of Pangu6 and an OTI of 6 h produces perturbations with decaying DTE in the upper troposphere compared to the initial forecast inference ($\zeta(t) < 1$). For both forecast combinations, the relative increase in global-mean DTE is greater at vertical levels below 700 hPa than above (not shown).

Latitudinal dependence of the DTE

In addition to the global mean, the DTE can be examined with respect to latitudinal dependencies. Thus, the globe is subdivided into multiple latitudinal bands. For each hemisphere, the polar regions are defined to be all grid-points with a latitude φ of $|\varphi| > 75°$. The mid-latitudes of the respective hemisphere are defined as $35° < |\varphi| < 75°$. The subtropics

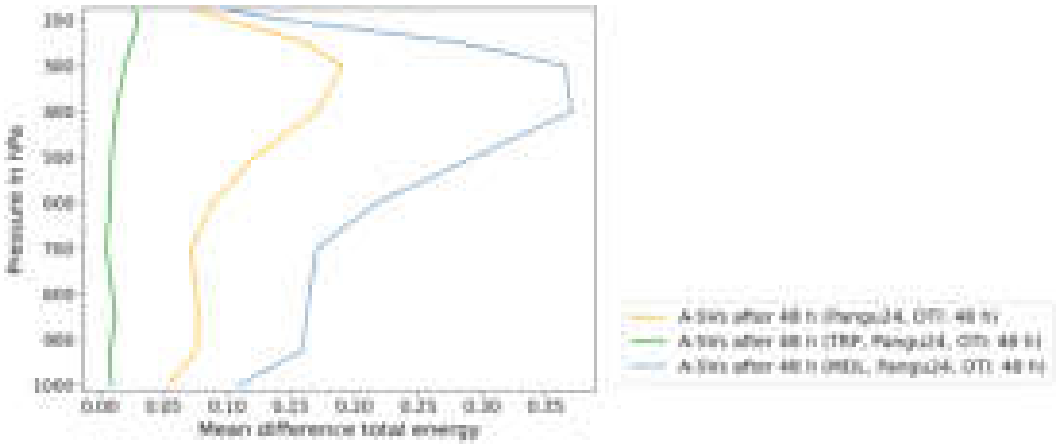


Figure 4.4: Vertical profile of the global-mean (yellow), the tropical-mean (TRP, green), and the northern mid-latitude-mean DTE (MDL, blue) at final time (03 January 2022, 00 UTC) of the leading ten A-SVs for forecasts initialized on 01 January 2022, 00 UTC. A-SVs are optimized globally for the forecast combination of Pangu24 and an OTI of 48 h with the dry total energy norm.

encompass all latitudes $15° < |\varphi| < 35°$. The tropics contain all grid-points with a latitude $|\varphi| < 15°$.

Figure 4.4 shows that, in the mid-latitudes, the domain-specific DTE mean is significantly higher than the global-mean DTE. The maxima at jet-level and at the top of the boundary layer are more pronounced in the mid latitudes. In addition, the maximum at jet level is located lower in the troposphere. In contrast, the tropics show a much smaller domain-mean DTE compared to the global-mean DTE. The low tropical DTE values suggest that similar perturbation growth mechanisms, as described by Baumgart et al. (2019), are active in Pangu-Weather because perturbation growth is focused on the mid-latitudes. Low tropical DTE values are related to low variances of u, v, and T in the tropics. In physics-based NWP models, peak L-SV perturbation growth is also typically found in mid-latitudes (Buizza and Palmer, 1995). Figure 4.4 shows that the upper-level DTE maximum in the tropics is located at 200 hPa, which is above the global-mean maximum. The latitudinally specific $\zeta(t)$ values show that, for the mid-latitudes, the DTE growth is larger than that of the global mean. In the tropics, contrasting behavior is again observed (not shown).

An intricate pattern is present in the respective subtropical vertical profiles. As mentioned, all vertical profiles are computed from forecasts initialized on 01 January 2022, 00 UTC. Hence, the southern hemisphere subtropics are located in the summer hemisphere, whereas the northern hemisphere subtropics are situated in the winter hemisphere. The northern hemisphere subtropics have higher DTE values than the global mean. The vertical profile resembles that of the mid-latitudes more strongly than that of the tropics (not shown). In contrast, the southern subtropics have much smaller domain-mean DTE values. In addition, the vertical profiles are much more akin to the tropical-mean vertical profile (not shown). This specific behavior can be attributed to the seasonal dependence of the Hadley cell. On average, the ascending branch of the Hadley cell is located in the summer hemisphere (Hoskins et al., 2020). Furthermore, the subtropical jet is stronger and is located closer to

the equator in the winter hemisphere (Hoskins et al., 2020). Hence, in January, the southern subtropical domain is strongly influenced by the ascending branches of the Hadley cells. For the northern hemisphere subtropics, the subtropical jet has a more significant influence on the domain. These general circulation features likely shape the hemispheric-specific vertical profiles of the DTE in the subtropics. The polar regions show the most erratic vertical profiles, possibly related to the polar vortex and interactions with the stratosphere (not shown).

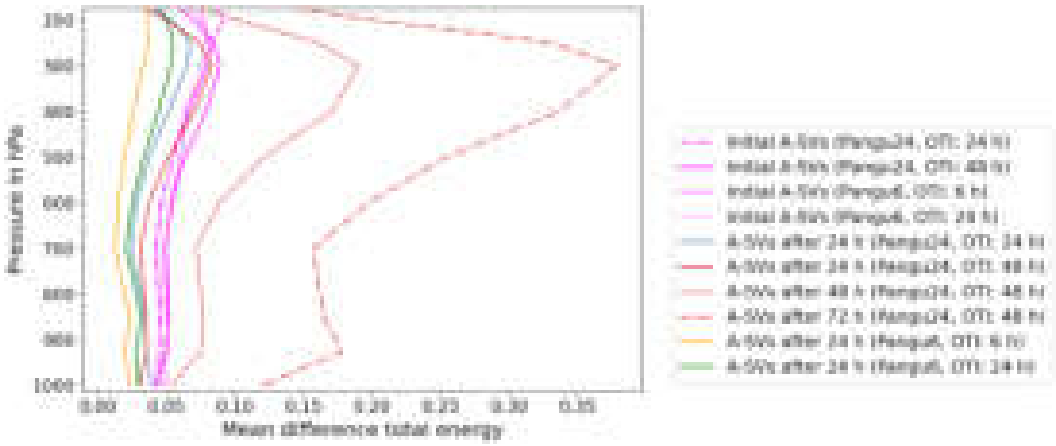
Perturbation growth relative to initial time

Figure 4.5 shows that at the initial time DTE is greater than the DTE at a lead time of 24 h. This observation pertains to all forecast combinations. Only the forecast combination of Pangu24 and an OTI of 48 h has a similar DTE in the upper part of the troposphere as the initial time DTE. According to Winkler and Denhard (2025), this signals that the chosen amplitude h is too large. It is important to note that Pangu24 inferences greatly smooth out perturbations (Selz and Craig, 2023). However, repeated forecasts with Pangu6 yield similar perturbation growth. At initial time, A-SV perturbations are still contaminated with white noise from the Arnoldi algorithm initialization. Due to the limited number of forward integrations in the Block Arnoldi method, remnants of this initial white noise transcend through the algorithm and impact the initial time A-SVs. White noise grants a large DTE at initial time due to the increased variance across the domain. This white noise induced DTE is subsequently removed by the smoothing of Pangu-Weather forecasts. Hence, DTE declines with the initial forecast inferences of Pangu-Weather. Aside from amplitude h, an increased size of matrix $\mathbf{H}$ can reduce the initial time white noise contamination of A-SVs. The analysis of Winkler and Denhard (2025) shows that, with their parameter choices, the amplitude of A-SVs grows slightly at a lead time of 24 h.

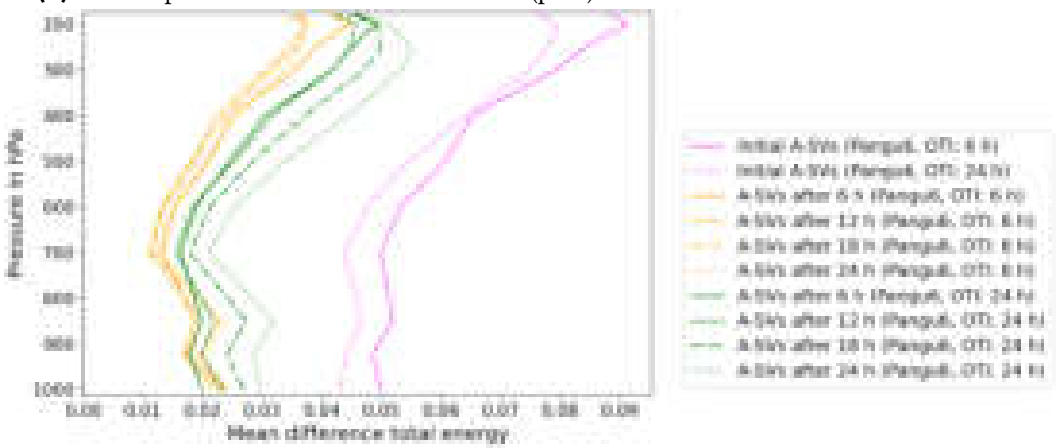
The Pangu6 forecast combinations shown in Figure 4.5(b) indicate that on day 1, the DTE remains below the respective initial time DTE values for all lead times. For an OTI of 6 h, the DTE is smaller for all lead times on day 1 than the DTE of the 6 h forecast for the forecast combination with an OTI of 24 h (with the exception of the lower troposphere). The relative increase of DTE is significantly lower for forecasts with an OTI of 6 h than for forecasts with an OTI of 24 h (not shown). The weak A-SV perturbation growth on day 1 may also suggest that the one-time shift of the off-attractor initial time A-SVs is dampened by the long forecast time of a single Pangu-Weather forecast.

4.2.2 Fraction of difference kinetic energy

The relative contributions of the DKE and DPE to the DTE are given by the FrDKE (see Equation 4.4). As mentioned, L-SVs in physics-based NWP models show a characteristic shift from a fraction of kinetic energy of around 0.25 at initial time to around 0.75 at final time (Diaconescu and Laprise, 2012). In the global-mean shown in Figure 4.6, the initial time FrDKE of A-SVs in Pangu-Weather is between 0.35 near the surface and around 0.5 in



(a) Vertical profiles of the DTE at initial time (pink) and for lead times between 24 and 72 h

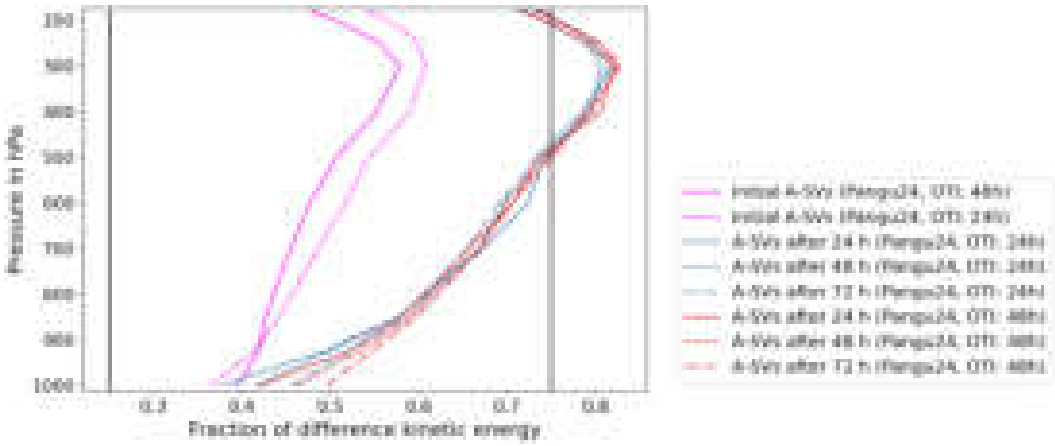


(b) Vertical profiles of the DTE at initial time (pink) and for lead times between 6 and 24 h

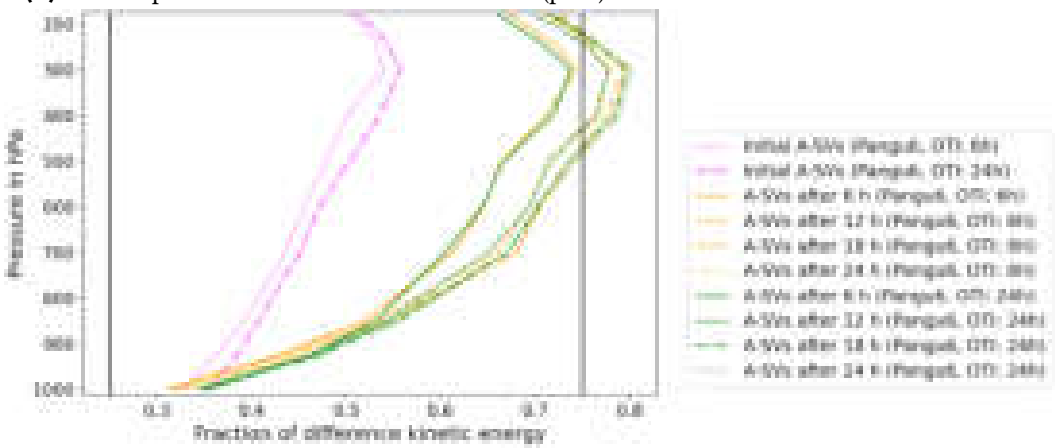
Figure 4.5: Vertical profiles of the DTE of the leading ten A-SVs at initial time (01 January 2022, 00 UTC) and for different lead times showing the growth of A-SVs. Vertical profiles for the forecast combination of Pangu24 and an OTI of 24 h are shown in dash-dotted pink at initial time and in blue for the forecast. For Pangu24 and an OTI of 48 h, vertical profiles are indicated by the continuous pink line at initial time and by a red line for the forecast. For Pangu6 and an OTI of 6 h, the initial time profiles are shown in dashed pink and in yellow for the forecast. Pangu6 and an OTI of 24 h vertical profiles are indicated by a dotted pink line at initial time and by green lines for the forecast. With the first lead time, the vertical profiles are shown by a continuous line, the second are shown with a dashed line, the third with a dash-dotted line, and the fourth with a dotted line (if shown).

the upper troposphere. This implies that DKE and DPE are of similar magnitude at the jet level. Thus, the initial time FrDKE is much higher for the A-SVs in Pangu-Weather than for the L-SVs in physics-based NWP models.

At final time, a significant shift towards larger FrDKEs is apparent. The vertical profiles in Figure 4.6 show that the increase of FrDKE for SVs in Pangu-Weather is largest at jet-level. At final time, the FrDKE here approaches values of 0.8. Near the surface, the FrDKE changes very little compared to initial time. The contrasting behavior of A-SVs may be caused by the different SV calculation methods (Arnoldi vs. Lanczos algorithm) or the different forecast



(a) Vertical profiles of the FrDKE at initial time (pink) and for lead times between 24 and 72 h



(b) Vertical profiles of the FrDKE at initial time (pink) and for lead times between 6 and 24 h

Figure 4.6: Vertical profiles of the FrDKE of the leading ten A-SVs at initial time (01 January 2022, 00 UTC) and for different lead times showing the change in FrDKE for A-SVs. Vertical profiles for the forecast combination of Pangu24 and an OTI of 24 h are shown in dash-dotted pink at initial time and in blue for the forecast. For Pangu24 and an OTI of 48 h, vertical profiles are indicated by the continuous pink line at initial time and by red lines for the forecast. For Pangu6 and an OTI of 6 h, the initial time profiles are highlighted in dashed pink and in yellow for the forecast. Pangu6 and an OTI of 24 h vertical profiles are shown in dotted pink at initial time and in green for the forecast. With the first lead time, the vertical profiles are shown by a continuous line, the second are shown with a dashed line, the third with a dash-dotted line, and the fourth with a dotted line (if shown). The vertical black lines indicate the observed fractions of kinetic energy for L-SVs in physics-based NWP models. The left line indicates the initial time fraction of kinetic energy of 0.25 and the right line shows the expected final time fraction of kinetic energy of 0.75.

inference methods (MLWP vs. NWP model), or a combination of both. It is beyond the scope of this work to find a satisfactory answer to this question.

In the forecast, the differences between the various forecast combinations are rather small. Only near the surface are significant differences in the FrDKE apparent in Figure 4.6. The forecast combinations with Pangu6 tend to have smaller FrDKE values near the surface

compared to Pangu24. Like L-SVs in physics-based NWP models, the shift to predominant DKE occurs within the first forecast inference. For subsequent forecast inferences, only minor shifts towards larger FrDKE values are seen in Figure 4.6. In Figure 4.6(b), the behavior of A-SVs on day 1 can be analyzed. These profiles show that the shift to predominantly DKE also occurs within the first forecast inference of Pangu6. For subsequent Pangu6 inferences, the increase in FrDKE is less than the initial increase in FrDKE. However, the relative increase per forecast inference is larger on day 1 than for lead times beyond day 1, where the FrDKE remains nearly constant. This is an indication that A-SV perturbations in Pangu-Weather are unbalanced at initial time. A similar turn of A-SVs towards the attractor of the system may also be present in Pangu-Weather, similar to L-SVs in physics-based NWP models (Kalnay, 2002).

At initial time, there are significant differences between the FrDKE of the two Pangu6 forecast combinations (Figure 4.6(b)). These differences may be attributed to the different perturbation growth mechanisms presented by Baumgart et al. (2019). Up to a lead time of 12 h, perturbation growth is focused on regions with convective instability. These perturbations are then projected onto mid-latitude jets and lead to large-scale perturbation growth for lead times of up to 2 d. For an OTI of 6 h, A-SVs are optimized for an OTI in the first perturbation growth regime. In contrast, all other forecast combinations optimize A-SVs for OTIs within the second perturbation growth regime. Hence, A-SVs of the forecast combination of Pangu6 and an OTI of 6 h target different sensitivities in the initial conditions if perturbation growth mechanisms similar to Baumgart et al. (2019) are active in Pangu-Weather. The Pangu24 forecast combinations also show an OTI-dependent initial time FrDKE. Similarly, the respective A-SVs likely target different sensitivities in the initial conditions, resulting in different initial time FrDKE profiles.

Latitudinal dependence of the FrDKE

The same latitude-specific domains introduced in Section 4.2.1 can also be applied to the FrDKE. This allows analysis of latitude-related dependencies of the FrDKE. Figure 4.7 shows that, in the Northern Hemisphere mid-latitudes, there are larger FrDKE values at initial time than in the global mean. Hence, the initial time DKE contribution is greater compared to the global mean. At final time, the FrDKE in the mid-latitudes is smaller than the global average. In addition, the FrDKE maximum in the upper troposphere is located at a lower pressure level in the mid-latitudes than in the global mean. The lower extent of the troposphere in the mid-latitudes may lead to a compression of the vertical FrDKE profile (Peixoto and Oort, 1992). The lower FrDKE values at final time also indicate that less of the available DPE is converted to DKE in the mid-latitudes.

In contrast, the tropics have a lower FrDKE at initial time, as shown by the dashed pink line in Figure 4.7. The largest deviations from the global mean are again located in the upper troposphere. Unlike in the mid-latitudes, the vertical FrDKE maxima in the tropics are located above the global vertical maxima. At final time, the tropics show opposite deviations from the global mean compared to the mid-latitudes. Especially in the upper troposphere, there is a significant increase in FrDKE. The FrDKE values reach more than 0.9 at 200 hPa

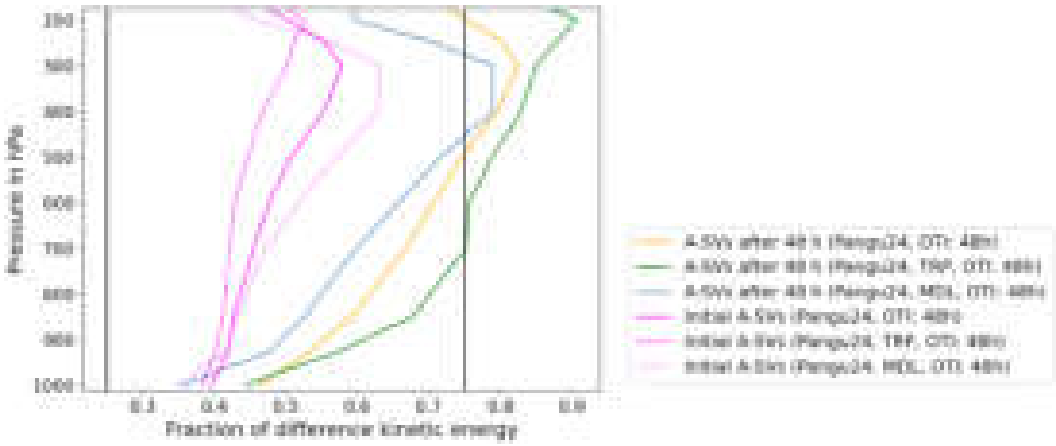


Figure 4.7: Vertical profiles of the FrDKE of the leading ten A-SVs showing the global mean (pink at initial time, yellow at final time), the tropical mean (TRP, dashed pink at initial time, green at final time), and the northern mid-latitudes FrDKE (MDL, dotted pink at initial time, blue at final time). A-SVs are optimized globally for the forecast combination of Pangu24 and an OTI of 48 h with the dry total energy norm. Vertical black lines indicate the observed fractions of kinetic energy for L-SVs in physics-based NWP models. The left line indicates the initial time fraction of kinetic energy of 0.25 and the right line shows the expected final time fraction of kinetic energy of 0.75.

in the tropics. These are significantly higher than both FrDKE values in the mid-latitudes, as well as the observed fraction of kinetic energy of L-SVs in physics-based NWP models. Near the surface, the FrDKE values remain nearly constant, independent of the domain. The larger shift of FrDKE values in the tropics implies that a larger fraction of the available DPE is converted into DKE at final time. The reason, therefore, is the weaker meridional temperature gradients in the tropics. In the mid-latitudes, the meridional temperature gradients are greater (Peixoto and Oort, 1992). Thus, the final time A-SVs have a greater contribution of DPE.

Section 4.2.1 outlines that the subtropics show hemisphere-specific dependencies in the vertical profiles of DTE. For FrDKE, there is no hemispheric dependency of the respective subtropical means (not shown). For both hemispheres, the vertical profiles of FrDKE follow the global-mean vertical profile. Similarly to the tropics, the vertical profiles of the subtropics also show a larger shift of FrDKE than the global mean. Additionally, the vertical maximum of FrDKE is located above the global-mean vertical maximum, as in the tropics. The hemispheric dependence of the subtropical DTE vertical profiles is attributed to the seasonality of the Hadley cell (see Section 4.2.1). For the FrDKE, small meridional temperature gradients throughout the tropical belt (Peixoto and Oort, 1992) lead to large FrDKE values at final time, independent of the hemisphere. Around the poles, the vertical profiles of the FrDKE are the most erratic, exhibiting the largest deviations from the global mean (not shown).

Regardless of the domain, model, and OTI, Figure 4.6 and 4.7 show that the FrDKE values are always smallest near the surface. Complex interactions of the boundary layer with the basic flow provide a possible explanation for the lack of DKE near the surface if similar perturbation growth mechanisms, as in NWP models, are also present in Pangu-Weather

(Montani and Thorpe, 2002). Further research on this question is beyond the scope of this work.

In summary, this section provides an overview of the vertical profiles of the perturbation energy and energy partition of A-SVs in Pangu-Weather. In contrast to L-SVs in physics-based NWP models, A-SVs in Pangu-Weather exhibit an upper-tropospheric DTE maximum already at initial time. At final time, the vertical profiles of the DTE are similar to those of L-SVs in physics-based NWP models, with an upper-tropospheric maximum and a secondary maximum in the lower troposphere. The FrDKE profiles show that A-SVs show a similar transition from predominantly DPE at initial time to predominantly DKE at final time. However, the relative increase of FrDKE from initial to final time is smaller than that of L-SVs in physics-based NWP models. The latitude-dependent vertical profiles of the DTE and FrDKE show that A-SVs are consistent with the general dynamics of the atmosphere.

4.3 Spectral analysis of Arnoldi-SVs in Pangu-Weather

Section 2.3.3 describes the horizontal scales of L-SV perturbations in physics-based NWP models. Similarly, the following section provides a spectral analysis of A-SV perturbations in Pangu-Weather following Selz and Craig (2023). A spectral analysis of the logarithmic densities of DKE, DPE, and DTE is performed (Selz et al., 2022; Augier and Lindborg, 2013). The spectral density is defined as the energy per unit $\log(k)$. The spectral analysis is limited to the mid-latitudes of the Northern Hemisphere. All grid points with a latitude $30\,°\mathrm{N} < \varphi < 60\,°\mathrm{N}$ are included in the spectral analysis. The spectrum is normalized by dividing the grid point value by the sum across all grid points. Spectral analysis is conducted using *Numpy's* Discrete Fourier Transform (Harris et al., 2020) with the efficient Fast Fourier Transform algorithm (Cooley and Tukey, 1965). In contrast to Selz and Craig (2023), the input data is retained on its native 0.25° equiangular grid. Without the truncation, the spectral analysis can be extended to larger wavenumbers (i. e., smaller wavelengths). The overall impact of omitting this truncation is expected to be minor. Negligible reductions of the spectral densities at higher wavenumbers are possible. In addition, this analysis evaluates the 500 hPa DKE spectrum in contrast to the 300 hPa DKE spectral analysis performed by Selz and Craig (2023).

In Figure 4.8, the spectral density recedes at both ends of the spectrum. The general spectral pattern of the DKE is similar to that of the analysis of Selz and Craig (2023). At wavelengths greater than 1,000 km, the spectral density of DKE does not decrease as sharply as it does in Selz and Craig (2023). For A-SVs in Pangu-Weather, the spectral density decreases by less than one order of magnitude at planetary wavelengths. For typical atmospheric background spectra, the spectral density of DKE falls by more than two orders of magnitude (Selz and Craig, 2023). In both spectral analyzes, the peak spectral density is found at wavelengths between 1,000 and 100 km. At wavelengths below 100 km, Figure 4.8 shows that the spectral density of the DKE remains nearly constant. This could be attributed to the construction of SVs using the Arnoldi method. There, SV construction is initialized by a perturbation vector consisting of pure white noise (see Section 2.4.1). The limited number of forecast

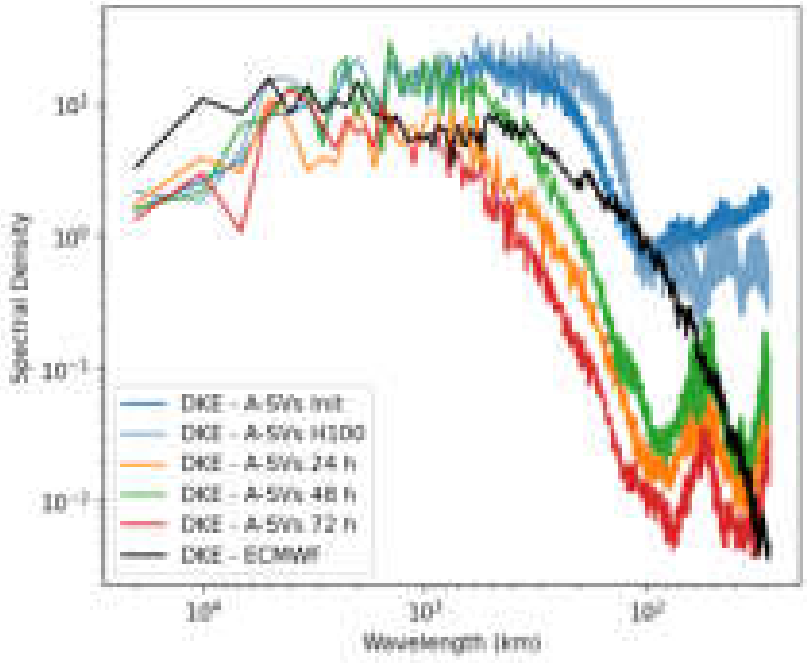


Figure 4.8: Spectral analysis of the 500 hPa mid-latitude DKE (30 °N $< \varphi <$ 60 °N) in Pangu-Weather at initial time (blue), at a lead time of 24 h (orange), at a lead time of 48 h (green), and at a lead time of 72 h (red) for forecasts initialized on 01 January 2022, 00 UTC. A-SVs are optimized globally for the forecast combination of Pangu24 and an OTI of 48 h with the dry total energy norm. The light blue line indicates the initial time DKE spectrum for an $\mathbf{H}$ matrix of size $m = 100$ (H100). The initial time spectral density of the 500 hPa DKE of the ECMWF IFS ensemble forecast, initialized on 01 January 2022, 00 UTC, is shown in black.

inferences with the Block Arnoldi method retains a non-negligible contribution from initial white noise that transcends into the SVs. Section 4.2.1 already highlights the impact of transcending white noise on A-SVs in Pangu-Weather by high initial time DTE values.

With increasing lead time, the spectral density of the DKE in Figure 4.8 decreases across all wavelengths. The only exception is the final time spectral density at a lead time of 48 h, for which the A-SVs are optimized. Here, the spectral density for the 48 h forecast increases compared to the 24 h forecast. Selz and Craig (2023) show that the spectral density increases with lead time for perturbations initialized from typical atmospheric background spectra. The contrasting behavior is likely induced by the differing initial condition perturbations given to Pangu-Weather. At wavelengths below 100 km, the spectral densities of DKE form a characteristic feature of undulating spectral density. The same pattern appears in the spectral densities of DPE and DTE (not shown). In addition, the beginning of the same pattern is visible in Selz and Craig (2023, Fig. 3). However, these plots terminate at horizontal wavelengths of 70 km due to truncation. The same pattern does not appear for physics-based NWP models (Selz and Craig, 2023). Hence, this undulating feature is a characteristic of Pangu-Weather forecasts and develops independently of the imposed initial time perturbations.

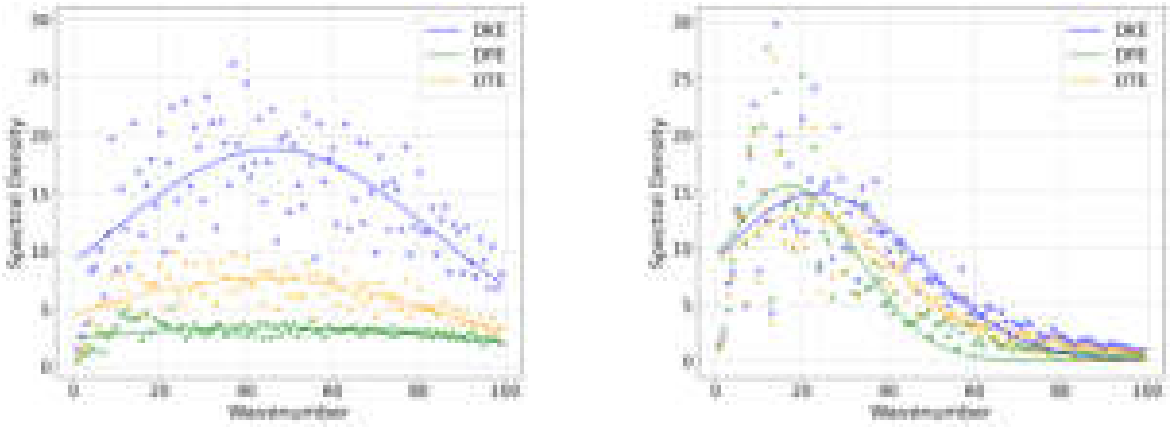
Figure 4.8 includes the initial time DKE spectral analysis for ensemble perturbations computed from A-SVs for an $\mathbf{H}$ matrix of size $m = 100$ ($l = 10$ and $b = 10$). This gives an indication for the potential reduction of transcending white noise in the initial A-SVs due to increased size of the $\mathbf{H}$ matrix. At horizontal wavelengths between 200 and 100 km, the spectral decay starts at smaller horizontal wavelengths for $\mathbf{H}_{100,100}$ than for A-SVs computed from a $\mathbf{H}$ matrix of size $m = 35$. Below horizontal wavelengths of 100 km, the DKE spectrum of $\mathbf{H}_{100,100}$ has lower spectral densities. This implies that the contribution of transcending white noise from the Arnoldi method initialization can be reduced by increasing the size of the $\mathbf{H}$ matrix. In addition, the described undulating feature of Pangu-Weather forecasts seems to also appear in the initial time DKE spectrum. An increased number of forward integrations with the Arnoldi method could imply that the inherent spectrum of Pangu-Weather is also engraved in the initial time A-SVs.

4.3.1 Spectral comparison to ECMWF initial time ensemble perturbations

Due to interest in ensemble generation with A-SVs, Figure 4.8 includes the DKE spectral analysis of the initial time L-SV-based ensemble perturbations of the ECMWF IFS ensemble forecast. The spectral densities of the IFS ensemble perturbations are derived using the methodology described.

The spectral analysis of the ECMWF IFS initial time ensemble perturbations has a shape similar to the A-SV-based ensemble perturbations in Pangu-Weather at larger wavelengths (for details on the construction of ensemble perturbations with Pangu-Weather, see Section 3.2.2). These also show only a weak decay of spectral density at planetary scales, in contrast to Selz and Craig (2023). The contrasting behavior may be attributed to the different NWP models and their ensemble initialization methods used for the spectral analysis (Leutbecher and Palmer, 2008; Zängl et al., 2015). Below horizontal wavelengths of 200 km, significant differences between the initial time ensemble perturbations of the IFS and Pangu-Weather develop. The spectral density of the IFS ensemble decays rapidly at wavelengths below 100 km. The rapid decay of spectral density in this wavelength range is in accordance with the basic energy spectra of the atmosphere (Lorenz, 1969). In contrast, Pangu-Weather initial time ensemble perturbations exhibit a short decline in spectral density between horizontal wavelengths of 200 and 100 km. In this wavelength range, the decay in spectral density in Pangu-Weather is sharper than the respective decay of the IFS in this wavelength bin. Figure 4.8 shows that at wavelengths below 100 km, the spectral densities of the DKE remain constant, in contrast to further decay in the IFS ensemble. The comparison to the initial time DKE spectral analysis for $\mathbf{H}_{100,100}$ at wavelengths below 100 km implies that A-SVs lead to a fundamentally different spectrum of the initial time ensemble perturbations than the L-SV-based IFS ensemble perturbations.

The spectral densities of both DKE and DPE shift towards the spectrum of the ECMWF IFS initial time ensemble perturbations during the forecast. Upon forecast inference, the spectral density of Pangu-Weather shows a decay close to the basic energy spectrum of Lorenz (1969). With increasing lead time, the decay of spectral density occurs at larger wavelengths, similar to the analysis of Selz and Craig (2023). At wavelengths below 100 km,



(a) Spectral density analysis at initial time

(b) Spectral density analysis at a lead time of 48 h

Figure 4.9: Spectral density analysis of the DKE (blue), DPE (green), and DTE (orange) showing a strong upscale growth of the perturbation energy from initial to final time. Forecasts are initialized on 01 January 2022, 00 UTC. An exponential quadratic fit (continuous lines) is applied to the first 100 wavenumbers (points).

the same undulating feature of Pangu-Weather forecasts is preserved without any change to the wavelengths at which it occurs. Similar curious patterns in the small-scale spectral density also appear for other MLWP models (Rodwell et al., 2025).

4.3.2 Peak spectral density

Section 2.3.3 shows that L-SVs in physics-based NWP models have characteristic spectral maxima. The comparison between the initial and final time spectra shows that the total perturbation energy undergoes upscale growth. At initial time, the maximum potential perturbation energy is located at higher wavenumbers than the peak of the total perturbation energy. The final time potential perturbation energy maximum shifts to smaller wavenumbers, as does the peak of the final time total perturbation energy.

An exponential quadratic fit is applied to the first 100 wavenumbers of the spectral analysis to assess whether similar shifts in the peak spectral density are present for A-SVs in Pangu-Weather. For raw A-SV perturbations, the spectral density spectra are much broader. Therefore, these are not suitable for the proposed exponential quadratic fit. This is again caused by the contamination of transcending white noise through the Arnoldi algorithm. A linear combination of A-SVs filters part of the contribution of transcending white noise to the spectral analysis. Hence, we analyze the A-SV-based ensemble perturbations (see Section 3.2.2).

Figure 4.9 shows the spectral analysis of DKE, DPE, and DTE at initial and final time (OTI of 48 h). In contrast to Figure 2.4, both spectra are normalized with the respective sum over all grid points. As a result, the spectral density remains constant and does not amplify, as in Figure 2.4. At initial time, the peak DTE spectral density is located at wavenumber 42. The maximum of the DTE spectrum shifts to wavenumber 20 at final time. This shows a strong

upscale growth of the DTE. For L-SVs in physics-based NWP models, the initial time total perturbation energy spectral maximum is located between wavenumbers 16 and 18. This shows that the A-SV perturbations have a much broader spectrum at initial time. At final time, the total perturbation energy spectral density of L-SVs peaks at wavenumbers 10 to 13. The larger shift in the final time DTE maximum of A-SVs shows that the upscale growth of SV perturbations is more pronounced for A-SVs.

Section 2.3.3 also highlights the different shifts in kinetic and potential perturbation energy. Figure 4.9 shows that for the DKE and DPE in Pangu-Weather, the shift between the two has similar characteristics. The spectral density of the DPE and DKE has an initial time maximum near wavenumber 45. At final time, significant differences between the DPE and DKE spectral maxima develop. Similar to L-SVs in NWP models, the DPE spectral maximum (wavenumber 16) of the A-SVs is located at a lower wavenumber than the final time DKE spectral maximum (wavenumber 22). As mentioned, all spectral maxima of A-SV perturbations in Pangu-Weather are located at higher wavenumbers than the spectral maxima of L-SV perturbations in physics-based NWP models. The general patterns in the upscale growth of the DPE and DKE are similar for both the L-SV perturbations in NWP models and the A-SV perturbations in Pangu-Weather. This suggests that similar adjustment processes to SV perturbations seem to be present in both, independent of the forecast inference method (NWP vs. MLWP) and SV computation method (L-SVs vs. A-SVs). At lead times beyond 48 h, the peak spectral densities of A-SV perturbations shift to the wavenumbers similar to L-SV perturbations in NWP models.

In summary, the A-SV-based initial time ensemble perturbations in Pangu-Weather show spectral densities similar to those of the ECMWF IFS initial time ensemble perturbations. Differences in spectral density decay develop at horizontal wavelengths below 100 km. One origin of this is likely the transcending white noise of the Arnoldi algorithm initialization. However, intrinsic properties of MLWP models might also cause such a property to some extent. An increase in the size of the $\mathbf{H}$ matrix indicates a reduction in the impact of this transcending white noise on A-SVs. In addition, the spectral analysis develops a characteristic spectral feature of Pangu-Weather forecasts known from previous analyzes. A-SVs in Pangu-Weather exhibit a larger upscale growth than L-SVs in physics-based NWP models. A detailed analysis reveals that, like for L-SVs in physics-based NWP models, the DPE performs a larger upscale shift of the spectral maximum than the DKE.

5 Sensitivity analysis

In the following chapter, A-SVs are used for a sensitivity analysis of Cyclone Xynthia. The results of the sensitivity analysis are compared with the study of Baño-Medina et al. (2025), who analyze the sensitivity pattern of Cyclone Xynthia in both NWP and MLWP models. The following chapter first discusses the synoptic development of Cyclone Xynthia. Then, a sensitivity analysis is conducted with A-SVs in Pangu-Weather.

5.1 Synopsis of Cyclone Xynthia

Cyclone Xynthia was a high-impact extratropical cyclone in late February 2010, causing more than 1 billion € in damage in Western and Central Europe (Bissolli, 2010). Cyclone Xynthia formed near the Azores. At cyclogenesis time, a large trough was located over the Central North Atlantic. The surface low that would mature into Cyclone Xynthia quickly moved along the southern flank of the main upper-level trough. Located ahead of the developing secondary trough, Xynthia diverted onto a northeasterly trajectory. The cyclone moved rapidly northeast towards the Iberian Peninsula due to strong upper-level winds. The upper-level wind and divergence pattern on 27 February 2010 indicates that Cyclone Xynthia propagated within a complex structure of two minor jet streaks (not shown).

Cyclone Xynthia quickly intensified with an intensification rate of 20 hPa/24 h on 27 February 2010. Thus, the intensification of Cyclone Xynthia exceeds the latitude-dependent threshold for explosive cyclogenesis (Sanders and Gyakum, 1980). During intensification, diabatic processes made an unusually large contribution to the explosive cyclogenesis of Cyclone Xynthia (Fink et al., 2012). The cyclone crossed the northwestern part of the Iberian Peninsula, bringing high winds and heavy precipitation on 27 February 2010. In the second half of the day, Xynthia traversed the Bay of Biscay and moved on-shore near the mouth of the Loire River in Western France early on 28 February 2010. Onshore, the cyclone brought hurricane force winds with wind gusts of up to 160 $\mathrm{km\,h^{-1}}$. High wind speeds and a severe storm surge caused significant damage along the French Atlantic coast. The cyclone passed through Northern France and tracked along the coast of the North Sea. On 01 March 2010, Cyclone Xynthia reached the Baltic Sea and ultimately dissipated over Southern Finland.

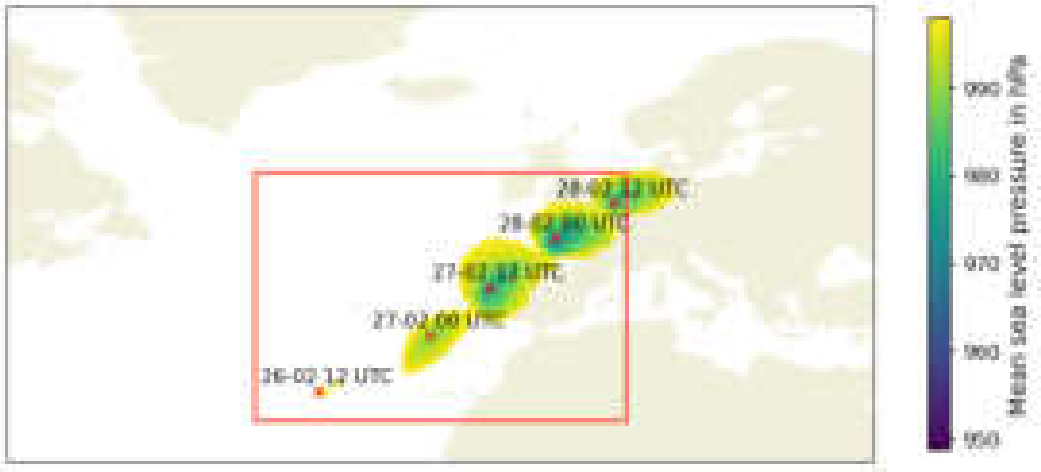


Figure 5.1: Trajectory of Cyclone Xynthia over Western Europe from 26 February 2010, 12 UTC to 28 February 2010, 12 UTC is indicated by the MSLP of the cyclone in ERA5. The red crosses show the central positions, determined by the lowest pressure, of Cyclone Xynthia at the corresponding dates. The maroon box indicates the extend of local projection operator **P** during the A-SV calculation.

5.2 Sensitivity analysis of Cyclone Xynthia

Baño-Medina et al. (2025) study the sensitivity fields of Cyclone Xynthia with a NWP and a MLWP model. In their study, Baño-Medina et al. (2025) use a sensitivity algorithm to derive the initial time atmospheric sensitivities of the kinetic energy for Cyclone Xynthia. The kinetic energy is chosen because it serves as a proxy for the strength of the cyclone. For both models, the forecasts initialized on 26 February 2010, 12 UTC are evaluated on 28 February 2010, 00 UTC. The sensitivities of the kinetic energy are calculated for a limited domain encompassing the Bay of Biscay (Baño-Medina et al., 2025, green boxes in Fig. 2) where Cyclone Xynthia is located on 28 February 2010, 00 UTC. Baño-Medina et al. (2025) show that the derived kinetic energy sensitivities for both NWP and MLWP models are similar.

Chapter 4 outlines that A-SV perturbations in Pangu-Weather are consistent with L-SV perturbations in physics-based NWP models and exhibit promising cross-variable consistency. Hence, A-SVs can also be used to evaluate the sensitivities of Cyclone Xynthia in Pangu-Weather. Comparison with the sensitivity analysis of Baño-Medina et al. (2025) highlights the consistency of the A-SVs in Pangu-Weather. Chapter 3 denotes that Pangu-Weather is trained on the ERA5 data set, with the training period spanning from 1979 to 2017. Thus, Cyclone Xynthia is part of the training data. Therefore, Pangu-Weather has already seen the data during the training period. The Spherical Fourier Neural Operator (SFNO) model (Pathak et al., 2022), which Baño-Medina et al. (2025) use, also includes Cyclone Xynthia in their training data set. Baño-Medina et al. (2025) justify the use of the MLWP model on data that the SFNO model has already seen during training due to the limited contribution of Cyclone Xynthia to the overall model loss function. Analogous arguments are applicable to Pangu-Weather. First, Pangu-Weather is not specifically optimized to generate forecasts

with a lead time of 36 h (Bi et al., 2023). In addition, Cyclone Xynthia was short-lived and affected only a minor part of the global domain (Baño-Medina et al., 2025). Further, Cyclone Xynthia had an unusual track for intense cyclones over the North Atlantic (Fink et al., 2012). Hence, the overall contribution of Cyclone Xynthia to the model loss function is likely negligible (Baño-Medina et al., 2025). This justifies the use of Pangu-Weather on data that has already been presented to the model during training.

Changes in the parameter choices of the Arnoldi method for SV calculation are necessary to compute A-SVs for Cyclone Xynthia. First, the forecasts are initialized from ERA5 for a valid time on 26 February 2010, 12 UTC. The OTI is set to 36 h. For this, Pangu6 is used as the forward propagating model. The Arnoldi method, in the used implementation, does not incorporate the Pangu-Weather temporal aggregation algorithm for the forward propagation of the initial time perturbations. Hence, the forward propagation of initial time perturbations is achieved by six consecutive Pangu6 inferences. The Arnoldi method can be easily combined with a local projection operator $\mathbf{P}$ that limits perturbation growth to a certain domain. However, such a local projection operator $\mathbf{P}$ is the same at initial and final time. Thus, limiting $\mathbf{P}$ to the Bay of Biscay would certainly not identify any perturbation growth related to Cyclone Xynthia in the initial conditions on 26 February 2010, 12 UTC. Hence, $\mathbf{P}$ must include a larger domain than just the Bay of Biscay. Figure 5.1 shows the path of Cyclone Xynthia from 26 February 2010, 12 UTC to 28 February 2010, 12 UTC, and the chosen extent of $\mathbf{P}$ is indicated by the maroon box. The respective extent of $\mathbf{P}$ is much larger than the extent of the domain for which the initial time sensitivities of the kinetic energy of Cyclone Xynthia are calculated (Baño-Medina et al., 2025, green boxes in Fig. 2). Thus, A-SVs in Pangu-Weather identify atmospheric sensitivities that are not exclusively attributed to Cyclone Xynthia.

Winkler and Denhard (2025) outline that the perturbation amplitude h depends on the perturbed state vector $\mathbf{x}_0'$ itself. The limitation of perturbation growth by local projection operator $\mathbf{P}$ therefore influences the choice of h, as $\mathbf{P}$ reduces the norm of $\mathbf{x}_0'$. Hence, a smaller amplitude h must be chosen to ensure the consistency of the Arnoldi method. For the case study of Cyclone Xynthia, $h = 500$ is chosen to account for the local projection operator $\mathbf{P}$. For this case study, the leading ten A-SVs are calculated. However, the following analysis focuses on the leading A-SV. Forecasts are computed for a lead time of up to 48 h to assess the impact of A-SVs perturbations on the development of Cyclone Xynthia.

The following section first presents the initial time sensitivities of Cyclone Xynthia on 26 February 2010, 12 UTC, and compares them with the initial time sensitivities presented by Baño-Medina et al. (2025). In a second step, the impact of A-SV perturbations on the development of Cyclone Xynthia is assessed. Similarities and differences to Baño-Medina et al. (2025) are described.

5.2.1 Initial time sensitivities

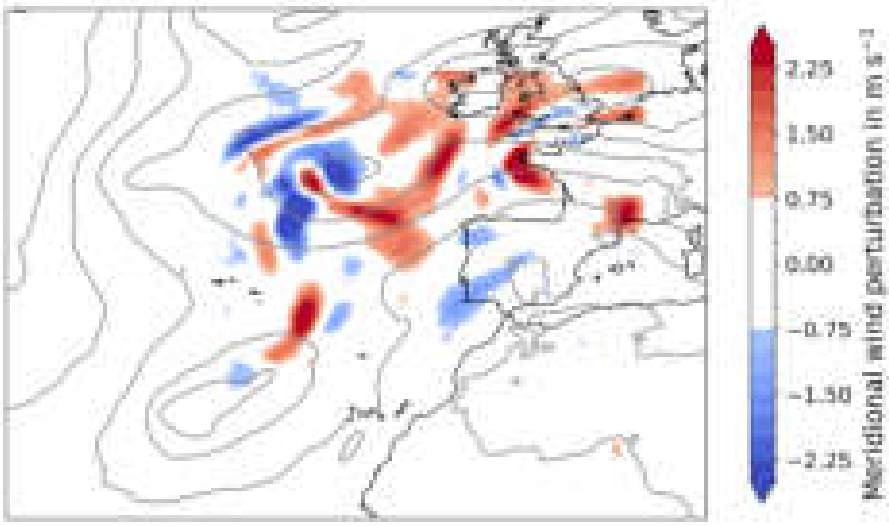
Baño-Medina et al. (2025) show that both the NWP and MLWP models identify similar initial time sensitivities of the kinetic energy for Cyclone Xynthia. Their analysis shows

high sensitivities along a mesoscale mid-tropospheric moisture filament near the Azores. The moisture filament is located along an atmospheric river that extends from the subtropics towards the Iberian Peninsula. The high sensitivities of the moisture field around Cyclone Xynthia are consistent with previous studies highlighting an anomalously large contribution of diabatic processes during the intensification of Cyclone Xynthia (Fink et al., 2012). For A-SVs, the moisture variable of Pangu-Weather is not perturbed (see Equation 2.7). Hence, initial time moisture sensitivities cannot be studied with A-SVs. However, Baño-Medina et al. (2025) also include atmospheric sensitivities of temperature (potential temperature for the NWP model) and meridional wind. For these forecast variables, the sensitivity patterns can be compared.

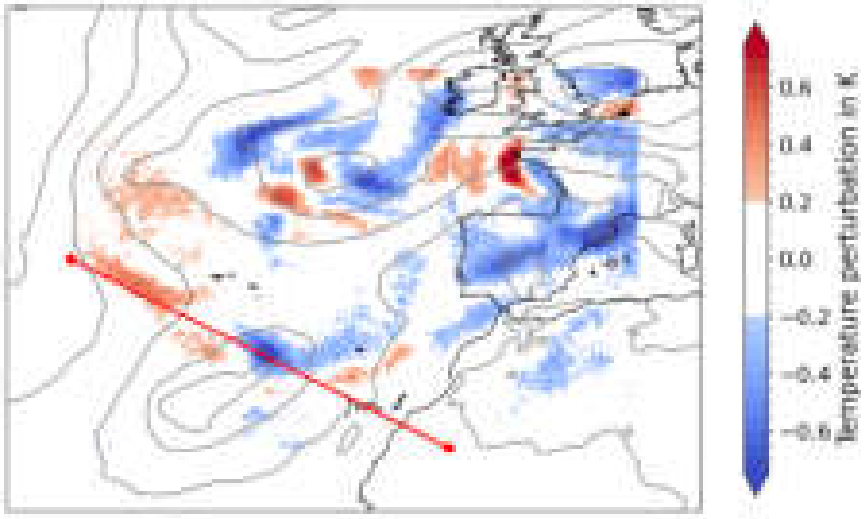
Figure 5.2 shows the 700 hPa sensitivity fields of temperature and meridional wind for the leading A-SV in Pangu-Weather. As mentioned, the larger local projection operator **P** causes A-SVs to identify atmospheric sensitivities that are not directly attributed to Cyclone Xynthia. At initial time, Cyclone Xynthia is located south of the Azores. The sensitivity patterns within the domain of **P** are dominated by perturbations attributed to the central low east of England and a secondary low southwest of the British Isles. Nevertheless, significant sensitivities are also identified to the northeast of Cyclone Xynthia. Both the 700 hPa meridional wind and the 700 hPa temperature show a dipole sensitivity structure along the track of Cyclone Xynthia. These dipole-structured sensitivities for temperature and meridional wind are also found independently of the forecast models (NWP vs. MLWP model) by Baño-Medina et al. (2025). The agreement of our approach with Baño-Medina et al. (2025) highlights that, despite the much larger optimization domain, A-SVs trigger similar sensitivities attributed to Cyclone Xynthia. In addition, this also shows that NWP and MLWP models exhibit similar sensitivities. Furthermore, the atmospheric sensitivities of Cyclone Xynthia appear to be consistent between different MLWP models, despite the fundamentally different model architectures of the SFNO model and Pangu-Weather.

Figure 5.3 shows the vertical cross-section taken along the red line in Figure 5.2(b), which follows a trajectory similar to the cross-sections shown by Baño-Medina et al. (2025). The cross-section extends from the North Atlantic, ahead of Cyclone Xynthia, into Northern Africa. Figure 5.3 shows a dipole structure for the temperature perturbations. Sensitivities are predominantly located below 400 hPa, similar to Baño-Medina et al. (2025). However, it must be noted that Baño-Medina et al. (2025) show the sensitivity patterns of the moisture variable (specific humidity for the NWP model and relative humidity for the SFNO model). In contrast, Figure 5.3 shows the temperature perturbation of the leading A-SV in Pangu-Weather, because A-SVs are optimized for the dry total energy norm (see Equation 2.7). Thus, the initial time A-SVs do not contain any moisture perturbations.

The dipole structure indicates that the atmospheric state ahead of Cyclone Xynthia is sensitive to the vertical temperature gradient along the atmospheric river stretching towards the Iberian Peninsula. The sensitivity structure of the leading A-SV in Pangu-Weather is very similar to the kinetic energy sensitivities of the SFNO model. Both MLWP models show atmospheric sensitivities attributed to Cyclone Xynthia between 25 and 15 °W. In Pangu-Weather, the sensitivity patterns are slightly shifted to the west compared to the sensitivities in the SFNO model. In addition, the leading A-SV also identifies spurious stratospheric



(a) 700 hPa meridional wind



(b) 700 hPa temperature

Figure 5.2: Sensitivity patterns of the 700 hPa meridional wind and temperature for the leading A-SV in Pangu-Weather. A-SVs are optimized for an OTI of 36 h. Forecasts are initialized on 26 February 2010, 12 UTC. The MSLP field is indicated by the gray contour lines at every 5 hPa between 980 and 1,020 hPa, showing the general synoptic set-up over the Northern Atlantic and Western Europe during forecast initialization. Cyclone Xynthia is the low-pressure system located south of the Azores. The red line indicates the trajectory of the vertical cross-section.

sensitivities similar to the SFNO model. These are already indicated by Baño-Medina et al. (2025). It is intriguing that two different MLWP models exhibit similar sensitivities in the stratosphere, which gives rise to a potential atmospheric connection learned by both MLWP models. In addition, the A-SVs in Pangu-Weather show sensitive areas above Cyclone Xynthia near the tropopause. This indicates a potential interaction of the cyclone with the upper troposphere. In total, the sensitivity patterns in Pangu-Weather have a larger vertical extent, with significant signals also at the tropopause. The cross-section of the meridional

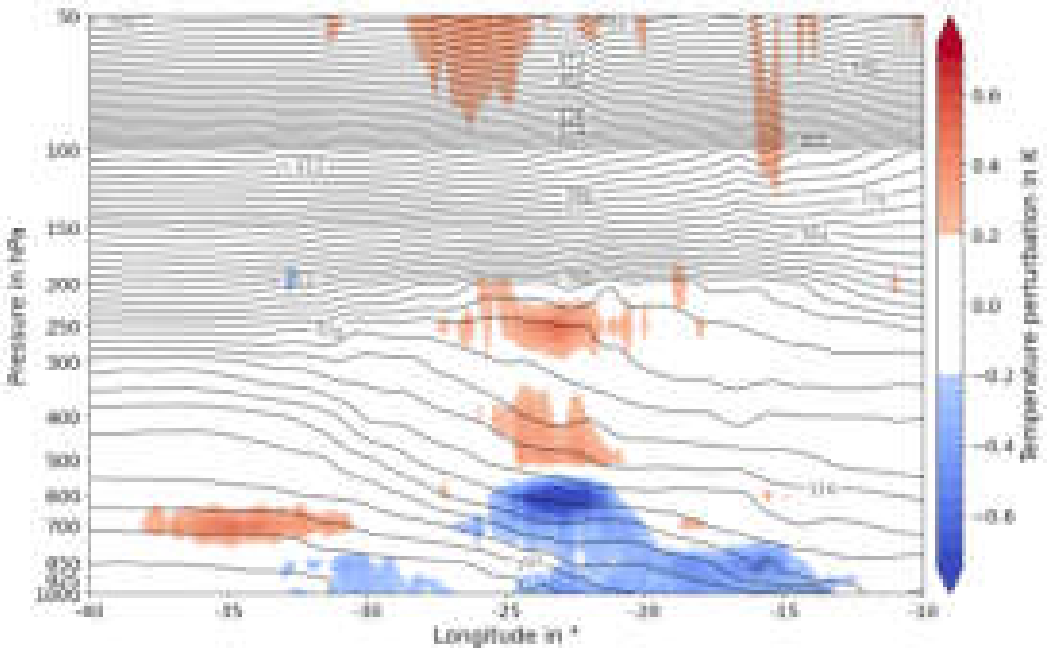


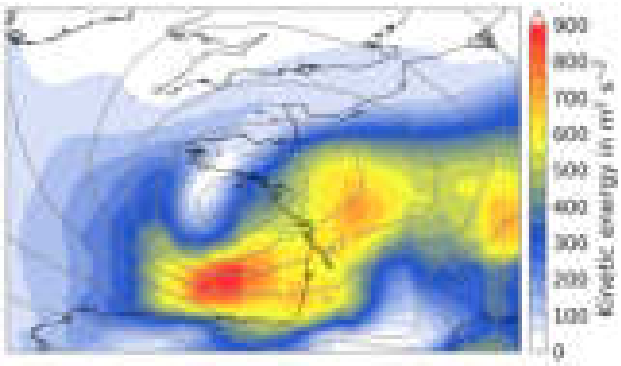
Figure 5.3: Vertical cross-section of the initial time temperature perturbation of the leading A-SV on 26 February 2010, 12 UTC. The trajectory of the cross-section is indicated by the red line in Figure 5.2(b). The contour lines show the potential temperature of ERA5 every 4 K between 280 and 516 K.

wind also exhibits a dipole structure between 25 and 15 °W ahead of Cyclone Xynthia (not shown). The dipole structure is restricted to vertical levels below 400 hPa. These signals are related to a short-wave trough located north of the surface low. Similarly to Baño-Medina et al. (2025), the sensitivity patterns suggest that mid-level southerly winds impact the development of Cyclone Xynthia.

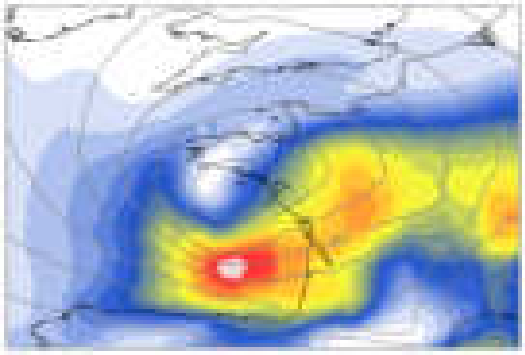
In summary, the initial time sensitivity patterns of the leading A-SV in Pangu-Weather are consistent with the kinetic energy sensitivity patterns for Cyclone Xynthia. This indicates that the sensitivity patterns are independent of the forecasting model class (NWP vs. MLWP model). The fact that both MLWP models identify similar spurious stratospheric sensitivities suggests that both MLWP models have learned similar atmospheric physics. A-SV-based sensitivity patterns show that Cyclone Xynthia is sensitive to the horizontal and vertical temperature gradients along the atmospheric river that extends towards the Iberian Peninsula. The magnitude of the mid-level southerly winds of the short-wave trough north of Cyclone Xynthia exhibits similar sensitivity patterns to both NWP and MLWP models.

5.2.2 Evolution of the Arnoldi-SV-based sensitivities for Cyclone Xynthia

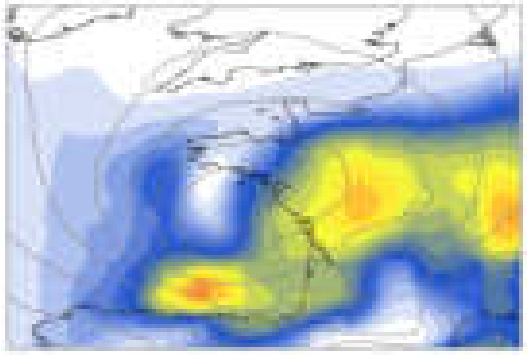
The following section evaluates the impact of A-SV-based sensitivities on the forecast of Cyclone Xynthia. Therefore, forecasts with a lead time of 36 h are evaluated. At this valid time, Cyclone Xynthia is located over the Bay of Biscay. The 850 hPa kinetic energy is



(a) Unperturbed forecast



(b) Positively perturbed forecast



(c) Negatively perturbed forecast

Figure 5.4: 850 hPa kinetic energy of the unperturbed, positively perturbed, and negatively perturbed forecast with a lead time of 36 h over the Bay of Biscay. The gray contour lines represent the MSLP field every 4 hPa between 966 and 998 hPa. The forecasts are perturbed by the leading A-SV.

used as a metric to measure the impact of the A-SV-based sensitivities on the intensity of Cyclone Xynthia. The 850 hPa pressure level is chosen because it captures the development and intensity of low-level jets around Cyclone Xynthia. For this analysis, multiple forecast inferences are used. First, an unperturbed model forecast is computed. In a second step, the initial conditions are positively perturbed (i. e., the A-SV perturbation is added to the initial conditions). Finally, the same A-SV perturbation is subtracted from the initial conditions (i. e., the initial conditions are negatively perturbed) to analyze whether negatively perturbing the initial conditions generates antisymmetric perturbations in the forecast. Positive perturbations in the context of this analysis refer to A-SV-perturbed forecasts that increase the 850 hPa kinetic energy of Cyclone Xynthia. Negative perturbations are defined as leading to a decrease in the intensity of Cyclone Xynthia. Baño-Medina et al. (2025) show that positively perturbing the initial conditions around Cyclone Xynthia generates an even stronger cyclone, while negatively perturbing the initial conditions leads to a weaker cyclone.

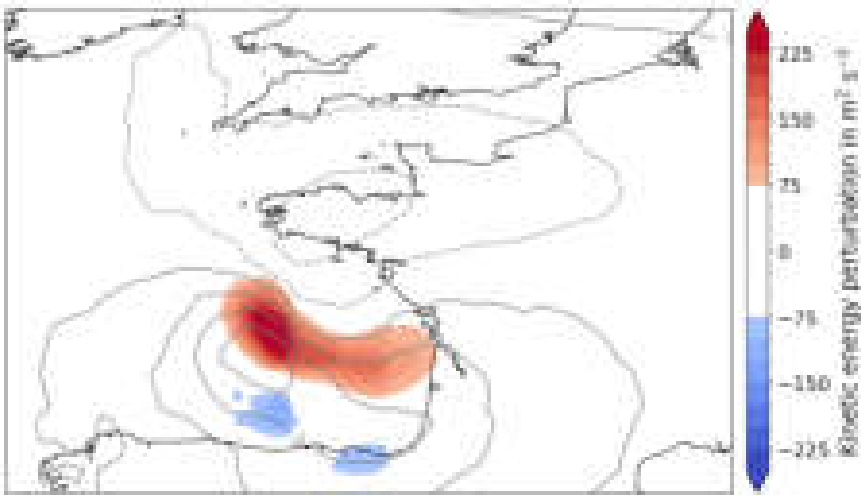
Figure 5.4 shows that the 850 hPa kinetic energy is focused along the southern flank of Cyclone Xynthia, where the low-level jet wraps around the cyclone center. Here, the horizontal pressure gradients are the largest as well. Upstream of Cyclone Xynthia, high kinetic energy values are present over Central France. However, the impact of A-SV-based

sensitivities is focused on the southeastern flank of Cyclone Xynthia, where the highest kinetic energy values are located. For the positively perturbed forecast, the extent and magnitude of the kinetic energy maximum north of the Spanish Coast increase. The increase coincides with stronger horizontal pressure gradients over the Bay of Biscay and a deeper cyclone. Both indicate, from theory, that the wind speeds of Cyclone Xynthia should increase in the positively perturbed forecast. Indeed, the increase in 850 hPa kinetic energy shows that A-SVs lead to a strengthening of Cyclone Xynthia and hence impose physically coherent perturbations in Pangu-Weather.

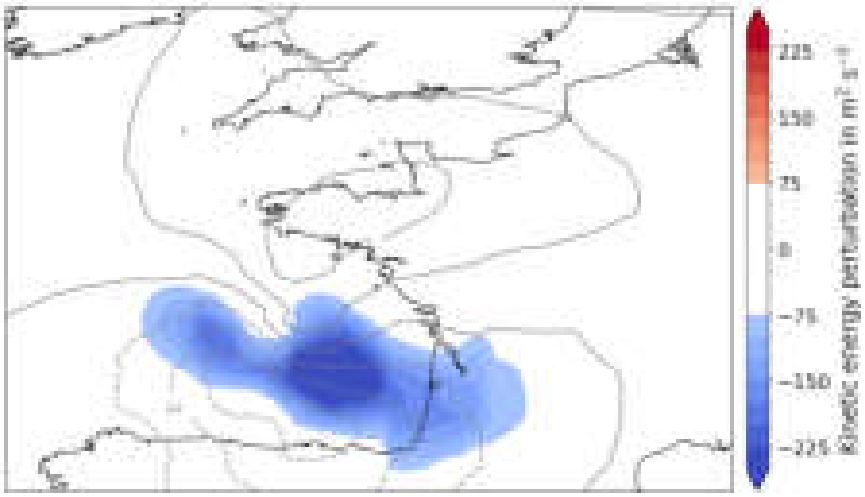
For the negatively perturbed forecast, the 850 hPa kinetic energy and thus the intensity of Cyclone Xynthia decrease. This coincides with a significantly lower horizontal pressure gradient and a weaker cyclone. For both unperturbed and positively perturbed forecasts, the central pressure of Cyclone Xynthia is below 970 hPa. In the case of the negatively perturbed forecast, the central pressure remains above 970 hPa at a lead time of 36 h. In addition to the smaller and weaker 850 hPa kinetic energy maximum, the maximum is also shifted to the southeast and is located closer to the Iberian Peninsula for the negatively perturbed forecast. Baño-Medina et al. (2025) show very similar results for the positively and negatively perturbed forecasts of the sensitivities of Cyclone Xynthia in the SFNO model. The positively perturbed forecast increases the kinetic energy over the Southern Bay of Biscay, while the negatively perturbed forecast decreases the kinetic energy in this part of the domain. This shows that despite the much larger optimization domain of A-SVs, A-SVs indeed trigger similar atmospheric sensitivities attributed to Cyclone Xynthia.

Figure 5.5 shows the difference in kinetic energy between the perturbed and unperturbed forecasts. Note that the MSLP field is unperturbed at initial time (see Section 3.2.1) and thus the MSLP field adjusts freely to the A-SV perturbations of the perturbed forecast variables. For the positively perturbed forecast, the kinetic energy strongly increases southeast of the cyclone center, with a minor decrease along the Spanish Coast. The extent of the increase in kinetic energy is more confined for the positively perturbed forecast compared to the decrease in kinetic energy of the negatively perturbed forecast. The magnitude of the kinetic energy perturbation is fairly similar, with peak values of more than 250 $m^2\ s^{-2}$. Baño-Medina et al. (2025) also show that perturbing atmospheric sensitivities antisymmetrically leads to similar patterns of the kinetic energy perturbation. For the positively perturbed forecast, A-SVs lead to a larger kinetic energy perturbation over the central part of the Bay of Biscay. The positive perturbation of the atmospheric sensitivities of Cyclone Xynthia in the SFNO model leads to a maximum increase in kinetic energy closer to the French Coast near Bordeaux. Both A-SVs in Pangu-Weather and the kinetic energy sensitivity patterns in the SFNO model lead to an increase/decrease in the maximum kinetic energy values over the Bay of Biscay of around one third of the maximum kinetic energy values of the unperturbed forecast (Baño-Medina et al., 2025).

The MSLP perturbations of the positively and negatively perturbed forecasts are almost completely anticorrelated, with enhanced/decreased horizontal pressure gradients over the Bay of Biscay for the positively/negatively perturbed forecasts. Baño-Medina et al. (2025) show very similar results for the patterns of the MSLP perturbations. In the SFNO model, the magnitude of the MSLP perturbations is slightly larger compared to the A-SV



(a) Positively perturbed forecast



(b) Negatively perturbed forecast

Figure 5.5: Difference in kinetic energy between the positively/negatively perturbed forecast and the unperturbed forecast over the Bay of Biscay. The initial conditions are perturbed by the leading A-SV. Gray contour lines represent the difference in MSLP between the positively/negatively perturbed forecast and the unperturbed forecast with contour lines every 1 hPa.

perturbations in Pangu-Weather. Otherwise, the similarities between both are striking. This further solidifies the assumption that A-SVs in Pangu-Weather trigger similar atmospheric sensitivities compared to the kinetic energy sensitivities in the SFNO model. This is despite the fact that the A-SVs are not optimized for the kinetic energy (see Equation 2.7) and the much larger optimization domain. Nevertheless, the evolved A-SVs lead to kinetic energy perturbations with magnitudes and patterns similar to those in the study of Baño-Medina et al. (2025).

Positively perturbing the initial conditions leads to a much higher intensity of Cyclone Xynthia. Whereas in the negatively perturbed forecast, the intensity of Cyclone Xynthia decreases. The general similarity of the results here with the study of Baño-Medina et al.

(2025) is striking. This confirms that similar atmospheric sensitivities are present in both MLWP models. The assumption is valid despite the fact that A-SVs do not perturb the specific humidity forecast variable of Pangu-Weather at initial time. The similar patterns of the evolved A-SV perturbations present for Cyclone Xynthia also indicate that Pangu-Weather adjusts to perturbations of different forecast variables. Hence, the final time A-SVs exhibit strong cross-variable consistency. This also demonstrates that, despite the important contribution of diabatic processes during the intensification of Cyclone Xynthia (Fink et al., 2012; Baño-Medina et al., 2025), A-SVs accurately capture the atmospheric sensitivities of Cyclone Xynthia. In addition, this case study validates the results of Baño-Medina et al. (2025). Moreover, these results solidify the assumption that similar atmospheric sensitivities attributed to Cyclone Xynthia are present in both NWP and MLWP models. Finally, this shows that Cyclone Xynthia could have been more intense under more favorable atmospheric conditions.

6 Ensemble evaluation

Multiple parameters must be chosen a priori to generate an ensemble forecast with A-SVs in Pangu-Weather. Section 3.2.1 discusses the parameter choices necessary for the Block Arnoldi method, such as the perturbation amplitude h, which is set to 4,000. The size of $\mathbf{H} \in \mathbb{R}^{m \times m}$ is determined by the choice of block size $b = 5$ and the number of loops $l = 7$ of the Block Arnoldi method, explored in Section 3.2.2, with $m = b \cdot l = 35$. The different trained forecast models of Pangu-Weather provide the opportunity to choose various combinations of OTIs and forecast models $\mathbf{M}$ for the forward propagation of initial time perturbations in the Arnoldi method (see Equation 2.11). The analysis in Chapter 4 details that the forecast combination of Pangu24 and an OTI of 48 h gives the largest perturbation growth in terms of DTE. The ECMWF IFS ensemble forecast uses the same OTI during the L-SV computation (Leutbecher and Palmer, 2008).

Without discussion, Section 3.2.2 states that the a posteriori amplitude scaling factor γ of the A-SV perturbations is set to 0.2. In this chapter, first, an analysis of the ensemble forecasts for three different γ for the months of January and July 2022 is discussed to justify the choice of $\gamma = 0.2$. Similarly to Bülte et al. (2025), the lead time step is 6 h, and the lead times are grouped into four bins to facilitate the analysis of forecast skill for the different scaling factors. The early lead time bin contains lead times between 6 and 48 h. In the intermediate lead time bin, forecasts with lead times between 48 and 120 h are included. Lead times of 120 to 186 h are grouped in the long lead time bin. In addition, the analysis introduces the medium-range lead time bin with lead times between 186 and 240 h.

Several a posteriori amplitude scaling factors γ for an ensemble setup are tested (see Section 6.1). For the specific scaling factor $\gamma = 0.2$, daily ensemble forecasts for the year 2022 are evaluated in the following chapter. A-SVs are computed from the ERA5 00 UTC analysis field, serving also as the initial conditions for the ensemble. The forecast skill of the ensemble is evaluated using the ensemble evaluation metrics presented in Section 3.3. In addition, a simple post-processing method is proposed and analyzed. Post-processing methods are known to improve the forecast skill of an ensemble (Gneiting et al., 2005). Subsequently, the forecast skill of the A-SV-based Pangu-Weather ensemble is evaluated against the ECMWF IFS ensemble forecast. Finally, the analysis concludes with an inter-comparison of the Pangu-Weather-based ensemble forecasts presented by Bülte et al. (2025). For these comparisons, only lead times up to 186 h are included to match the analysis of Bülte et al. (2025). For the entire ensemble evaluation, the analysis is restricted to the European domain, which extends from 35 to 75 °N and 12.5 °W to 42.5 °E. Like Bülte et al. (2025), only a subset of the Pangu-Weather output variables is evaluated. These include the U10M, V10M, T2M, 850 hPa Temperature (T850), and Z500.

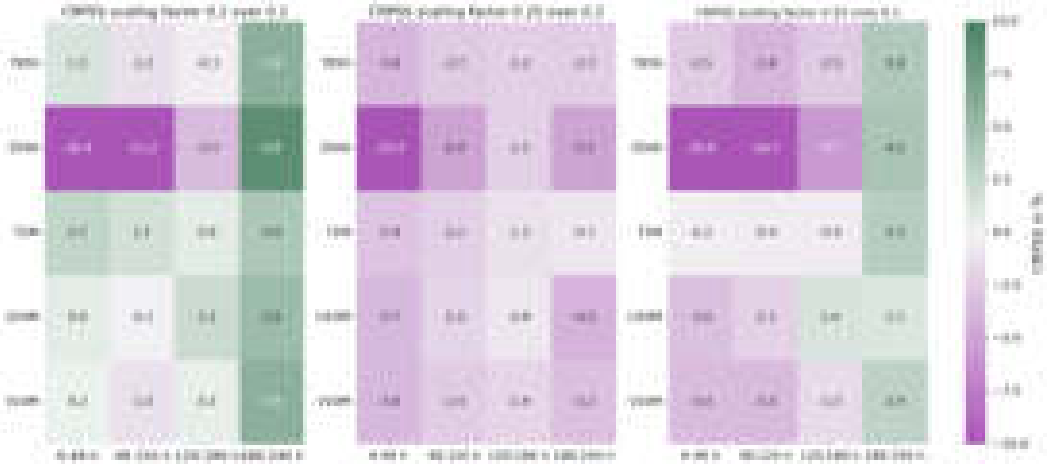
6.1 A posteriori amplitude scaling factor comparison

The a posteriori amplitude scaling factor γ is the last parameter choice that is yet to be discussed in this analysis. For this choice, multiple scaling factors are tested. As mentioned, $\gamma = 0.2$ provides the best ensemble forecast, as determined by the lowest average CRPS for the test months of January and July 2022. For both months, daily ensemble forecasts are computed using three different $\gamma \in \{0.1, 0.2, 0.25\}$. The forecast skill evaluation focuses on a subset of forecast variables. Figure 6.1 shows the respective CRPSS values for the different γ. The CRPSS in the respective heat maps is derived from the CRPS of the ensemble of the first scaling factor, $CRPS_{\mathrm{f}}$, and the CRPS of the ensemble of the second scaling factor, $CRPS_{\mathrm{ref}}$ (see Equation 3.5).

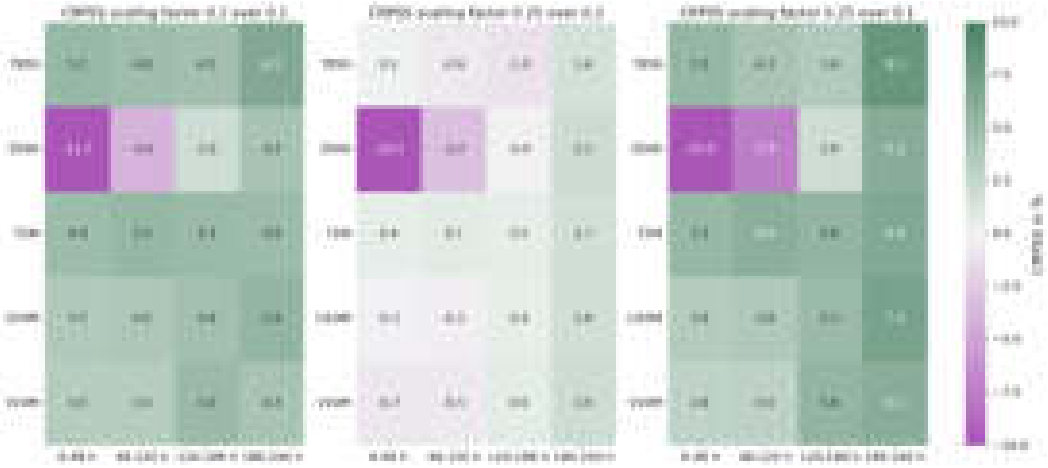
In most cases, the mean CRPSS between the different scaling factors is driven by the CRPSS of the Z500 ensemble. At early lead times, Z500 is strongly over-dispersive, with an increasing spread in winter and for larger γ (not shown). The lower CRPSS values for the Z500 ensemble are mainly driven by a smaller ensemble spread for smaller γ (see Figure 6.1). At later lead times, the CRPSS for a larger γ becomes positive. This is attributed to an increase in under-dispersion for smaller γ.

In the dynamically more active month of January (Buizza and Palmer, 1995), the best of the three scaling factors tested is $\gamma = 0.2$. The $\gamma = 0.25$ ensemble has always negative CRPSS values over the $\gamma = 0.2$ ensemble, indicating that the $\gamma = 0.2$ has forecast skill over the $\gamma = 0.25$ ensemble. The lowest CRPSS values are again connected to the Z500 ensemble. In January, Z500 has a very negative CRPSS at early lead times for $\gamma = 0.2$ and skews the comparison in favor of $\gamma = 0.1$ (see Figure 6.1(a)). Only at lead times greater than 186 h does this advantage revert, with positive CRPSS values for the $\gamma = 0.2$ ensemble. Generally, the CRPSS of $\gamma = 0.2$ increases with increasing lead time. In total, the CRPS of the $\gamma = 0.1$ ensemble is marginally lower. If the Z500 ensemble is excluded, $\gamma = 0.2$ has a lower average CRPS. Similarly, the comparison of $\gamma = 0.25$ and $\gamma = 0.1$ shows a lower average CRPS of the $\gamma = 0.1$ ensemble when including Z500. Excluding Z500 gives forecast skill to the $\gamma = 0.25$ ensemble. Again, the CRPSS of the $\gamma = 0.25$ ensemble increases with increasing lead time owing to the strong under-dispersion of the $\gamma = 0.1$ ensemble at later lead times.

In July, CRPSS relationships for the different γ become more complex. July is dynamically less active than January (Buizza and Palmer, 1995). So one could expect that a larger γ might be necessary to account for the smaller A-SV perturbations over the European domain in July. Indeed, $\gamma = 0.2$ has, on average, positive CRPSS over $\gamma = 0.1$ (see Figure 6.1(b)). This is attributed to the strong under-dispersion of the $\gamma = 0.1$ ensemble (not shown). For $\gamma = 0.25$, the CRPSS relationships are very interesting. There is a strong negative CRPSS at early lead times for Z500. For all other forecast variables, the CRPS is closely matched to the $\gamma = 0.2$ ensemble. At early lead times, the $\gamma = 0.2$ ensemble has a slightly lower CRPS. At lead times beyond 120 h, the CRPS of the $\gamma = 0.25$ ensemble is lower. Averaged across all forecast variables and lead time bins, the $\gamma = 0.2$ ensemble has marginally lower average CRPS than the $\gamma = 0.25$ ensemble. Excluding Z500 yields a marginal forecast skill for the $\gamma = 0.25$ ensemble. The comparison between $\gamma = 0.25$ and $\gamma = 0.1$ shows that the $\gamma = 0.25$ ensemble has, on average, a positive CRPSS. Here, the CRPSS increases with lead time.



(a) CRPSS for January 2022



(b) CRPSS for July 2022

Figure 6.1: CRPSS in percent for different scaling factors γ of the A-SV-based Pangu-Weather ensemble for the month of January and July 2022 and the forecast variables T850, Z500, T2M, U10M, and V10M. The left heat map gives the CRPSS of the $\gamma = 0.2$ ensemble over the $\gamma = 0.1$ ensemble, the middle heat map gives the CRPSS of the $\gamma = 0.25$ ensemble over the $\gamma = 0.2$ ensemble and the right heat map gives the CRPSS of the $\gamma = 0.25$ ensemble over the $\gamma = 0.1$ ensemble. Purple indicates that the first factor has worse forecast skill than the second factor, green indicates that the first factor has forecast skill over the second factor. Per variable, the forecasts are grouped into four lead time bins to highlight lead time dependencies of the forecast skill.

If the analysis averages over both months, the $\gamma = 0.2$ ensemble has the lowest CRPS values in the European domain among the three scaling factors tested. Hence, $\gamma = 0.2$ is the scaling factor chosen for the ensemble generation. Future work should consider that the CRPSS between the different γ exhibits a seasonal cycle due to changes in the spread-RMSE relationship. Thus, the CRPS of the A-SV-based Pangu-Weather ensemble forecast could be further reduced by including a seasonal dependency in γ during ensemble initialization.

6.2 Ensemble metrics

In this section, standard ensemble evaluation metrics are applied to the A-SV-based Pangu-Weather ensemble. These include the RMSE, the spread, and the bias of the ensemble forecast. Previous studies show that Pangu-Weather exhibits significant biases for the forecast variables tested (e. g., Ben Bouallègue et al., 2024). For a well calibrated ensemble, RMSE and spread values should match (Fortin et al., 2014). First, this section evaluates the spread-RMSE relationship. In a second step, the seasonal dependence of spread − RMSE is analyzed. Third, the ensemble mean bias of A-SV-based Pangu-Weather ensemble is evaluated.

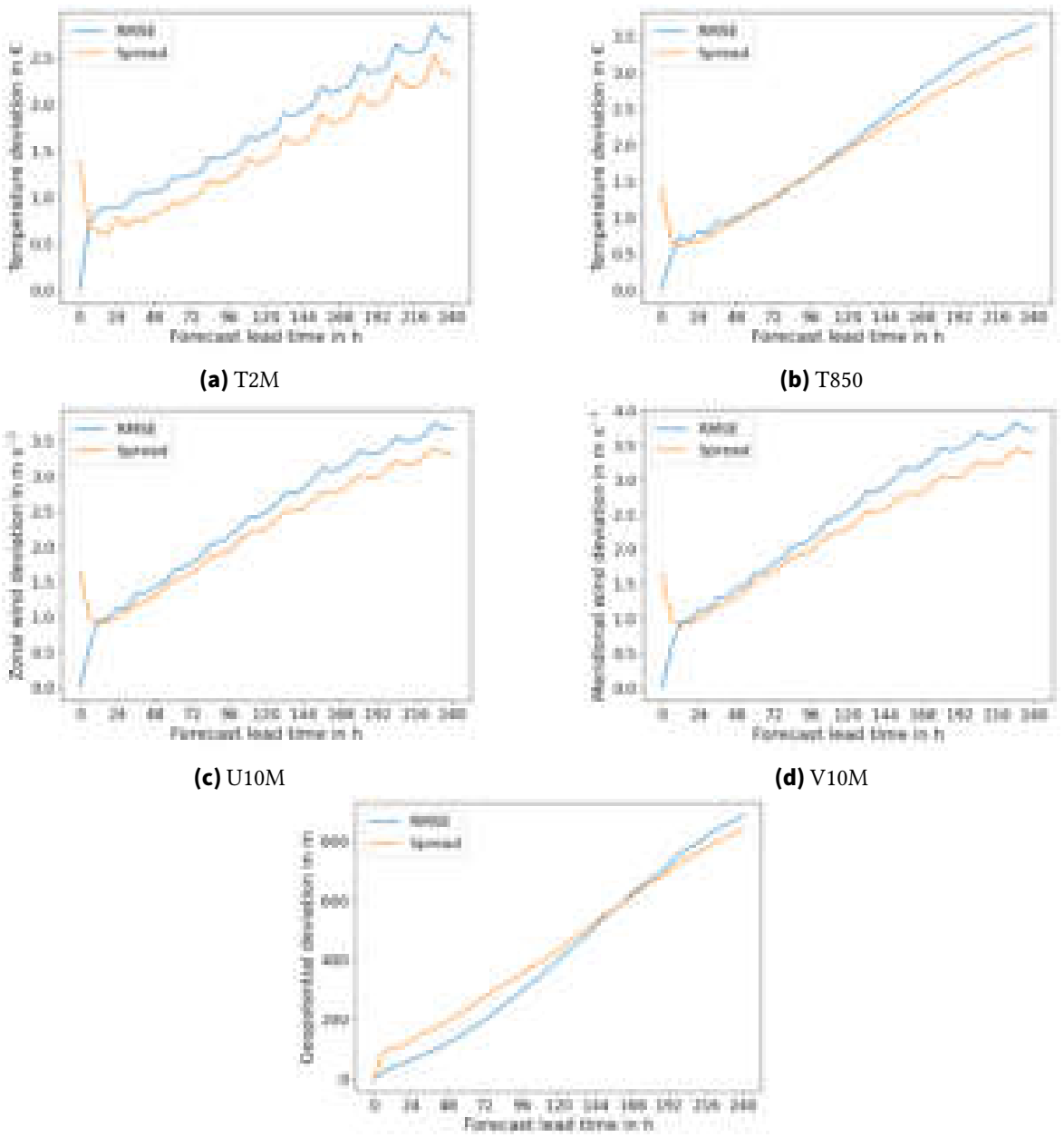
6.2.1 Spread-RMSE relationship

An ensemble with similar spread and RMSE values is desirable because, in this case, high spread values indicate that the forecast is uncertain and large deviations are possible, i. e., large RMSE values can occur. Analyzing the spread-RMSE relationships for all Pangu-Weather forecast variables is beyond the scope of this work. Thus, Figure 6.2 shows the domain-averaged spread and RMSE values for the five forecast variables tested.

All forecast variables tested exhibit a similar spread-RMSE pattern, depending on the lead time. At initial time, the spread is significantly larger than the RMSE. With the first Pangu6 inference, the spread of the ensemble decreases significantly. The initial spread reduction is caused by a denoising process connected to the internal structure of Pangu-Weather, due to down-sampling and subsequent up-sampling in the vision encoder (Bi et al., 2023). Hence, the initial time perturbations must be significantly larger than the desired spread at early lead times to account for the initial denoising of Pangu-Weather. After the initial reduction in spread, the spread increases, as expected, with lead time. For perturbed forecast variables, it takes four to six days for the spread to reach the initial time values again. Still, the ensemble is typically over-dispersive at early lead times. Beyond day 1, the ensemble becomes under-dispersive because the spread increases more slowly than the RMSE.

The exception here is Z500. Figure 6.2(e) shows that for Z500, the ensemble remains over-dispersive for lead times up to 150 h. At longer lead times, the Z500 ensemble is also under-dispersive. In the Arnoldi method for SV calculation, the A-SVs are optimized for the dry total energy norm (see Equation 2.7). The dry total energy norm is independent of the geopotential. Hence, Z500 is unperturbed at initial time and adjusts freely to the perturbations of the other forecast variables. Hakim and Masanam (2024) denote that Pangu-Weather produces mixed results for tests of geostrophic adjustment processes in Pangu-Weather. Hence, Z500 may also produce a deficient response to A-SV perturbations of the 500 hPa wind field, which might lead to the initial over-dispersion of the Z500 ensemble. It is beyond the scope of this work to assess this issue in more detail.

In general, the spread-RMSE patterns for all forecast variables tested indicate that the A-SV-based Pangu-Weather ensemble is close to the desired calibration of equal spread and RMSE values (Fortin et al., 2014). The tests for January and July 2022 show that, on average,



(a) T2M

(b) T850

(c) U10M

(d) V10M

(e) Z500

Figure 6.2: Time series of the domain-averaged spread (orange) and RMSE (blue) for the A-SV-based Pangu-Weather ensemble averaged across the year 2022 for the forecast variables T2M, T850, U10M, V10M, and Z500. A larger spread than RMSE indicates that the ensemble is over-dispersive. Smaller spread than RMSE shows that the ensemble is under-dispersive. For a well-calibrated ensemble, spread and RMSE should match.

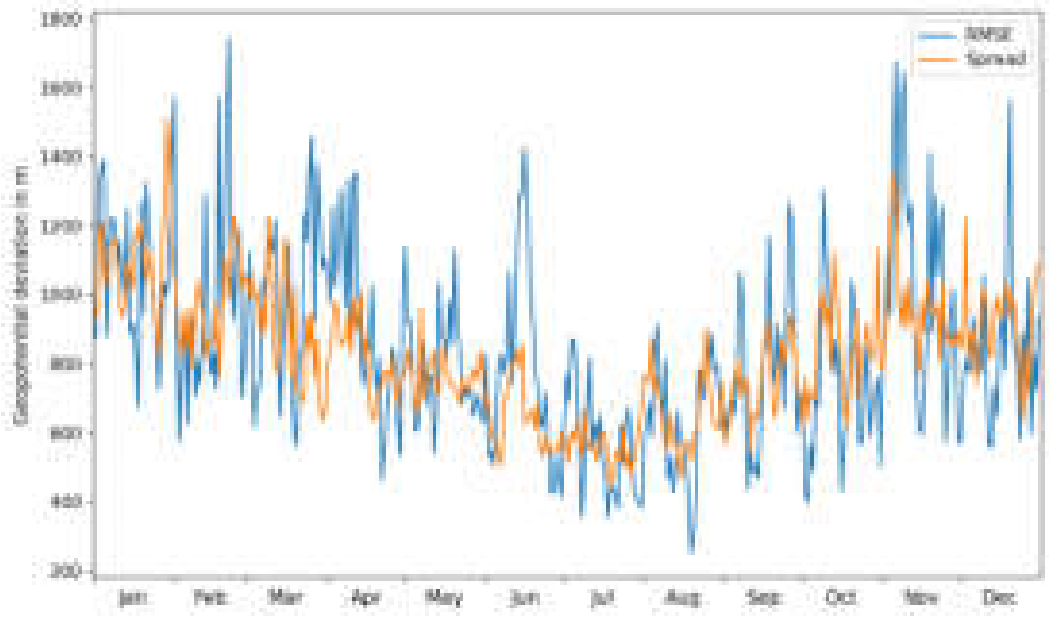


Figure 6.3: Time series of the domain-averaged spread (orange) and RMSE (blue) of the Z500 ensemble at a lead time of 240 h. Each point in the time series corresponds to one of the daily forecast initializations in the year 2022.

the scaling factors $\gamma \in \{0.1, 0.25\}$ perform worse. For $\gamma = 0.1$, the ensemble is strongly under-dispersive in January. In July, the spread is still significantly smaller than the RMSE. In contrast, the $\gamma = 0.25$ ensemble is slightly over-dispersive in January. In July, the ensemble is significantly over-dispersive. Even for $\gamma = 0.25$, the ensemble is under-dispersive at lead times beyond 200 h.

6.2.2 Seasonal dependence of spread and RMSE

Section 6.2.1 highlights that the A-SV-based Pangu-Weather ensemble is over-dispersive on day 1. Independent of the forecast variable, the ensemble becomes under-dispersive at a certain lead time horizon. The following section provides an overview of the seasonal dependency of the spread and RMSE. Figure 6.3 shows both the spread and RMSE at a lead time of 240 h for all forecasts in the year 2022. Both exhibit a general seasonal cycle, with higher values in the winter half of the year and lower values in the summer. This seasonality is consistent with previous studies (Ben Bouallègue et al., 2024). Similar patterns are also present in other forecast variables tested (not shown).

Figure 6.3 also shows that, for Z500, the ensemble is generally well calibrated. In the domain-average, periods of elevated RMSE are mostly matched by increased spread values. In both late-March to early-April and June, there are periods with significantly greater RMSE than spread. In late-March to early-April, the RMSE and spread show contrasting behavior. In this period, the spread at a lead time of 240 h is fairly low in many forecasts, indicating that the ensemble is fairly confident. However, the elevated RMSE values indicate that the

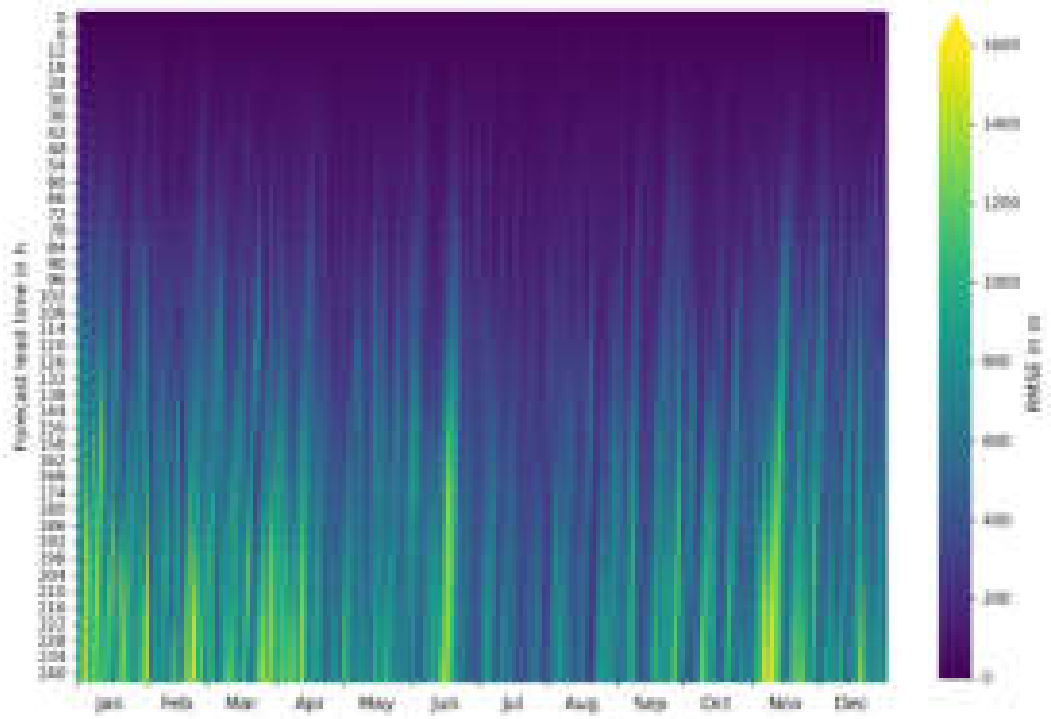


Figure 6.4: Heat map of the domain-averaged RMSE of the Z500 ensemble for the daily ensemble forecasts in the year 2022. One column in the heat map corresponds to one forecast initialization. The forecast lead time increases from top to bottom.

ensemble mean shows significant deviations from ERA5. Hence, the ensemble is not well calibrated during this period. In general, the RMSE peaks are typically much larger than the spread peaks.

The analysis of seasonal dependencies of the spread and RMSE patterns in Figure 6.3 is limited to a single lead time. Figure 6.4 shows the daily domain-averaged RMSE for all lead times. From long lead times, RMSE peaks extend to shorter lead times. High RMSE values occur consistently in many consecutive forecast initializations. At times, periods of high/low RMSE span multiple weeks. For some forecast initializations, there are *windows of opportunity* in which the RMSE remains very low into the medium-range. Typically, these occur in the summer months. Occasionally, *windows of opportunity* also occur in other seasons. Figure 6.4 generally shows that the variability of the RMSE is mainly driven by synoptic events.

Similarly to Figure 6.4, the heat map of spread – RMSE shows that, for the year 2022, alternating periods of higher spread or higher RMSE values occur on synoptic timescales. Up to lead times of 72 h, the Z500 ensemble is never under-dispersive. At long lead times, Figure 6.5 shows that under-dispersive and over-dispersive forecasts quickly alternate. There are two striking periods where the Z500 ensemble is strongly under-dispersive (late-March to early April and mid-June). Both are located in the first half of the year. In the second half of the year, the Z500 ensemble is more often over-dispersive. Similar patterns are also observed for the T850 ensemble (not shown). To a lesser extent, this seasonality of

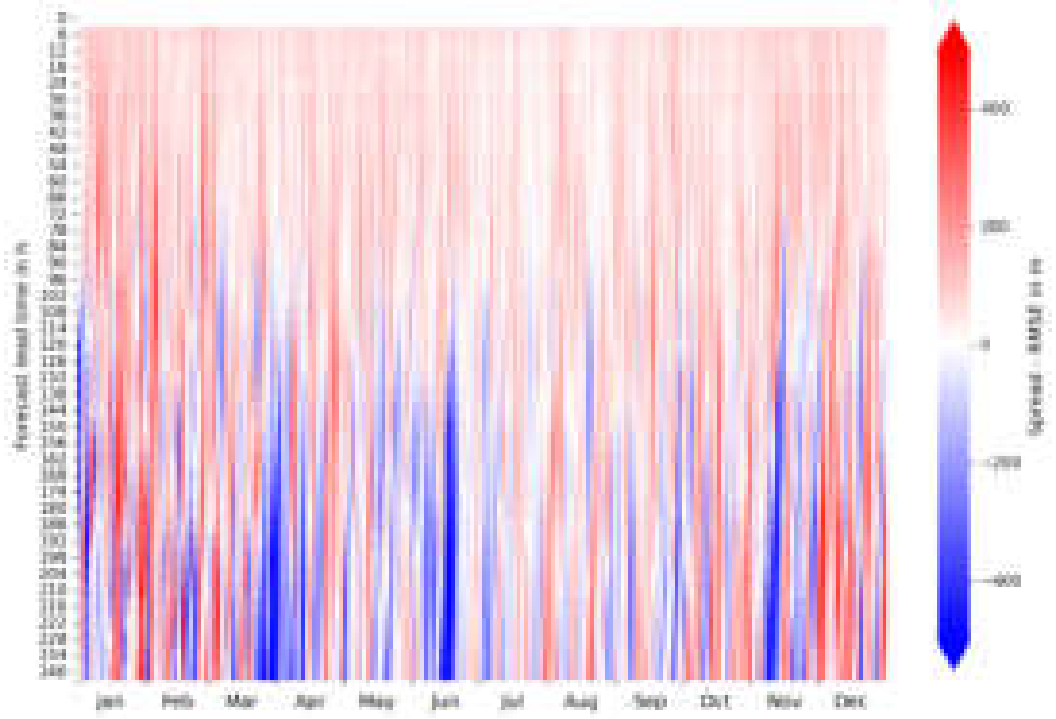


Figure 6.5: Heat map of the domain-averaged spread − RMSE of the Z500 ensemble for the daily ensemble forecasts in the year 2022. One column in the heat map corresponds to one forecast initialization. The forecast lead time increases from top to bottom. Red values indicate larger spread than RMSE values (i. e., the ensemble is over-dispersive). Blue values show larger RMSE than spread (i. e., the ensemble is under-dispersive).

the spread − RMSE relation also occurs for the surface variables tested. Like the RMSE, Figure 6.5 shows that if the ensemble is either under- or over-dispersive, this state is retained across the entire forecast horizon.

The Spread of the T2M ensemble shows a distinct diurnal spread pattern at later lead times. Figure 6.6 shows that this diurnal spread pattern is more pronounced in the summer months. In contrast, the spread is dominated by synoptic events during the winter months. Thus, the diurnal spread cycle is not as prominent. This spread pattern can be attributed to the Pangu-Weather temporal aggregation algorithm (Bi et al., 2023; Selz and Craig, 2023). Each 00 UTC forecast is performed by Pangu24, which leads to a strong smoothing of the forecast (Selz and Craig, 2023). The first Pangu6 inference leads to a significant increase in spread at 06 UTC. At 12 UTC, the spread cycle reaches its diurnal maximum. With the 18 UTC forecast, the spread of the ensemble typically decreases marginally. The new 00 UTC forecast is performed again by Pangu24 using the previous 00 UTC forecast as initial conditions. Subsequently, the spread of the ensemble decreases considerably. Different Pangu-Weather forecast models are known to cause discontinuities in the forecast (Hakim and Masanam, 2024; Bonavita, 2024). The seasonal cycle of the diurnal spread cycle is also interpretable in a meteorological context. In summer, changes in diurnal insolation affect midday to afternoon temperatures most strongly. Thus, the temperature spread around midday is highest in

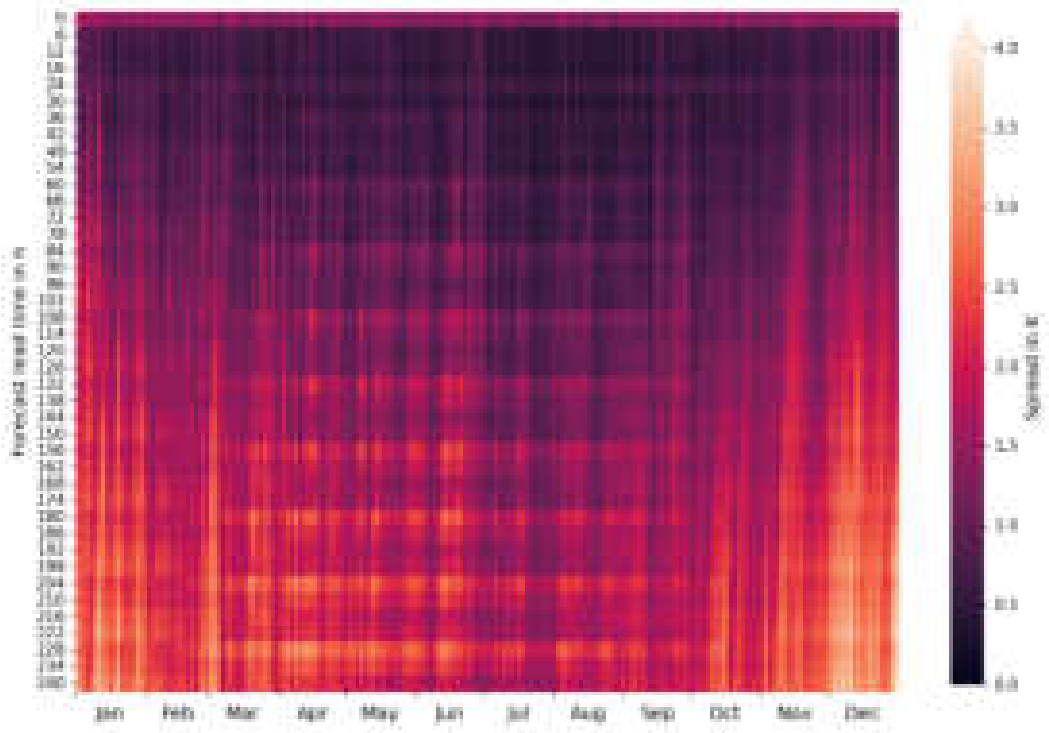


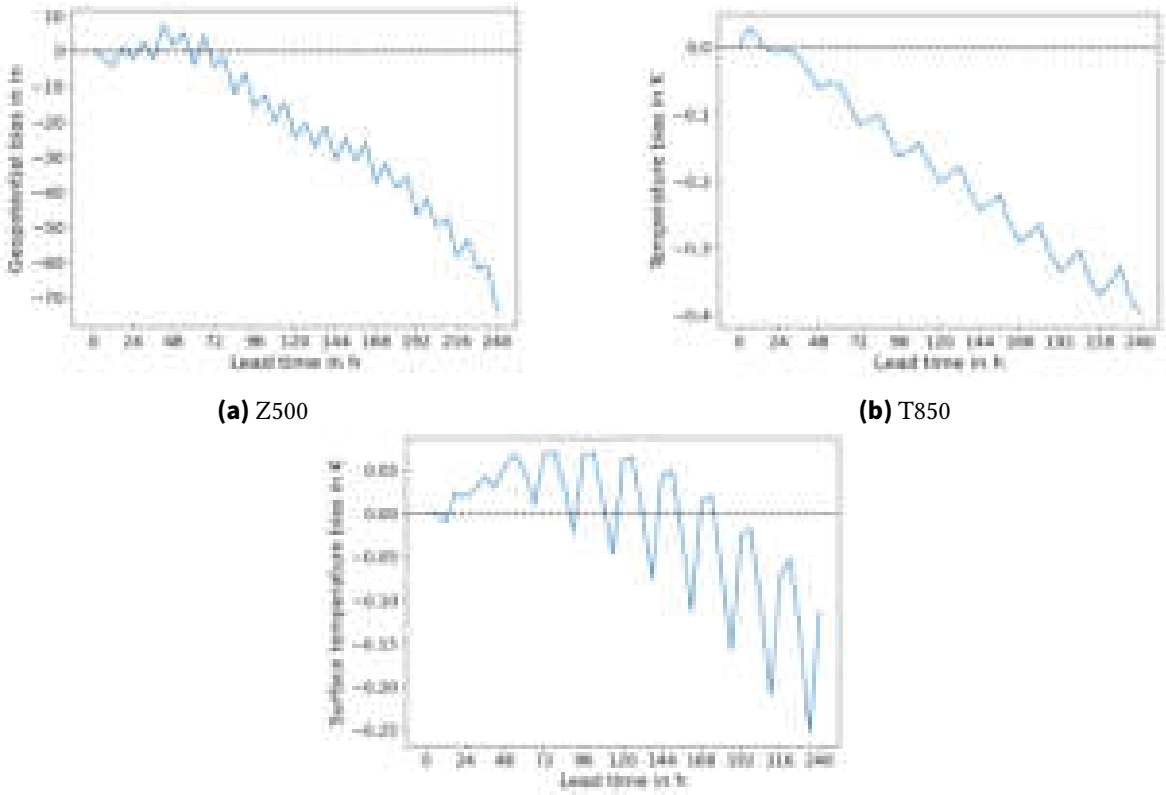
Figure 6.6: Heat map of the domain-averaged spread of the T2M ensemble for the daily ensemble forecasts in the year 2022. One column in the heat map corresponds to one forecast initialization. The forecast lead time increases from top to bottom.

summer compared to that in the other seasons. The existence of the diurnal spread cycle shows that this interaction is represented in the ML-based Pangu-Weather model.

For the upper-atmospheric forecast variables, the domain-specific spread – RMSE patterns show little coherent signal. For the surface variables, there are characteristic spread – RMSE patterns (not shown). For the surface wind, U10M and V10M, the ensemble is often over-dispersive over the ocean at early lead times. In contrast, the spread of the ensemble forecast over land is typically too small. At early lead times, the T2M ensemble shows persistent signals of over-dispersion over the ocean and under-dispersion over land (not shown). Hence, Pangu-Weather underestimates, on average, the temperature range over land. Over the ocean, the temperature range is typically too large. This implies that Pangu-Weather is overly sensitive to the general atmospheric dynamics over the ocean.

6.2.3 Bias

Previous studies have shown that Pangu-Weather exhibits significant biases that develop with forecast lead time. These are considerably larger than the biases of state-of-the-art NWP models (Ben Bouallègue et al., 2024). Very similar bias patterns for T850 and Z500 are also present in the ensemble mean (see Figure 6.7). The Z500 ensemble has little bias up to a lead time of 100 h. For longer lead times, a strong negative bias develops in Figure 6.7(a).



(a) Z500

(b) T850

(c) T2M

Figure 6.7: Time series of the domain-averaged ensemble-mean bias of the Pangu-Weather ensemble averaged across the year 2022 for the forecast variables Z500, T850, and T2M. The dashed black line indicates the zero line of the bias.

For T850, the negative bias develops beyond day 1. The linear increase in the T850 bias is superimposed by a diurnal bias cycle that develops at lead times beyond 50 h. The surface wind variables, U10M and V10M, do not show a significant bias ($< 0.1\,\mathrm{m\,s^{-1}}$) for all lead times.

The T2M ensemble exhibits a stronger diurnal bias cycle than the T850 ensemble. The daily average bias remains close to zero for the first seven days of the forecast, as shown in Figure 6.7(c). Beyond day 7, the bias becomes increasingly negative in the daily average. The diurnal bias cycle increases with lead time. At lead times beyond 200 h, the diurnal bias cycle reaches an amplitude of up to 0.2 K. Similarly to Figure 6.6, the T2M ensemble-mean bias for every ensemble forecast is shown in Figure 6.8. The heat map shows that the diurnal bias cycle shares a strong similarity with the diurnal spread cycle of the T2M ensemble (see Figure 6.6). The similarities extend to both the cycle itself and the seasonal dependence of these diurnal cycles. Figure 6.8 shows that, at 00 UTC, the bias is closest to zero. With the first diurnal Pangu6 inference, a negative bias develops. At 12 UTC, the negative bias of the forecast increases further. With the third diurnal Pangu6 inference, the diurnal minimum of the bias cycle is reached at 18 UTC. Due to the temporal aggregation algorithm, the next forecast is generated by Pangu24. Pangu24 uses the previous 00 UTC forecast as initial

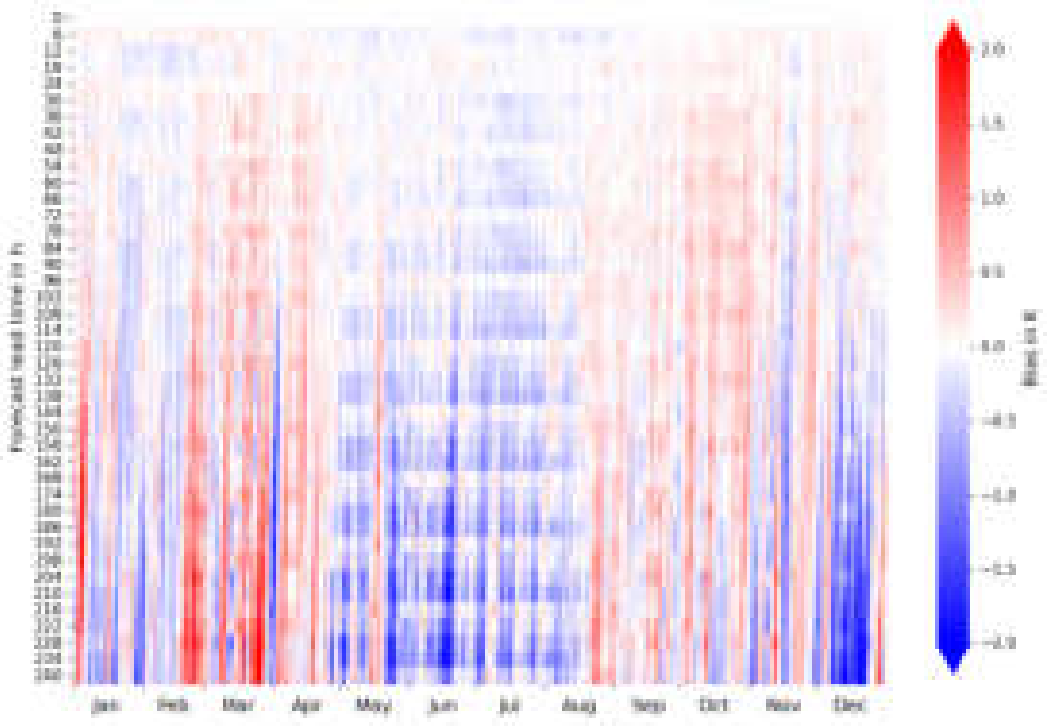
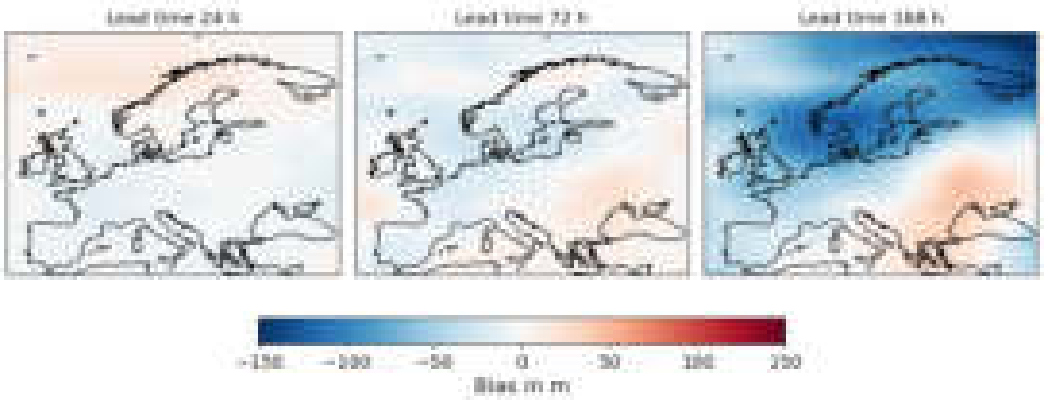


Figure 6.8: Heat map of the domain-averaged bias of the T2M ensemble for the daily ensemble forecasts in the year 2022. One column in the heat map corresponds to one forecast initialization. The lead time of the forecast increases from top to bottom. Red values indicate that the mean of the Pangu-Weather ensemble is warmer than ERA5 and blue values indicate that ERA5 is warmer than the ensemble mean.

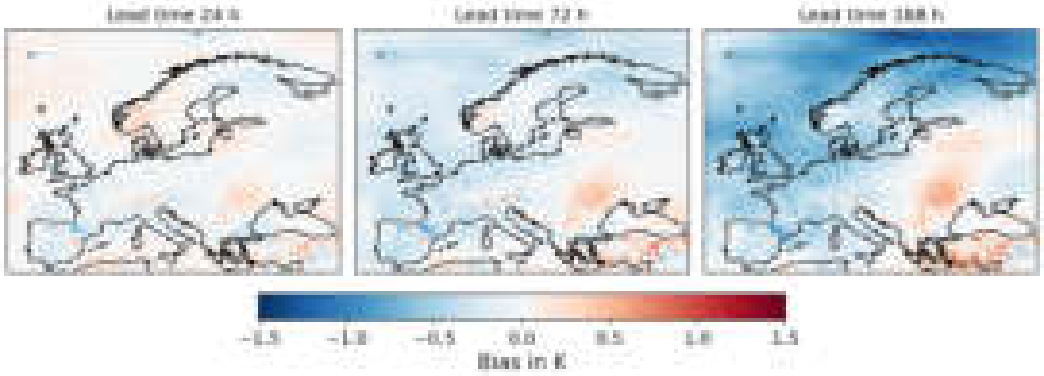
conditions (Bi et al., 2023). For this forecast, the negative bias is strongly reduced, similar to the diurnal spread cycle. Thus, the diurnal bias cycle can also be attributed to the temporal aggregation algorithm.

In addition, the diurnal bias cycle exhibits a strong seasonality (see Figure 6.8). In the summer half of the year, a strong negative bias develops at long lead times. In contrast, late winter to early spring shows mostly a strong positive bias at later lead times. In autumn, there is a tendency for positive biases at intermediate lead times. Whether this seasonality of the bias is related only to the year 2022 can be determined only by extending the ensemble forecast across more years. This is beyond the scope of this work.

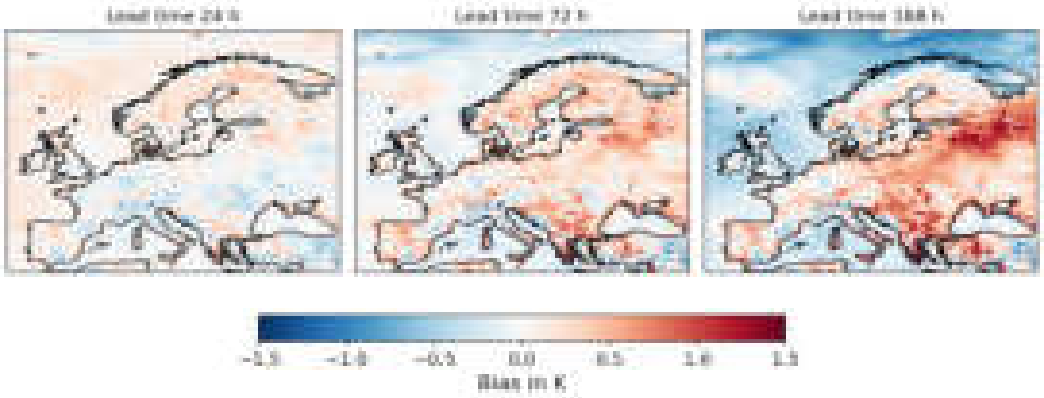
The previous analysis in Section 6.2.1 already shows that cross-variable consistency in Pangu-Weather can be a concern. Figure 6.9 shows the averaged ensemble-mean bias of Pangu-Weather in 2022. Both the Z500 and T850 ensembles have positive biases that extend from the Central Mediterranean to Western Russia. In addition, both have negative biases that stretch from the Iberian Peninsula to Scandinavia at longer lead times. Both the Z500 and T850 biases are co-located. This indicates cross-variable consistency in terms of the ensemble-mean bias.



(a) Z500



(b) T850



(c) T2M

Figure 6.9: Ensemble-mean bias of the Pangu-Weather ensemble over ERA5 averaged across all forecasts for the year 2022 for the forecast variables Z500, T850, and T2M.

Similarly to the positive Z500 and T850 biases over Eastern Europe, the strong positive T2M biases extend from the Balkans into Western Russia at longer lead times (see Figure 6.9(c)). A positive geopotential anomaly over Eastern Europe can be caused by an anomalous occurrence frequency of ridges in the region. These ridges lead to increased warm air advection, which can also induce a positive surface temperature anomaly in the region.

In addition to under-dispersion, Figure 6.9(c) highlights that the T2M ensemble also has a negative bias over the ocean.

The analysis of spread, RMSE, and bias shows that the A-SV-based Pangu-Weather ensemble has some deficiencies in terms of ensemble calibration. However, the analysis outlines the dependence of the ensemble forecast on synoptic patterns, which shows that A-SVs impose flow-dependent perturbations in the initial conditions of the ensemble. Furthermore, the analysis highlights that the principles of the ensemble initialization process of the ECMWF IFS, discussed in Section 2.5, are successfully transferred to the ML-based Pangu-Weather framework. The ensemble-mean biases of the A-SV-based Pangu-Weather ensemble are consistent with previous studies (Ben Bouallègue et al., 2024). In addition, the biases exhibit promising cross-variable consistency.

6.3 Post-processing

There are known intrinsic biases in Pangu-Weather (Ben Bouallègue et al., 2024). Section 6.2.3 shows that the A-SV-based Pangu-Weather ensemble exhibits similar biases in the ensemble mean. As the CRPS penalizes biases (Gneiting and Raftery, 2007), correcting for the ensemble-mean bias could reduce the average CRPS of the ensemble and thus improve the forecast skill. The bias correction of forecast variable χ with the bias of said forecast variable $\hat{\chi}$ is defined as

$$\tilde{\chi} = \chi - \hat{\chi} = \chi - \sum_{t=1}^{T} (\bar{\chi} - \psi). \tag{6.1}$$

T is the number of forecast initializations, $\bar{\chi}$ is the ensemble mean averaged among all ensemble members, and ψ is an observation. As mentioned, the analysis uses ERA5 as the ground truth ψ. This bias correction is a simplistic post-processing approach that is applied to the A-SV-based Pangu-Weather ensemble (Gneiting et al., 2005). The CRPSS of the post-processed ensemble forecast over the raw ensemble forecast is calculated, to assess the impact of the bias correction on the forecast skill (see Equation 6.1). Positive CRPSS values indicate that the bias-corrected ensemble forecast performs better than the raw ensemble forecast (see Equation 3.5).

Figure 6.10 shows that the greatest improvements in forecast skill, in terms of CRPSS, are derived for the T2M ensemble. Here, the CRPS of the post-processed ensemble forecast decreases by up to 5 % due to the bias correction. For all surface variables tested, the largest CRPSS values are found in the early lead time bin. For Z500, the bias correction provides only a minor decrease in CRPS (< 1 %). Figure 6.7 shows that Z500 and T850 both have significant biases toward increasing lead times. Hence, it is not trivial why the CRPSS values of the post-processed T850 ensemble are about three times as large as those for the Z500 ensemble. For both, the respective CRPSS values remain nearly constant, independent of the lead time bin. On average, the CRPSS values are the smallest in the medium-range lead time bin between 186 and 240 h. This implies that, at early lead times, the impact of the bias

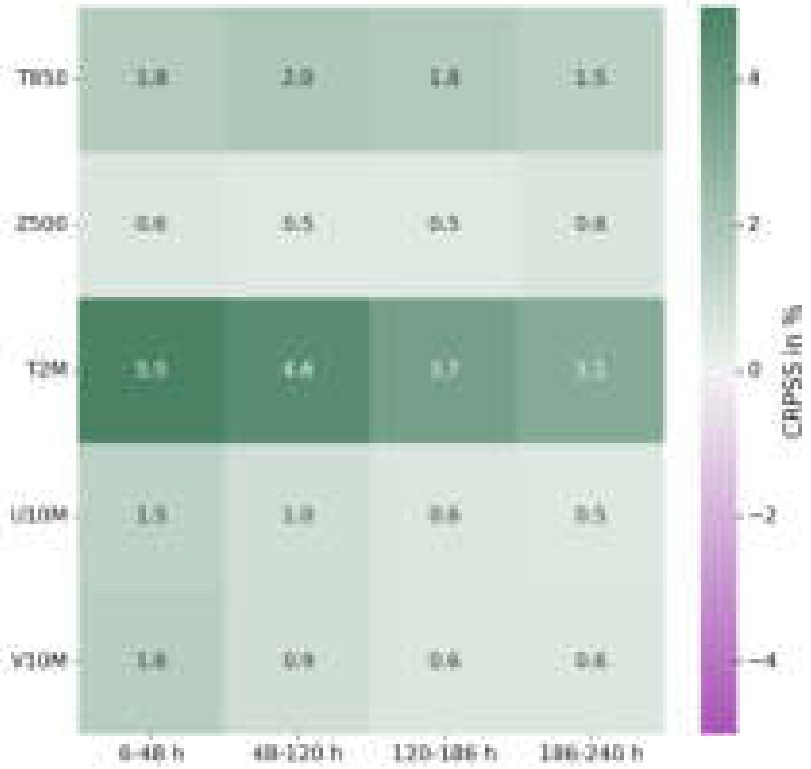


Figure 6.10: CRPSS in percent of the bias-corrected ensemble forecast over the raw ensemble forecast for the year 2022 and the forecast variables T850, Z500, T2M, U10M, and V10M. Per variable, the forecast is grouped into four lead time bins.

on the CRPS and thus the ensemble forecast skill is the largest. At longer lead times, the ensemble-mean bias becomes a smaller contributor to the CRPS.

The bias correction of the ensemble-mean bias demonstrates that even a simple post-processing method can significantly improve the forecast skill of the A-SV-based ensemble. More sophisticated post-processing approaches are likely to further improve the forecast skill of the ensemble by lowering its CRPS values (Gneiting et al., 2005).

6.4 Comparison to ECMWF IFS ensemble forecast skill

In the sections above, the A-SV-based Pangu-Weather ensemble is evaluated using a variety of metrics. A simple post-processing approach is proposed in Section 6.3. Based on the forecast variable, the bias correction provides a significant decrease in the CRPS values. In particular, the CRPS of the T2M ensemble decreases by more than 5 % at early lead times. The following section evaluates the A-SV-based Pangu-Weather ensemble against the ECMWF IFS ensemble. The IFS also uses SVs to generate an initial condition ensemble (Leutbecher and Palmer, 2008). Like in Section 6.3, the evaluation first focuses on the headline CRPSS heat maps. From these, the analysis shifts to the time series of the domain-averaged CRPS

to further examine the lead time dependent forecast skill of the A-SV-based Pangu-Weather ensemble. For selected lead times, the CRPSS of the Pangu-Weather ensemble over the IFS ensemble is analyzed with respect to the local forecast skill of the Pangu-Weather ensemble.

6.4.1 CRPSS heat maps

Similarly to Section 6.3, the forecast skill of the A-SV-based Pangu-Weather ensemble is validated against the ECMWF IFS ensemble forecast. Thus, the IFS ensemble serves as the reference ensemble $CRPSS_{\text{ref}}$ in Equation 3.5. Figure 6.11 shows that the five tested forecast variables form three groups of relative forecast skill. By far, the worst forecast skill is achieved for the Z500 ensemble in Pangu-Weather. In the early lead time bin, the CRPSS reaches values close to -0.3. With increasing lead time, the CRPSS of the Z500 ensemble remains negative. However, there is an increasing CRPSS trend. This indicates that the Pangu-Weather ensemble forecast improves compared to the IFS ensemble forecast. The poor performance of the Z500 ensemble is related to the strong over-dispersion of the Z500 ensemble at early lead times (Hersbach, 2000). This over-dispersion of the Z500 ensemble is attributed to the A-SV computation. In the Arnoldi method, A-SVs are optimized for the dry total energy norm (see Equation 2.7), which is independent of the geopotential (Winkler and Denhard, 2025). A deficient response of the geopotential field to the A-SV perturbations may induce this over-dispersion. The bias correction leads only to a marginal decrease in CRPS (see Figure 6.10). Thus, the post-processed Z500 ensemble still has significantly worse forecast skill compared to the IFS ensemble at early lead times.

In stark contrast, the T2M ensemble in Pangu-Weather has significant forecast skill over the ECMWF IFS ensemble at early lead times. The forecast skill persists into intermediate lead times. At long lead times, the raw A-SV-based Pangu-Weather ensemble exhibits slightly negative CRPSS values. Due to the significant CRPS decrease in the post-processed T2M ensemble, the forecast skill persists in the long lead time bin for the post-processed ensemble. The forecast skill for the T2M ensemble implies that Pangu-Weather has a better ensemble forecast for the T2M without the need for extensive surface parameterization schemes used in the IFS ensemble (e. g., Viterbo and Beljaars, 1995).

The remaining forecast variables form the third group in terms of their CRPSS values. These have a fairly constant CRPSS independent of the lead time bin. All variables have a negative CRPSS over the IFS ensemble forecast. However, the forecast skill is higher compared to the Z500 ensemble, with CRPSS values ranging between -12 and $-5\,\%$. In general, it is remarkable that the bias correction reduces the CRPS of the best forecast variable, T2M, the most. At early lead times, the post-processed T2M ensemble has a domain-averaged CRPSS of 0.2 over the IFS ensemble. For Z500, the CRPS reduction of the post-processed ensemble is only minor, despite the strong bias and the strongly negative CRPSS of the Z500 ensemble.

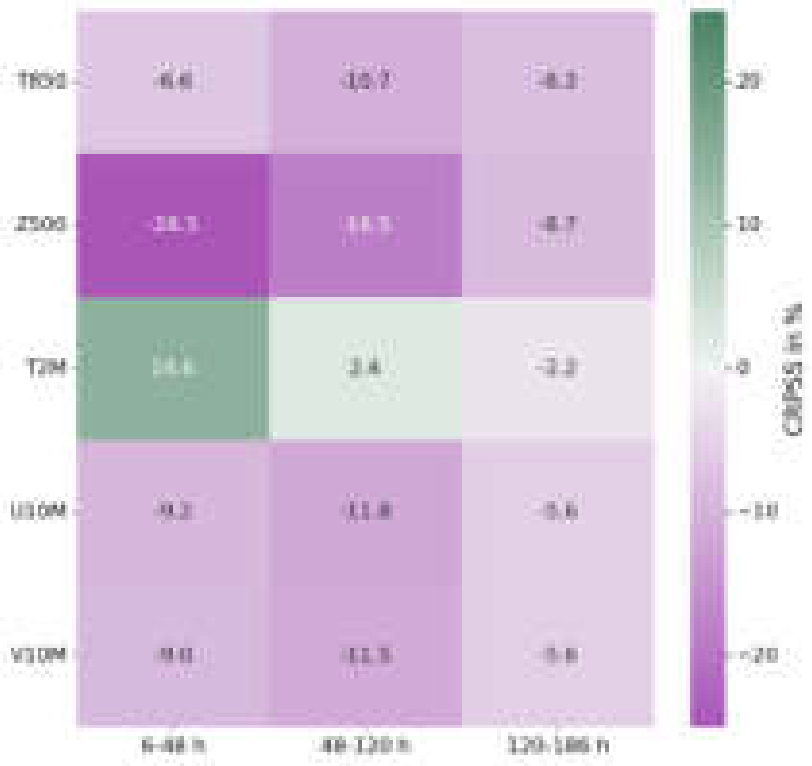
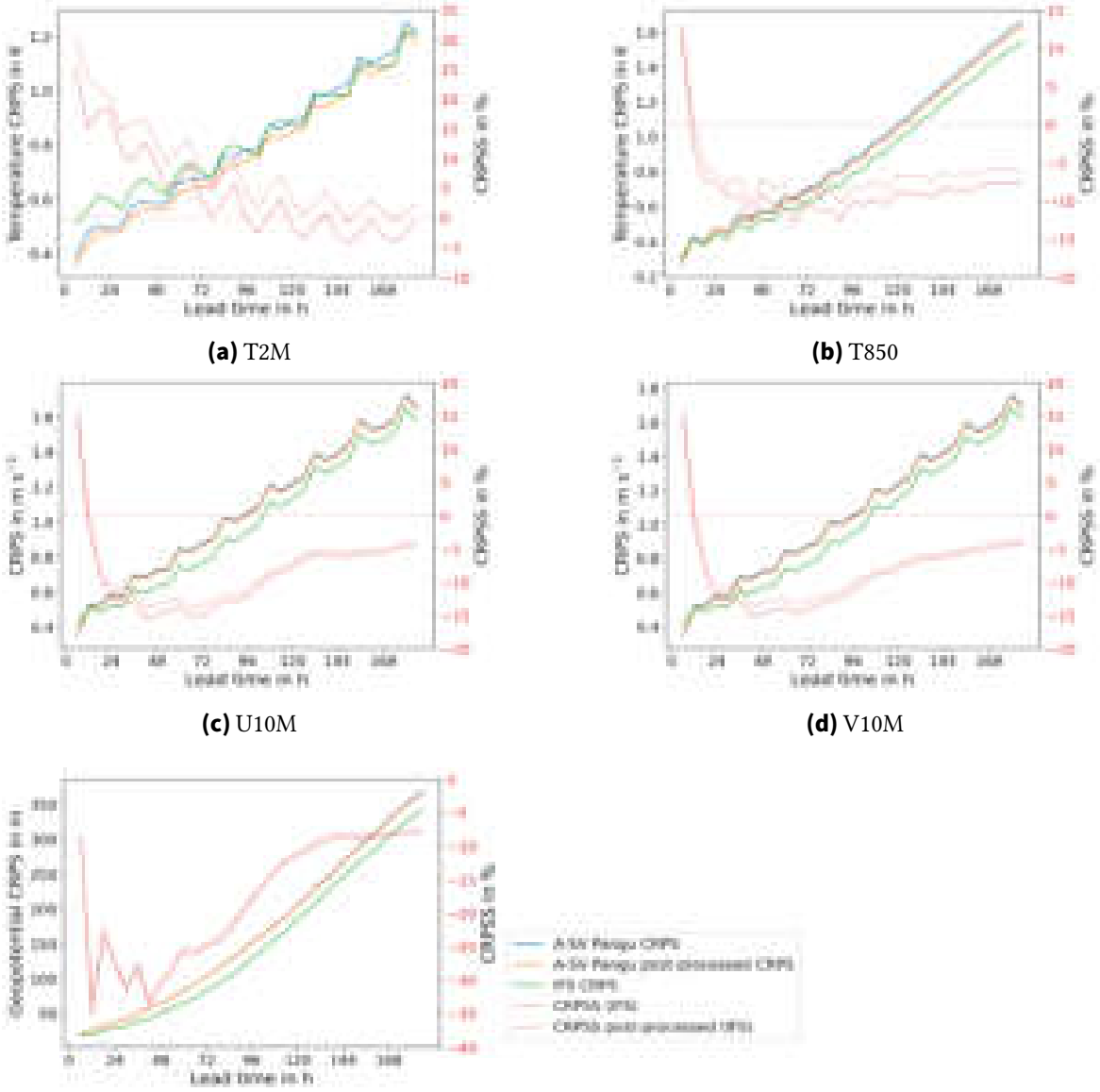


Figure 6.11: CRPSS in percent of the A-SV-based Pangu-Weather ensemble forecast over the ECMWF IFS ensemble forecast for the year 2022 and the forecast variables T850, Z500, T2M, U10M, and V10M. Per variable, the forecasts are grouped into three lead time bins.

6.4.2 CRPS time series comparison to ECMWF IFS

The analysis in Section 6.4.1 shows that the A-SV-based Pangu-Weather ensemble achieves forecast skill for some forecast variables over the ECMWF IFS ensemble. The heat map in Figure 6.11 averages the CRPSS across different lead time bins. This method provides a quick overview of the headline scores. A detailed analysis of the CRPSS of the A-SV-based Pangu-Weather ensemble is achieved through a comprehensive investigation of the lead time dependence of the CRPSS. As anticipated, the performance of the different forecast variables can be categorized into three distinct patterns. The three groups consist of the same forecast variables as before.

The first group is given by the overall best-performing variable, T2M. Previous analysis shows that the T2M ensemble has forecast skill over the ECMWF IFS ensemble in the early and intermediate lead time bins. With the first Pangu6 inference, the CRPSS of the Pangu-Weather ensemble is approximately 0.25, indicating significant forecast skill for the Pangu-Weather ensemble. After day 1, the CRPSS of the Pangu-Weather ensemble rapidly decreases, as shown in Figure 6.12(a). At a lead time of 78 h, the CRPSS reaches zero for the first time. This implies that the A-SV-based Pangu-Weather ensemble has equal forecast skill compared to the IFS ensemble, averaged over the European domain.



(a) T2M

(b) T850

(c) U10M

(d) V10M

(e) Z500

Figure 6.12: Time series of the domain-averaged CRPS for the raw (blue) and the bias-corrected (orange) Pangu-Weather ensemble up to a lead time of 186 h averaged across the year 2022 for the forecast variables T2M, T850, U10M, V10M, and Z500. The CRPS of the ECMWF IFS ensemble is indicated by the green lines. The CRPSS of the raw Pangu-Weather ensemble over the IFS ensemble is shown by the red line and the CRPSS of the bias-corrected ensemble is shown by the dashed red line. If shown, the dotted red line indicates where $CRPSS = 0$. Values above this line imply that Pangu-Weather has forecast skill over the IFS ensemble forecast. Values below this line indicate that the Pangu-Weather ensemble forecast is less accurate than the IFS ensemble forecast.

Additionally, the CRPS time series for the T2M Pangu-Weather ensemble shows a distinct diurnal cycle. All forecasts are initialized at 00 UTC. Due to the Pangu-Weather temporal aggregation algorithm, the forecasts for the valid times 06, 12 and 18 UTC are predicted by Pangu6 (Bi et al., 2023). The 00 UTC forecast is computed by Pangu24, using the 00 UTC forecast from the previous day as initial conditions. The diurnal CRPS cycle in Figure 6.12(a) starts with a decreasing CRPS at early lead times due to the Pangu24 inference at 00 UTC. The first Pangu6 inference, i. e., valid time 06 UTC, leads to a minor CRPS increase. At 12 UTC, the CRPS spikes upward. With the third daily Pangu6 inference (18 UTC), the CRPS increases further at early lead times. Hence, the resultant CRPS time series is a superposition of two different forecast models. It is known that the different Pangu-Weather forecast models lead to discontinuities in the forecast (Hakim and Masanam, 2024; Bonavita, 2024). At later lead times, the daily CRPS maximum occurs at 12 UTC as the diurnal CRPS cycle shifts. At lead times beyond 120 h, the diurnal CRPS cycle of both the Pangu-Weather and the ECMWF IFS ensemble has their respective daily CRPS maximum at 12 UTC. At early lead times, both ensemble forecasts tend to have a daily CRPS maximum at 18 UTC. The Pangu-Weather ensemble shifts faster to the 12 UTC diurnal CRPS maximum than the IFS ensemble. The similar behavior of the diurnal T2M CRPS cycle in both models implies that the ML-based Pangu-Weather model can represent atmospheric dynamics similar to the NWP model.

The offset of the diurnal CRPS cycles leads to a significant CRPSS decrease for Pangu-Weather. Hence, the CRPSS of Pangu-Weather falls below zero for the first time at a lead time of 102 h. This coincides with the first Pangu6 inference on day 4 of the forecast. At later lead times, a clear diurnal CRPSS cycle develops. At 06 UTC, the CRPSS of Pangu-Weather is at its lowest. In contrast, the CRPSS of Pangu-Weather is at its highest at 18 UTC. Up to a lead time of 120 h, the CRPSS of the Pangu-Weather ensemble shows a clear decreasing trend, which is superimposed on the diurnal CRPSS cycle. At lead times beyond 120 h, there is little trend in the CRPSS. The diurnal CRPSS cycle indicates that Pangu-Weather has forecast skill over the ECMWF IFS ensemble for the afternoon/evening temperature forecast. In contrast, the IFS ensemble has forecast skill for the morning temperatures (06 UTC). In the European domain, this diurnal CRPSS cycle might be attributed to the representation of dynamic processes in the boundary layer. Pangu-Weather appears to better capture the impact of afternoon/evening processes on the T2M and shows weaker skill compared to the IFS regarding the impact of late-night/morning processes on the T2M.

The post-processed Pangu-Weather ensemble shows a consistent CRPSS increase over the raw Pangu-Weather ensemble. The key patterns of the CRPSS time series are retained through the bias correction. Only after a lead time of 150 h does the CRPSS of the post-processed Pangu-Weather ensemble drop below zero for the first time. Thus, the post-processed T2M ensemble has consistent forecast skill over the ECMWF IFS ensemble until day 7. At lead times beyond 150 h, the diurnal CRPSS cycle leads to lead time-dependent forecast skill of the post-processed T2M ensemble. At 06 UTC, the diurnal CRPSS cycle leads to negative CRPSS values of the bias-corrected T2M ensemble beyond day 7.

In contrast, the worst performing forecast variable, Z500, has, on average, no skill compared to the ECMWF IFS ensemble. On day 1, the CRPSS of Pangu-Weather quickly decreases to

values as low as -0.35. This decrease shows that the lowest CRPSS values do not occur at the time of strongest over-dispersion[1], but during the subsequent adjustment of Pangu-Weather to the initial time perturbations. Up to a lead time of 72 h, the CRPSS of Pangu-Weather averages around -0.3. Between day 3 and day 5, the CRPSS of Pangu-Weather increases to about -0.07 and remains constant thereafter. For all lead times, there is no diurnal CRPS or CRPSS cycle. As expected, the bias correction has little impact on the CRPSS time series, as shown in Figure 6.12(e).

The remaining forecast variables, T850, U10M, and V10M, form the final group. For these, the CRPSS time series of the Pangu-Weather ensemble follows a very similar pattern. On day 1, this group exhibits forecast skill over the ECMWF IFS ensemble for the first two Pangu6 inferences. The CRPSS reaches values of 0.1 to 0.15 at a lead time of 6 h. Subsequently, the CRPSS undergoes a rapid decline, dropping below -0.1. In the lead time interval between 24 and 96 h, the CRPSS reaches its minimum in the range of -0.15 to -0.1 for all forecast variables in this group. At lead times beyond 120 h, the CRPSS stabilizes at values around -0.1. There is a minor increase in CRPSS with increasing lead time. Similar CRPSS values are also observed for Z500 in this lead time range.

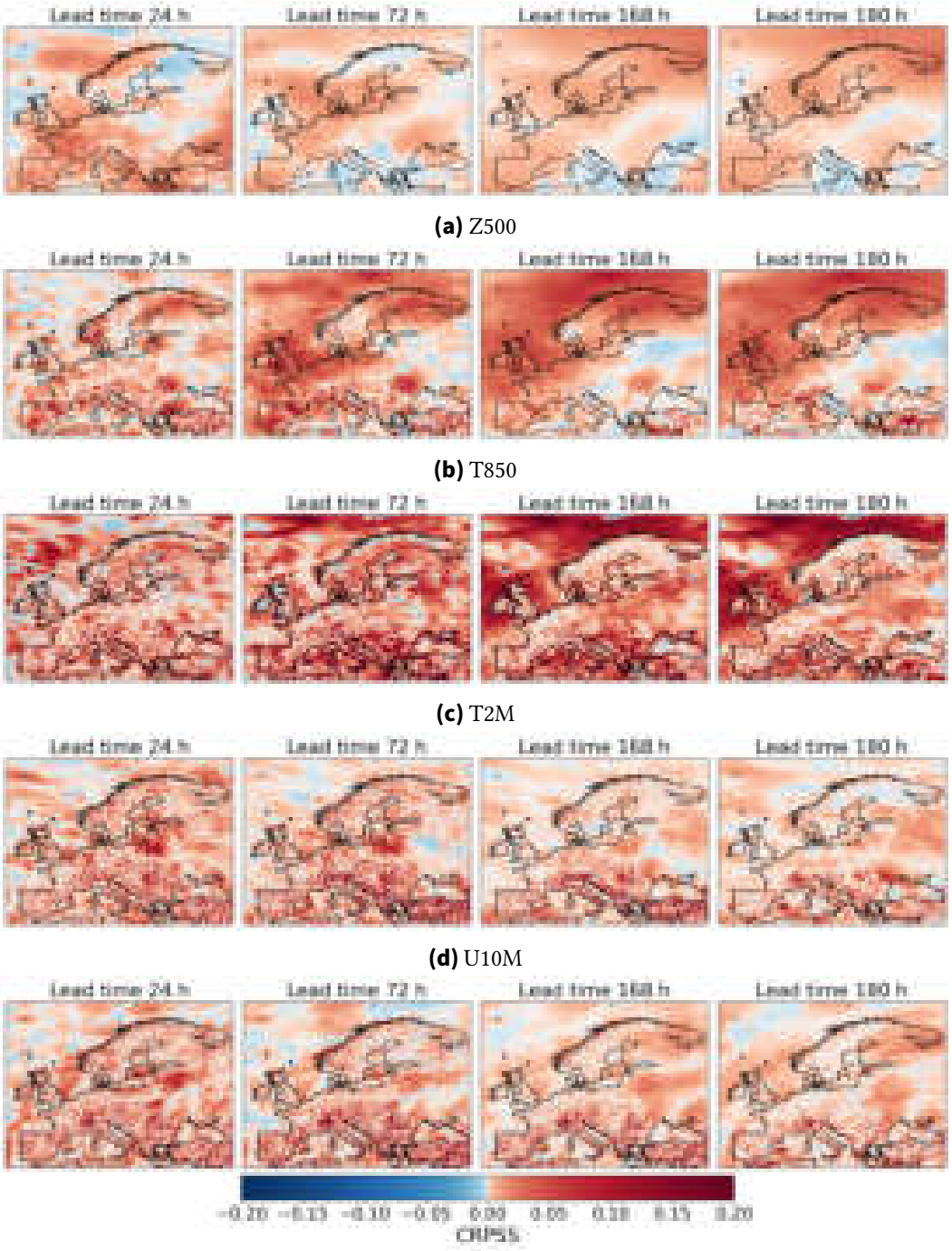
In contrast to the T2M ensemble, the third group does not show a diurnal CRPSS cycle. For the surface variables U10M and V10M, the CRPS typically reaches its diurnal maximum at 12 UTC, with lower values at 18 UTC. The ECMWF IFS ensemble shows a very similar diurnal CRPS cycle, visible in Figure 6.12. The IFS ensemble also has a midday CRPS maximum for the 10 m wind components. Both diurnal CRPS cycles superimpose. Consequently, the CRPSS does not have a significant diurnal cycle. Again, this shows that the diurnal CRPS cycle of the surface wind components is influenced by similar general dynamics, independent of the model architecture (MLWP vs. NWP model). For T850, there is a 12 h period CRPS cycle in the first days of the forecast, with respective maximums at 00 and 12 UTC and respective minimums at 06 and 18 UTC.

6.4.3 CRPSS maps

The previous section focuses on the domain-averaged CRPSS time series of the Pangu-Weather ensemble forecast over the ECMWF IFS ensemble forecast. Bias correction during post-processing shows the potential to improve the forecast skill of the A-SV-based Pangu-Weather ensemble. The analysis of the domain-averaged CRPSS time series gives no indication of the regional dependence of the CRPSS of the Pangu-Weather ensemble concerning both the bias correction and the IFS ensemble. The following section first discusses the CRPSS of the post-processed ensemble to highlight regional improvements in the forecast skill due to the bias correction. Then, we analyze the CRPSS of the Pangu-Weather ensemble forecast over the IFS ensemble forecast.

For Z500, Figure 6.13(a) shows positive CRPSS values across the domain without a distinct pattern developing at any lead time. For T850, there are strong positive CRPSS signals in the most mountainous regions of the domain. These include the Pyrenees, the Alps, and

[1]The largest ratio of spread/RMSE occurs with the first Pangu6 inference.



(a) Z500

(b) T850

(c) T2M

(d) U10M

(e) V10M

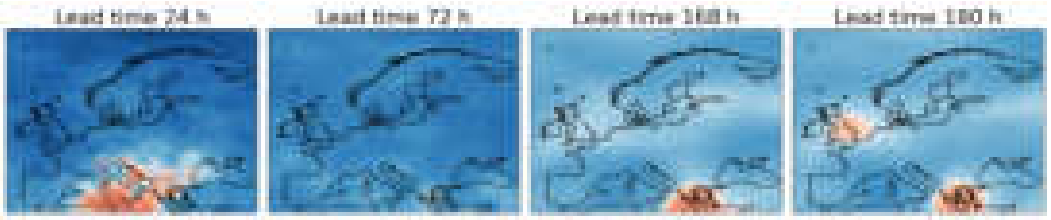
Figure 6.13: CRPSS of the post-processed Pangu-weather ensemble forecast over the raw Pangu-Weather ensemble forecast averaged across the year 2022 for the forecast variables Z500, T850, T2M, U10M, and V10M. Positive values (red) indicate forecast skill of the post-processed ensemble, negative values (blue) indicate worse forecast skill.

Central Turkey. Here, CRPSS values reach 0.2. This indicates that ensemble-mean biases in mountainous areas are an important contributor to the CRPS. The signals are independent of the lead time, though they tend to grow smaller with increasing lead time. Due to the internal architecture of Pangu-Weather, which includes patch embedding and recovery, the resulting compression may deteriorate forecasts at grid points in highly mountainous terrain. Aside from the signals over mountainous areas, higher CRPSS values are found more often over the ocean than over land. From a lead time of 72 h, a strong positive signal is present over the Norwegian Sea and the adjacent Arctic Ocean. In contrast, a broad area of negative CRPSS in the post-processed ensemble develops over Eastern Europe. Figure 6.9(b) shows a strong negative bias for T850 over the Norwegian Sea where the bias correction improves the forecast skill of the ensemble. The correction of the weaker positive bias over Eastern Europe appears to deteriorate the ensemble forecast in this part of the domain.

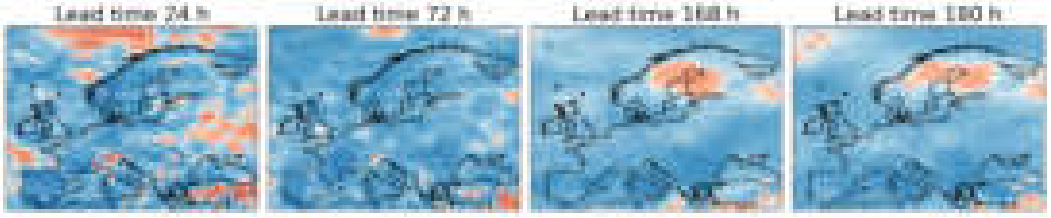
For U10M and V10M, the largest CRPSS signals are concentrated in the alpine regions of the domain. Figure 6.13 shows that these signals are often specific to only a single grid cell, i. e., the bias correction accounts for a strong local bias of Pangu-Weather. More coherent patterns of positive CRPSS are located in Eastern Europe at later lead times. As expected from the previous analysis, the most positive CRPSS patterns of the post-processed ensemble are apparent for T2M in Figure 6.13(c). Strong signals are located in mountainous areas, especially along the coastal mountain ranges in the Mediterranean. At later lead times, positive CRPSS values over land tend to decrease. In contrast, strong positive CRPSS signals develop around the British Isles and the Norwegian Sea. The most positive CRPSS signals of the post-processed Pangu-Weather ensemble are located in an area between Jan Mayen in the west and Bear Island in the east.

In these high latitudes, a potential explanation for these patterns is a bias in sea ice extent. However, even in March, when the Arctic sea ice extent is at its annual maximum, this part of the Norwegian Sea is, on average, south of the sea ice edge (Belchansky et al., 2004). The T2M bias (Figure 6.9(c)) shows a broad area of negative bias across the Norwegian Sea, which implies that the ensemble mean is too cold. The impact of a positive bias in sea ice extent in Pangu-Weather could lead to a negative T2M bias, since this sector of the Norwegian Sea saw below-average sea ice extent in March 2022 (Meier, 2022). More peculiar is the fact that the bias correction leads to a local deterioration of the T2M ensemble north of the Shetland Islands, where there is a local positive T2M bias at later lead times (see Figure 6.9(c)). Similarly to the bias correction of the T850 ensemble over Eastern Europe, the correction of the ensemble mean bias here appears to deteriorate the ensemble forecast skill on a local scale.

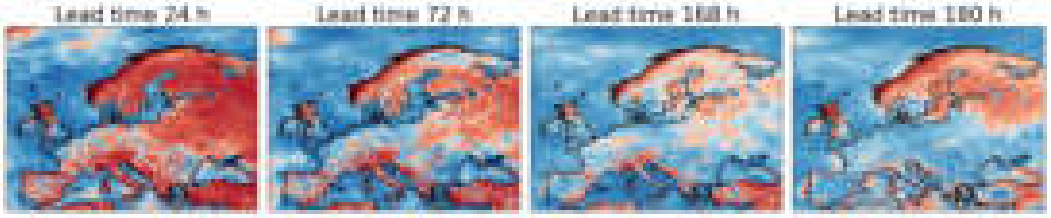
The previous paragraphs provide an evaluation of the regional impact of the ensemble-mean bias on the CRPS. Similarly, the regional patterns of forecast skill for Pangu-Weather are analyzed in the following paragraphs. The Z500 ensemble has some regional skill over the ECMWF IFS ensemble, as seen in Figure 6.14(a). Note that the color bar for Figure 6.14 is scaled from -1 to 1, in contrast to Figure 6.13. At early lead times, positive CRPSS values are observed across the Central Mediterranean. At later lead times, there are two minor positive CRPSS areas located over Greece and England. The analysis of Bülte et al. (2025)



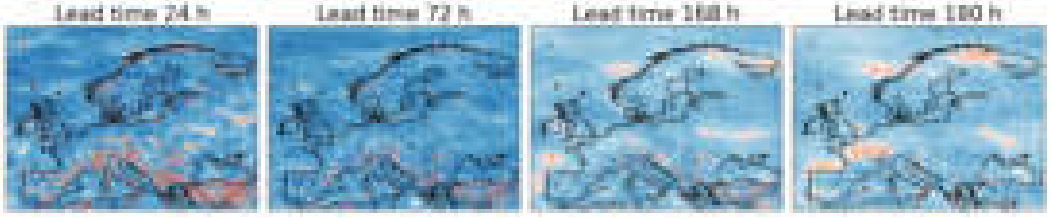
(a) Z500



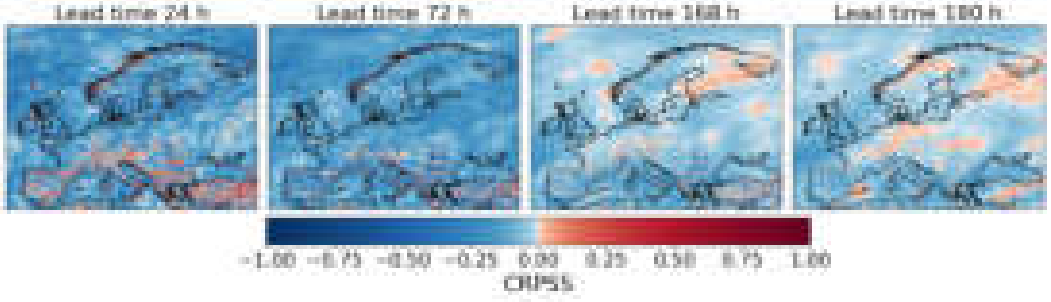
(b) T850



(c) T2M



(d) U10M



(e) V10M

Figure 6.14: CRPSS of the A-SV-based Pangu-weather ensemble forecast over the ECMWF IFS ensemble forecast averaged across the year 2022 for the forecast variables Z500, T850, T2M, U10M, and V10M. Positive values (red) indicate forecast skill of the A-SV-based Pangu-Weather ensemble, negative values (blue) indicate worse forecast skill of the A-SV-based ensemble compared to the IFS ensemble.

already indicates that Pangu-Weather-based ensembles tend to exhibit forecast skill for the Z500 ensemble across the Central Mediterranean at early lead times in the year 2022.

Both the surface wind components, U10M and V10M, share similar CRPSS patterns. Positive CRPSS values are concentrated mainly in the mountainous areas of the Pyrenees, the Alps, the Balkans, and Central Turkey. These disappear at later lead times. Bülte et al. (2025) already denote that the down-sampling of the IFS output from its native, higher-resolution grid to the coarser 0.25° equiangular grid of ERA5 and Pangu-Weather puts the IFS forecasts at an inherent disadvantage. This effect is most prominent in complex model topography. Hence, these strong positive CRPSS values of Pangu-Weather may only be an artifact of the down-sampling process of the IFS forecast. Over Continental Europe, some positive CRPSS values are found at later lead times. For U10M, positive CRPSS is located closer to the Atlantic Ocean than positive CRPSS of the V10M ensemble. Over the ocean, very few grid points exhibit positive CRPSS values.

At a lead time of 24 h, scattered areas of positive CRPSS are found for the T850 ensemble across the domain. These patches are located in parts of the Bay of Biscay, the Norwegian Sea between Jan Mayen and Bear Island, parts of Western Russia, and Central Turkey. From these initial areas, positive CRPSS values shift to Northern Scandinavia at the end of the forecast horizon. In the rest of the domain, negative CRPSS values dominate for the T850 ensemble. Figure 6.14(c) shows that for T2M, the extent of the positive CRPSS values is much larger compared to the other forecast variables at all lead times displayed. For every lead time, maritime grid cells have a lower average CRPSS. Thus, the forecast skill of Pangu-Weather is relatively worse over the ocean than over land, compared to the ECMWF IFS ensemble. Despite the complex surface parameterizations used in NWP models (e. g., Viterbo and Beljaars, 1995), the ML-based Pangu-Weather ensemble shows ample forecast skill over land. At early lead times, large coherent areas of positive CRPSS are found across Scandinavia, Western Russia, and Turkey. At a lead time of 180 h, positive CRPSS values decrease significantly in extent and magnitude. Positive CRPSS values are located in Scotland, Northern Scandinavia, and parts of Central Turkey.

Grid points with positive CRPSS values at later lead times are primarily located in parts of the European domain that have the longest average snow cover. Despite the vast amount of research on surface parameterizations, Pangu-Weather may better capture the dynamic processes involved in the persistence, formation, and melting of snow cover and their implications for the T2M forecast. Over the ocean, Pangu-Weather has significantly worse forecast skill than the IFS ensemble. Previous studies have shown that MLWP models tend to perform better in data-rich areas, such as Continental Europe (Baño-Medina et al., 2025). Data-rich areas in this context relate to the amount of observational data assimilated into ERA5 (Baño-Medina et al., 2025). Baño-Medina et al. (2025) show that the IFS tends to perform better in data-sparse areas such as the ocean. The T2M ensemble of Pangu-Weather exhibits very similar patterns. In addition, the previous analysis shows that the A-SV-based Pangu-Weather ensemble tends to be over-dispersive over the ocean (see Section 6.2.2). On a local scale, the forecast skill of Pangu-Weather is considerably worse over the Vuoksi lake system north-east of Saint Petersburg. Across these lakes, misrepresentation of lake ice

concentrations can significantly alter the T2M forecast and, thus, deteriorate the forecast skill of the ensemble on a local scale in this part of the domain.

Overall, the analysis shows that the A-SV-based Pangu-Weather ensemble performs best compared to the ECMWF IFS ensemble on day 1. These forecasts are performed solely by Pangu6 (Bi et al., 2023). At intermediate lead times between 24 and 96 h, there is a sharp drop in the forecast skill of Pangu-Weather. Beyond day 5, the CRPSS of Pangu-Weather stabilizes. For some forecast variables, the CRPSS improves slightly. Only the T2M ensemble has consistent skill over the IFS ensemble beyond day 1, with positive CRPSS values up to day 5. Positive CRPSS values are mostly restricted to land areas. Poorer forecast skill is derived over the ocean. Correction of the ensemble-mean bias provides significant forecast skill gains in terms of CRPSS for the T2M ensemble. For all other forecast variables tested, the post-processing provides only minor improvements (CRPSS $< 2.5\,\%$). Z500 is the only forecast variable for which the Pangu-Weather ensemble has no forecast skill over the IFS at any lead time. For the remaining forecast variables tested, the A-SV-based Pangu-Weather ensemble is competitive, but the CRPSS is persistently negative, with average values ranging from -0.1 to -0.05.

6.5 Comparison to different Pangu-Weather ensembles

The above analysis shows that the A-SV-based Pangu-Weather ensemble is competitive with the ECMWF IFS ensemble. Bülte et al. (2025) provide a selection of Pangu-Weather ensembles that were evaluated for the same European domain in the year 2022. These Pangu-Weather ensembles include additional initial condition ensembles, such as the A-SV-based ensemble. For these ensembles, the initial conditions of the model are perturbed and then multiple forecast trajectories are computed (Bülte et al., 2025). Additionally, Bülte et al. (2025) provide a selection of post-hoc approaches to compute an ensemble with Pangu-Weather. In these approaches, the Pangu-Weather forecast is taken, and then a pretrained statistical model is applied to the deterministic model output. The statistical model is a post-processing method that generates an ensemble of forecast states from the deterministic model output (Vannitsem et al., 2021). Bülte et al. (2025) also use the IFS ensemble forecast to evaluate the forecast skill of their Pangu-Weather ensembles. Before further analysis of the forecast skill of the A-SV-based Pangu-Weather ensemble, it should be noted that Bülte et al. (2025) have not used the domain-weighted mean of the CRPS in their analysis. Equally weighting all grid points in the mean creates unwanted emphasis on high-latitude grid points due to the contraction of the Pangu-Weather grid towards the poles. In order to ensure a fair comparison between the results of this thesis and Bülte et al. (2025), the subsequent analysis deliberately incorporates this error in the calculation of the averaged CRPS values for the A-SV-based Pangu-Weather ensemble.

Figure 6.15 shows the respective CRPS values for the five forecast variables analyzed in the previously defined lead time bins. The CRPSS values are computed using the respective ensemble forecasts in Figure 6.15 as the reference forecast $CRPSS_{\text{ref}}$ and the A-SV-based

Figure 6.15: Heat map of the CRPS values of the ECMWF IFS ensemble forecast (IFS), the A-SV-based Pangu-Weather ensemble forecast (A-SV), and the other Pangu-Weather-based ensemble forecasts (GNP, IFSP, RFP, UQ, DRN) of Bülte et al. (2025) in the predefined lead time bins for a subset of forecast variables (T850, Z500, T2M, U10M, and V10M) averaged across the year 2022. The first three ensembles (GNP, IFSP, RFP) of Bülte et al. (2025) are initial condition ensembles, the latter two (UQ and DRN) are post-hoc ensembles. The color coding gives the CRPSS values of the A-SV-based Pangu-Weather ensemble over the respective ensemble forecasts. Green boxes indicate positive CRPSS values and thus forecast skill of the A-SV-based Pangu-Weather ensemble. Magenta boxes highlight negative CRPSS values and thus worse forecast skill of the A-SV-based Pangu-Weather ensemble.

Pangu-Weather ensemble is always $CRPS_f$ in Equation 3.5. It is vital to compare the A-SV-based Pangu-Weather ensemble against the Gaussian Noise Perturbations (GNPs). The mathematical concept of SVs aims to identify the fastest growing perturbations in a system (Diaconescu and Laprise, 2012). By construction, the A-SV-based ensemble should have an advantage over GNPs because a A-SV-perturbed ensemble should have larger dispersion than the GNP ensemble (Winkler and Denhard, 2025; Bülte et al., 2025). The negative CRPSS values of GNPs imply that for every lead time bin and every forecast variable, the GNP ensemble performs worse than the A-SV-based ensemble. This shows that the Arnoldi algorithm indeed identifies rapidly growing perturbations in the initial conditions. In addition, this provides justification for the investment of extra computational resources to derive optimal perturbations of the initial conditions. In contrast, GNPs randomly perturb the initial conditions with white noise. Furthermore, it is remarkable that the CRPSS of the

A-SV-based ensemble over the GNP ensemble increases with lead time. This implies that knowledge of optimal perturbations in the initial conditions increases in importance with lead time. The ensemble that uses IFS Initial Conditions (IFSP) to perturb Pangu-Weather has even worse forecast skill. The poor forecast skill of the IFSP ensemble is discussed by Bülte et al. (2025).

The comparison with post-hoc approaches for ensemble generation provides a more comprehensive overview of the forecast skill of the A-SV-based Pangu-Weather ensemble. The Distributional Regression Network (DRN) is the overall best performing ensemble for the early lead time bin, as shown in Figure 6.15. In this lead time bin, the ensemble has forecast skill over the ECMWF IFS ensemble. Both post-hoc approaches provide forecast skill over the A-SV-based ensemble for all forecast variables in the early and intermediate lead time bins. For longer lead times, the A-SV-based ensemble outperforms both post-hoc approaches for every forecast variable tested. This suggests that knowledge of optimal perturbations of the initial conditions becomes increasingly important for longer lead times. At early lead times, the implication is that a better performing Pangu-Weather ensemble does not require knowledge of the optimal perturbations in the initial conditions. A simple statistical model is sufficient to provide a skillful ensemble forecast in these lead time bins.

To this point in the analysis, the comparison excludes the third and final initial condition ensemble provided by Bülte et al. (2025). The Random Field Perturbations (RFPs), proposed by Magnusson et al. (2009), use randomly selected, independent past atmospheric states that are scaled and used as initial condition perturbations. Details on the exact set-up of RFPs are provided in Bülte et al. (2025). For most forecast variables, the RFP ensemble has a lower CRPS than the statistical post-hoc approaches at intermediate lead times. At long lead times, RFPs are the Pangu-Weather ensemble with the lowest CRPS for all forecast variables (see Figure 6.15). In comparison, the A-SV-based ensemble has, in most instances, a higher CRPS. Thus, the A-SV-based ensemble has lower forecast skill than the RFP ensemble. Only in the early lead time bin is the CRPSS of the A-SV-based T2M ensemble positive. For all other forecast variables and lead time bins, the A-SV-based ensemble remains competitive. Yet, the CRPSS is always slightly negative ($\text{CRPSS} \in [-0.05, 0)$). Thus, the forecast skill of the A-SV-based ensemble is not significantly worse than that of the RFP ensemble.

However, these results give rise to a peculiar question in relation to the Pangu-Weather ensembles. A-SVs identify optimal perturbations of the initial conditions. RFPs provide knowledge about flow-dependent perturbations, including information on the underlying dynamics of the atmosphere (Magnusson et al., 2009). The above analysis suggests that, for the ML-based Pangu-Weather model, the knowledge of atmospheric dynamics based on past atmospheric states is of greater importance for the forecast skill of the ensemble. The small CRPS difference between both methods shows that further research in the set-up of the A-SV-based Pangu-Weather ensemble may provide sufficient improvements to gain forecast skill over RFPs. Additionally, extending the analysis to lead times beyond 186 h may grant the A-SV-based ensemble forecast skill over RFPs. An increasing CRPSS trend of the A-SV-based ensemble is observed for statistical post-hoc approaches. Similarly, the CRPSS of the A-SV-based ensemble over the RFP ensemble may also increase at medium-range

lead times. This would imply that knowledge of the optimal initial condition perturbations may increase in importance at lead times beyond 186 h.

The analysis shows that A-SVs provide initial condition perturbations that can be used to generate a skillful ensemble forecast with Pangu-Weather. The ensemble is generally well calibrated when choosing an appropriate a posteriori amplitude scaling factor γ. The forecast variables show a flow-dependent spread. The ensemble-mean bias is of a similar magnitude to that found in previous studies of Pangu-Weather. In addition, there is cross-variable consistency among the biases of different forecast variables. Accounting for the ensemble-mean bias shows that even simple post-processing approaches can significantly enhance the forecast skill of the A-SV-based Pangu-Weather ensemble. Comparison with the ECMWF IFS ensemble highlights that T2M is the forecast variable that performs best among the five tested. For the T2M ensemble, positive CRPSS persists at intermediate lead times. Positive CRPSS values are usually located over land. The other forecast variables tested show forecast skill on day 1, after which the CRPSS drops to negative values. At later lead times, the CRPSS generally stabilizes around -0.1. In addition, the A-SV-based ensemble is competitive with other Pangu-Weather-based ensemble approaches presented in Bülte et al. (2025). The A-SV-based ensemble has a lower CRPS than the GNP ensemble, which justifies the investment of additional computational resources for the A-SV computation. At later lead times, the A-SV-based ensemble demonstrates forecast skill over statistical approaches for ensemble generation. However, the A-SV-based ensemble has no forecast skill over RFPs. This shows that knowledge of the general dynamics inscribed in RFPs is of great importance for ensemble generation in Pangu-Weather.

7 Conclusion and outlook

MLWP models have revolutionized the field of NWP in the last few years. Advances in artificial intelligence have led to a vast increase in the forecast skill of MLWP models. Today, the forecast skill of MLWP models is on par with state-of-the-art NWP models. Once trained, a key advantage of MLWP models is their minimal computational cost for forecast inference, which is orders of magnitude smaller compared to NWP models. This opens the opportunity to make use of the adjoint-free Arnoldi method for SV calculation. This thesis evaluates the validity and potential value of A-SVs in Pangu-Weather.

First, the general structure of A-SV perturbations in Pangu-Weather is analyzed, focusing on the known patterns of L-SV perturbations in NWP models. Different combinations of Pangu-Weather forecast models and OTIs are used to outline potential dependencies of A-SVs on these parameters. A case study for Cyclone Xynthia is performed to compare A-SV-based sensitivities with the known sensitivity patterns of Cyclone Xynthia in NWP and MLWP models. In a final step, an extensive analysis of the forecast skill of the proposed A-SV-based Pangu-Weather ensemble is presented for a selected number of forecast variables across the European domain. The forecast skill is evaluated against the benchmark ECMWF IFS ensemble and other Pangu-Weather-based ensemble approaches.

In the introduction, three main research questions are formulated. In the following paragraphs, these research questions are answered with a concise summary of the results of this study.

1. **Do A-SVs in Pangu-Weather have similar structures to L-SVs in physics-based NWP models in terms of their vertical profiles, energy partitions, and horizontal scales at initial and final time?**

   The analysis shows that A-SV perturbations in Pangu-Weather have structures similar to L-SV perturbations in physics-based NWP models. At initial time, A-SV perturbations in Pangu-Weather are tilted westward with height. These shift to an eastward tilt with height at final time. This shift is more pronounced than in physics-based NWP models. Inter-variable comparisons have shown that the A-SV perturbations in Pangu-Weather are physically coherent. In contrast to L-SVs in NWP models, A-SVs in Pangu-Weather exhibit no upward energy propagation at final time because the initial time perturbation maximum is already located in the upper troposphere. A-SVs in Pangu-Weather show an increase of FrDKE from initial to final time. Compared to L-SVs in NWP models, this increase is smaller due to higher FrDKE values at initial and lower FrDKE values at final time. The latitudinal dependencies of the DTE and FrDKE are consistent with the general dynamics of the atmosphere. The horizontal scales of A-SV perturbations exhibit an upscale energy growth that is consistent with L-SVs in NWP models. The

upscale energy growth is more pronounced in Pangu-Weather as A-SV perturbations have spectral maxima at higher wavenumbers. However, transcending white noise from the Arnoldi method initialization contaminates the spectral analysis revealed by a different/non-existent decay of the spectral density at horizontal wavelengths below 100 km.

2. **Can A-SVs in Pangu-Weather identify sensitivity patterns similar to those previously analyzed for extra-tropical cyclones in the example of Cyclone Xynthia (2010)?**

   The case study of Cyclone Xynthia highlights that A-SVs in Pangu-Weather show physically coherent atmospheric sensitivities that are consistent with previous studies, regardless of the fact that the local projection operator **P** cannot be limited exclusively to the domain of Cyclone Xynthia. In addition, the use of the dry total energy norm prevents the study of the initial time moisture sensitivity field, despite the unusually large contribution of diabatic processes during the intensification of Cyclone Xynthia. The temperature sensitivity pattern shows a dipole structure upstream of Cyclone Xynthia. The vertical cross-section indicates sensitivity to the vertical temperature gradients in the troposphere. Similarly to the SFNO model, A-SVs in Pangu-Weather also hint at spurious stratospheric sensitivities of Cyclone Xynthia. The evolved A-SVs in Pangu-Weather lead to an intensification/weakening of Cyclone Xynthia of a magnitude akin to previous studies when the initial conditions are positively/negatively perturbed by the leading A-SV. The case study highlights not only the physical coherence of A-SVs in Pangu-Weather but also shows that the MLWP and NWP models derive similar atmospheric sensitivities for Cyclone Xynthia despite the completely different forecast inference methods. In addition, this case study further validates the hypothesis that Cyclone Xynthia could have been more intense in the presence of better atmospheric conditions surrounding the cyclone.

3. **Can A-SVs be used to generate a reliable ensemble forecast with Pangu-Weather that mimics the L-SV-based approach of the ECMWF IFS ensemble forecast? Does the A-SV-based ensemble have forecast skill over the ECMWF IFS ensemble or other Pangu-Weather-based ensemble approaches?**

   For the selected subset of forecast variables across the European domain, the A-SV-based Pangu-Weather ensemble is generally well calibrated. However, the choice of the correct a posteriori amplitude scaling factor γ for the ensemble perturbations is challenging. All variables exhibit a tendency to become under-dispersive with increasing lead time. Despite the significant over-dispersion at early lead times, even the Z500 ensemble is under-dispersive at later lead times. Detailed analysis shows that the A-SV-based initial condition perturbations depend on the general synoptic set-up over Europe and exhibit a significant seasonal cycle that matches the expectations of the general atmospheric dynamics. For the T2M ensemble, a characteristic Pangu-Weather deficiency is observed. Discontinuities between the different Pangu-Weather forecast models are present in the ensemble forecast. Pangu-Weather is known to have significant biases at later lead times. Very similar biases also develop in the ensemble mean. However, these biases exhibit cross-variable consistency across the domain during the tested time period.

These biases provide an opportunity for a simplistic post-processing approach. In the proposed post-processing approach, the ensemble mean bias of the reference period is corrected for. At early lead times, the post-processed ensemble sees a significant reduction in CRPS for the T2M ensemble. For all other forecast variables, the CRPS decreases only marginally, independent of the lead time. The CRPSS of the A-SV-based Pangu-Weather ensemble forecast over the ECMWF IFS ensemble forecast shows that for T2M, the ensemble has forecast skill over the IFS in the early and intermediate lead time bin. In contrast, the highly over-dispersive Z500 ensemble is by far the worst performing forecast variable with a CRPSS of -0.3 at early lead times. On day 1, all other forecast variables tested have skill over the IFS. This is followed by a sharp drop in CRPSS. In the long lead time bin, the CRPSS increases to values between -0.1 and -0.05, indicating a somewhat competitive ensemble forecast. Positive CRPSS values of Pangu-Weather are typically located in highly mountainous parts of the domain, which likely result from the re-gridding of the IFS forecast. For T2M, positive CRPSS values focus on land areas with negative CRPSS values over the ocean. At later lead times, positive CRPSS values are mostly confined to Northern Scandinavia and Western Russia. It is remarkable that a MLWP model has forecast skill over the state-of-the-art NWP model for a forecast variable, where vast amounts of research have been dedicated to improving surface parameterizations.

Comparison with other Pangu-Weather-based ensemble approaches shows that the A-SV-based ensemble is highly competitive. Crucially, the A-SV-based ensemble has a significantly lower CRPS than the GNP ensemble. This justifies the investment of additional computational resources spent on the calculation of optimal perturbations in the initial conditions, instead of randomly perturbing them. At early lead times, statistical ensemble forecasts have forecast skill over the A-SV-based ensemble. With increasing lead time, the forecast skill shifts towards the A-SV-based ensemble. This highlights that the knowledge of optimal perturbations in the initial conditions increases in importance for longer lead times. Remarkably, RFPs have a forecast skill over the A-SV-based ensemble. Thus, for the forecast variables and lead times tested, knowledge of the general atmospheric dynamics, imposed in the model by RFPs, is of greater importance to a skillful ensemble forecast with Pangu-Weather than knowledge of optimal perturbations of the initial conditions.

This work shows that A-SVs impose physically consistent perturbations in Pangu-Weather that share general characteristics of L-SVs in physics-based NWP models. The A-SV-based sensitivity patterns for Cyclone Xynthia are in agreement with previous studies. A-SVs also provide the starting ground for a highly competitive ensemble forecast with Pangu-Weather. This work should be considered a proof-of-concept study. Most certainly, the forecast skill of the A-SV-based Pangu-Weather ensemble can be improved by more elaborate studies of the parameter choices of the Arnoldi method for SV calculation. The following paragraphs outline potential starting points for future research on this topic.

Winkler and Denhard (2025) use a larger $\mathbf{H}$ matrix in their study. Expanding the $\mathbf{H}$ matrix grants a larger sample of the initial condition perturbations. A larger number of computed perturbation vectors v also reduces the risk of utilizing non-growing modes of the $\mathbf{H}$ matrix

due to the limited sample of perturbation vectors. In addition, a larger **H** matrix may reduce the impact of transcending white noise from the initialization of the Arnoldi method. This work also omits any study of the perturbation amplitude h. SV calculation with the Arnoldi method can also be initialized from the ensemble initial conditions of ICON. This could further reduce the contamination of white noise. The spread-RMSE relationship of the forecast variables shows a seasonal cycle. As such, a seasonal cycle of the a posteriori amplitude scaling factor γ could also improve the forecast skill of the ensemble. Implicitly, the ECMWF IFS also scales the L-SV-based ensemble perturbations with a scaling factor that has a minor seasonal dependency (Leutbecher and Palmer, 2008, Equation 25). With the knowledge from this study, a valid first guess for the seasonal dependence of γ would be a sinusoidal function with values ranging from around 0.15 in winter to 0.3 in summer. Furthermore, an argument can be made for variable-specific scaling factors.

The initialization scheme of the A-SV-based Pangu-Weather ensemble can be increased in complexity to more closely mimic the ensemble initialization scheme of the ECMWF IFS. Separate sets of the leading A-SVs per hemisphere could be used. For these, perturbation growth is localized to the respective hemispheres. In addition, ensemble initialization from pure A-SV perturbations only accounts for the limited knowledge of the initial conditions. NWP models induce a second type of error in their predictions of future atmospheric states. NWP models account for the impact of the model error by, e. g., perturbing parameterizations (e. g., Buizza et al., 1999; Plant and Craig, 2008). For a MLWP model like Pangu-Weather, this is not an option. However, the analysis shows that A-SV perturbations depend on the Pangu-Weather forecast models and OTIs. Hence, a potential routine that accounts for model errors could include A-SVs derived from different combinations of Pangu-Weather forecast models and OTIs.

Without the constraint of the linearization assumption for traditional L-SV calculation, the OTI of A-SVs can be extended to longer lead times. In the medium-range, the ensemble could potentially perform better for a medium-range OTI. In addition, the analysis above only considers a simplistic post-processing approach by accounting for the ensemble mean bias. More sophisticated post-processing approaches are likely to further increase the forecast skill of the ensemble. Additionally, the analysis above is limited to a selected number of forecast variables, a restricted domain, and a rather short forecast horizon. Future work can easily extend the analysis to all available forecast variables of Pangu-Weather, the global domain, and a forecast horizon spanning into the medium-range or potentially even the sub-seasonal range. This is intriguing since the CRPSS of the A-SV-based Pangu-Weather ensemble increases for most forecast variables at longer lead times. Future work could examine whether the CRPSS increases further in the medium-range.

The inexpensive cost of repeated forecast inferences with Pangu-Weather allows for broad sampling of forecast uncertainty. Hence, very large ensemble sizes are a realistic possibility with Pangu-Weather. With a very large ensemble, the forecast skill of the A-SV-based ensemble approach could be evaluated during extreme events. In addition, the analysis above shows that the A-SV-based ensemble exhibits periods of large under- or over-dispersion. A case study during these events could deepen the understanding of the interactions between A-SVs and Pangu-Weather. Furthermore, the Arnoldi method for SV calculation could be

applied to different MLWP models, or the MLWP-based A-SVs could be transferred to NWP models for ensemble initialization.

The above paragraphs provide a brief outline of the opportunities for future research on the topic of A-SVs in MLWP models. This study shows their potential value for ensemble generation with MLWP models. As mentioned, this work should be seen as a proof-of-concept study that clearly outlines that A-SVs can be used to generate a skillful ensemble forecast with MLWP models such as Pangu-Weather. The rapid progress of MLWP models in recent years has opened and will open many opportunities for future research.

Abbreviations

4D-Var	Four-Dimensional Variational Analysis
A-SV	Arnoldi-SV
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
DKE	Difference Kinetic Energy
DPE	Difference Potential Energy
DRN	Distributional Regression Network
DTE	Difference Total Energy
DWD	Deutscher Wetterdienst
ECMWF	European Centre for Medium-Range Weather Forecasts
EDA	Ensemble Data Assimilation
EIM	Evolved Increment Matrix
ERA5	ECMWF Reanalysis v5
FrDKE	Fraction of Difference Kinetic Energy
GNP	Gaussian Noise Perturbation
GPU	Graphical Processing Unit
ICON	ICOsahedral Nonhydrostatic
IFS	Integrated Forecasting System
IFSP	IFS Initial Conditions
L-SV	Lanczos-SV
ML	Machine Learning
MLWP	Machine Learning based Weather Prediction
MSLP	Mean Sea Level Pressure
NWP	Numerical Weather Prediction
OTI	Optimization Time Interval
Pangu24	Pangu-Weather's 24 h forecast model
Pangu6	Pangu-Weather's 6 h forecast model
RFP	Random Field Perturbation
RMSE	Root Mean Squared Error
SFNO	Spherical Fourier Neural Operator
SV	Singular Vector
SVD	Singular Value Decomposition
T2M	2 m Temperature
T850	850 hPa Temperature
U10M	10 m zonal Wind
V10M	10 m meridional Wind
Z500	500 hPa Geopotential

Bibliography

Arnoldi, W. E., 1951: The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, **9 (1)**, 17–29, https://doi.org/10.1090/qam/42792, URL: https://www.ams.org/qam/1951-09-01/S0033-569X-1951-42792-9/.

Augier, P., and E. Lindborg, 2013: A New Formulation of the Spectral Energy Budget of the Atmosphere, with Application to Two High-Resolution General Circulation Models. *Journal of the Atmospheric Sciences*, **70 (7)**, 2293–2308, https://doi.org/10.1175/JAS-D-12-0281.1, URL: https://journals.ametsoc.org/doi/10.1175/JAS-D-12-0281.1.

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

Baumgart, M., P. Ghinassi, V. Wirth, T. Selz, G. C. Craig, and M. Riemer, 2019: Quantitative View on the Processes Governing the Upscale Error Growth up to the Planetary Scale Using a Stochastic Convection Scheme. *Monthly Weather Review*, **147 (5)**, 1713–1731, https://doi.org/10.1175/MWR-D-18-0292.1, URL: http://journals.ametsoc.org/doi/10.1175/MWR-D-18-0292.1.

Baño-Medina, J., A. Sengupta, J. D. Doyle, C. A. Reynolds, D. Watson-Parris, and L. D. Monache, 2025: Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. *npj Climate and Atmospheric Science*, **8 (1)**, 92, https://doi.org/10.1038/s41612-025-00949-6, URL: https://www.nature.com/articles/s41612-025-00949-6.

Baño-Medina, J., A. Sengupta, D. Watson-Parris, W. Hu, and L. Delle Monache, 2025: Toward Calibrated Ensembles of Neural Weather Model Forecasts. *Journal of Advances in Modeling Earth Systems*, **17 (4)**, e2024MS004 734, https://doi.org/10.1029/2024MS004734, URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004734.

Belchansky, G. I., D. C. Douglas, and N. G. Platonov, 2004: Duration of the Arctic Sea Ice Melt Season: Regional and Interannual Variability, 1979–2001. *Journal of Climate*, **17 (1)**, 67–80, https://doi.org/10.1175/1520-0442(2004)017<0067:DOTASI>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0442(2004)017<0067:DOTASI>2.0.CO;2.

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

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

Bissolli, P., 2010: Orkansturm Xynthia über Südwest- und Westeuropa. Deutscher Wetterdienst, URL: https://www.dwd.de/DE/leistungen/besondereereignisse/stuerme/20100315_xynthia.pdf?__blob=publicationFile&v=4.

Bonavita, M., 2024: On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters*, **51 (12)**, e2023GL107 377, https://doi.org/10.1029/2023GL107377, URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL107377.

Buizza, R., 1994: Sensitivity of optimal unstable structures. *Quarterly Journal of the Royal Meteorological Society*, **120 (516)**, 429–451, https://doi.org/10.1002/qj.49712051609, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49712051609.

Buizza, R., 1998: Impact of Horizontal Diffusion on T21, T42, and T63 Singular Vectors. *Journal of the Atmospheric Sciences*, **55 (6)**, 1069–1083, https://doi.org/10.1175/1520-0469(1998)055<1069:IOHDOT>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0469(1998)055<1069:IOHDOT>2.0.CO;2.

Buizza, R., R. Gelaro, F. Molteni, and T. N. Palmer, 1997: The impact of increased resolution on predictability studies with singular vectors. *Quarterly Journal of the Royal Meteorological Society*, **123 (540)**, 1007–1033, https://doi.org/10.1002/qj.49712354011, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49712354011.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **125 (560)**, 2887–2908, https://doi.org/10.1002/qj.49712556006, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49712556006.

Buizza, R., and T. N. Palmer, 1995: The Singular-Vector Structure of the Atmospheric Global Circulation. *Journal of Atmospheric Sciences*, **52 (9)**, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/atsc/52/9/1520-0469_1995_052_1434_tsvsot_2_0_co_2.xml, place: Boston MA, USA Publisher: American Meteorological Society.

Bülte, C., N. Horat, J. Quinting, and S. Lerch, 2025: Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*, https://doi.org/10.1175/AIES-D-24-0049.1, URL: https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-24-0049.1/AIES-D-24-0049.1.xml.

Cohn, S. E., 1997: An Introduction to Estimation Theory. *Journal of the Meteorological Society of Japan. Ser. II*, **75 (1B)**, 257–288, https://doi.org/10.2151/jmsj1965.75.1B_257.

Cooley, J. W., and J. W. Tukey, 1965: An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, **19 (90)**, 297–301, https://doi.org/10.2307/2003354, URL: http://www.jstor.org/stable/2003354, publisher: American Mathematical

Society.

Coutinho, M. M., B. J. Hoskins, and R. Buizza, 2004: The Influence of Physical Processes on Extratropical Singular Vectors. *Journal of the Atmospheric Sciences*, **61 (2)**, 195–209, https://doi.org/10.1175/1520-0469(2004)061<0195:TIOPPO>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0469(2004)061<0195:TIOPPO>2.0.CO;2.

Diaconescu, E. P., and R. Laprise, 2012: Singular vectors in atmospheric sciences: A review. *Earth-Science Reviews*, **113 (3-4)**, 161–175, https://doi.org/10.1016/j.earscirev.2012.05.005, URL: https://linkinghub.elsevier.com/retrieve/pii/S0012825212000657.

Diaconescu, E. P., R. Laprise, and A. Zadra, 2012: Singular vector decomposition of the internal variability of the Canadian Regional Climate Model. *Climate Dynamics*, **38 (5-6)**, 1093–1113, https://doi.org/10.1007/s00382-011-1179-x, URL: http://link.springer.com/10.1007/s00382-011-1179-x.

Dosovitskiy, A., and Coauthors, 2021: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv, URL: http://arxiv.org/abs/2010.11929, arXiv:2010.11929 [cs], https://doi.org/10.48550/arXiv.2010.11929.

Errico, R. M., 1997: What Is an Adjoint Model? *Bulletin of the American Meteorological Society*, **78 (11)**, 2577–2591, https://doi.org/10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2.

Fink, A. H., S. Pohle, J. G. Pinto, and P. Knippertz, 2012: Diagnosing the influence of diabatic processes on the explosive deepening of extratropical cyclones. *Geophysical Research Letters*, **39 (7)**, 2012GL051 025, https://doi.org/10.1029/2012GL051025, URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2012GL051025.

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte, 2014: Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology*, **15 (4)**, 1708–1713, https://doi.org/10.1175/JHM-D-14-0008.1, URL: http://journals.ametsoc.org/doi/10.1175/JHM-D-14-0008.1.

Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear Regime Duration: Is 24 Hours a Long Time in Synoptic Weather Forecasting? *Journal of the Atmospheric Sciences*, **58 (22)**, 3525–3539, https://doi.org/10.1175/1520-0469(2001)058<3525:LRDIHA>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0469(2001)058<3525:LRDIHA>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2007: Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378, https://doi.org/10.1198/016214506000001437, URL: http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118, https://doi.org/10.1175/MWR2904.1, URL:

http://journals.ametsoc.org/doi/10.1175/MWR2904.1.

Hakim, G. J., and S. Masanam, 2024: Dynamical Tests of a Deep Learning Weather Prediction Model. *Artificial Intelligence for the Earth Systems*, **3 (3)**, e230 090, https://doi.org/10.1175/AIES-D-23-0090.1, URL: https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0090.1.xml.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585 (7825)**, 357–362, https://doi.org/10.1038/s41586-020-2649-2, URL: https://www.nature.com/articles/s41586-020-2649-2.

Hartmann, D. L., R. Buizza, and T. N. Palmer, 1995: Singular Vectors: The Effect of Spatial Scale on Linear Growth of Disturbances. *Journal of Atmospheric Sciences*, **52 (22)**, 3885–3894, https://doi.org/10.1175/1520-0469(1995)052<3885:SVTEOS>2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/atsc/52/22/1520-0469_1995_052_3885_svteos_2_0_co_2.xml, place: Boston MA, USA Publisher: American Meteorological Society.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, **15 (5)**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

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

Hoskins, B. J., R. Buizza, and J. Badger, 2000: The nature of singular vector growth and structure. *Quarterly Journal of the Royal Meteorological Society*, **126 (566)**, 1565–1580, https://doi.org/10.1002/qj.49712656601, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49712656601.

Hoskins, B. J., G. Yang, and R. M. Fonseca, 2020: The detailed dynamics of the June–August Hadley Cell. *Quarterly Journal of the Royal Meteorological Society*, **146 (727)**, 557–575, https://doi.org/10.1002/qj.3702, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3702.

Kalnay, E., 2002: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511802270, URL: https://www.cambridge.org/core/product/C5FD207439132836E85027754CE9BC1A.

Keisler, R., 2022: Forecasting Global Weather with Graph Neural Networks. arXiv, URL: http://arxiv.org/abs/2202.07575, arXiv:2202.07575 [physics], https://doi.org/10.48550/arXiv.2202.07575.

Kim, H. M., and B.-J. Jung, 2009a: Influence of Moist Physics and Norms on Singular Vectors for a Tropical Cyclone. *Monthly Weather Review*, **137 (2)**, 525–543, https://doi.org/10.1175/2008MWR2739.1, URL: http://journals.ametsoc.org/doi/10.1175/2008MWR2739.1.

Kim, H. M., and B.-J. Jung, 2009b: Singular Vector Structure and Evolution of a Recurving Tropical Cyclone. *Monthly Weather Review*, **137 (2)**, 505–524, https://doi.org/10.1175/2008MWR2643.1, URL: http://journals.ametsoc.org/doi/10.1175/2008MWR2643.1.

Kochkov, D., and Coauthors, 2024: Neural general circulation models for weather and climate. *Nature*, **632 (8027)**, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y, URL: https://www.nature.com/articles/s41586-024-07744-y.

Lam, R., and Coauthors, 2023: GraphCast: Learning skillful medium-range global weather forecasting. arXiv, URL: http://arxiv.org/abs/2212.12794, arXiv:2212.12794 [cs], https://doi.org/10.48550/arXiv.2212.12794.

Lanczos, C., 1950: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, **45 (4)**, 255, https://doi.org/10.6028/jres.045.026, URL: https://nvlpubs.nist.gov/nistpubs/jres/045/jresv45n4p255_A1b.pdf.

Lang, S., and Coauthors, 2024: AIFS – ECMWF's data-driven forecasting system. arXiv, URL: http://arxiv.org/abs/2406.01465, arXiv:2406.01465 [physics], https://doi.org/10.48550/arXiv.2406.01465.

Le Dimet, F.-X., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography*, **38 (2)**, 97, https://doi.org/10.3402/tellusa.v38i2.11706, URL: https://tellusjournal.org/articles/10.3402/tellusa.v38i2.11706.

Leutbecher, M., and T. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227 (7)**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, URL: https://linkinghub.elsevier.com/retrieve/pii/S0021999107000812.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, 2021: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv, URL: http://arxiv.org/abs/2103.14030, arXiv:2103.14030 [cs], https://doi.org/10.48550/arXiv.2103.14030.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17 (3)**, 321–333, https://doi.org/10.1111/j.2153-3490.1965.tb01424.x, URL: http://tellusa.net/index.php/tellusa/article/view/9076.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x, URL: http://tellusa.net/index.php/tellusa/article/view/10086.

Magnusson, L., J. Nycander, and E. Källén, 2009: Flow-dependent versus flow-independent initial perturbations for ensemble prediction. *Tellus A: Dynamic Meteorology and Oceanography*, **61 (2)**, 194, https://doi.org/10.1111/j.1600-0870.2008.00385.x, URL: https://a.tellusjournals.se/article/10.1111/j.1600-0870.2008.00385.x/.

Matheson, J. E., and R. L. Winkler, 1976: Scoring Rules for Continuous Probability Distributions. *Management Science*, **22 (10)**, 1087–1096, URL: http://www.jstor.org/stable/2629907, publisher: INFORMS.

Meier, W. N., 2022: Sea Ice. Tech. rep., Global Ocean Monitoring and Observing (U.S.). https://doi.org/10.25923/XYP2-VZ45, URL: https://repository.library.noaa.gov/view/noaa/48539.

Montani, A., and A. J. Thorpe, 2002: Mechanisms leading to singular-vector growth for FASTEX cyclones. *Quarterly Journal of the Royal Meteorological Society*, **128 (579)**, 131–148, https://doi.org/10.1256/00359000260498824, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1256/00359000260498824.

Palmer, T. N., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular Vectors, Metrics, and Adaptive Observations. *Journal of the Atmospheric Sciences*, **55 (4)**, 633–653, https://doi.org/10.1175/1520-0469(1998)055<0633:SVMAAO>2.0.CO;2, URL: http://journals.ametsoc.org/doi/10.1175/1520-0469(1998)055<0633:SVMAAO>2.0.CO;2.

Pathak, J., and Coauthors, 2022: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. arXiv, URL: http://arxiv.org/abs/2202.11214, arXiv:2202.11214 [physics], https://doi.org/10.48550/arXiv.2202.11214.

Peixoto, J. P., and A. H. Oort, 1992: *Physics of climate*. 1st ed., American Institude of Physics, Melville, NY.

Peng, M. S., and C. A. Reynolds, 2006: Sensitivity of Tropical Cyclone Forecasts as Revealed by Singular Vectors. *Journal of the Atmospheric Sciences*, **63 (10)**, 2508–2528, https://doi.org/10.1175/JAS3777.1, URL: https://journals.ametsoc.org/doi/10.1175/JAS3777.1.

Pereira, M. B., and L. Berre, 2006: The Use of an Ensemble Approach to Study the Background Error Covariances in a Global NWP Model. *Monthly Weather Review*, **134 (9)**, 2466–2489, https://doi.org/10.1175/MWR3189.1, URL: http://journals.ametsoc.org/doi/10.1175/MWR3189.1.

Plant, R. S., and G. C. Craig, 2008: A Stochastic Parameterization for Deep Convection Based on Equilibrium Statistics. *Journal of the Atmospheric Sciences*, **65 (1)**, 87–105, https://doi.org/10.1175/2007JAS2263.1, URL: https://journals.ametsoc.org/doi/10.1175/2007JAS2263.1.

Price, I., and Coauthors, 2024: Probabilistic weather forecasting with machine learning. *Nature*, https://doi.org/10.1038/s41586-024-08252-9, URL: https://www.nature.com/articles/s41586-024-08252-9.

Reynolds, C. A., M. S. Peng, and J.-H. Chen, 2009: Recurving Tropical Cyclones: Singular Vector Sensitivity and Downstream Impacts. *Monthly Weather Review*, **137 (4)**, 1320–1337, https://doi.org/10.1175/2008MWR2652.1, URL: http://journals.ametsoc.org/doi/10.1175/2008MWR2652.1.

Rodwell, M. J., M. C. A. Clare, S. Lock, K. Lonitz, and M. Chevallier, 2025: Power Spectra of Physics-Based and Data-Driven Ensembles. *Meteorological Applications*, **32 (5)**, e70 071, https://doi.org/10.1002/met.70071, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/met.70071.

Sanders, F., and J. R. Gyakum, 1980: Synoptic-Dynamic Climatology of the "Bomb". *Monthly Weather Review*, **108 (10)**, 1589–1606, https://doi.org/10.1175/1520-0493(1980)108<1589:SDCOT>2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/mwre/108/10/1520-0493_1980_108_1589_sdcot_2_0_co_2.xml, place: Boston MA, USA Publisher: American Meteorological Society.

Selz, T., and G. C. Craig, 2023: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters*, **50 (20)**, e2023GL105 747, https://doi.org/10.1029/2023GL105747, URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023GL105747.

Selz, T., M. Riemer, and G. C. Craig, 2022: The Transition from Practical to Intrinsic Predictability of Midlatitude Weather. *Journal of the Atmospheric Sciences*, **79 (8)**, 2013–2030, https://doi.org/10.1175/JAS-D-21-0271.1, URL: https://journals.ametsoc.org/view/journals/atsc/79/8/JAS-D-21-0271.1.xml.

Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov and optimal vectors in a low-resolution GCM. *Tellus A*, **49 (2)**, 200–227, https://doi.org/10.1034/j.1600-0870.1997.00004.x, URL: http://tellusa.net/index.php/tellusa/article/view/14467.

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

Viterbo, P., and A. C. M. Beljaars, 1995: An Improved Land Surface Parameterization Scheme in the ECMWF Model and Its Validation. *Journal of Climate*, **8 (11)**, 2716–2748, https://doi.org/10.1175/1520-0442(1995)008<2716:AILSPS>2.0.CO;2, URL: https://journals.ametsoc.org/view/journals/clim/8/11/1520-0442_1995_008_2716_ailsps_2_0_co_2.xml, place: Boston MA, USA Publisher: American Meteorological Society.

Winkler, J., and M. Denhard, 2025: Arnoldi Singular Vector perturbations for machine learning weather prediction. arXiv, URL: http://arxiv.org/abs/2506.22450, arXiv:2506.22450 [physics], https://doi.org/10.48550/arXiv.2506.22450.

Winkler, J., M. Denhard, and B. A. Schmitt, 2020: Krylov methods for adjoint-free singular vector based perturbations in dynamical systems. *Quarterly Journal of the Royal Meteorological Society*, **146 (726)**, 225–239, https://doi.org/10.1002/qj.3668, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3668.

Zadra, A., M. Buehner, S. Laroche, and J. Mahfouf, 2004: Impact of the GEM model simplified physics on extratropical singular vectors. *Quarterly Journal of the Royal*

Meteorological Society, **130 (602)**, 2541–2569, https://doi.org/10.1256/qj.03.208, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1256/qj.03.208.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale Predictability of Moist Baroclinic Waves: Convection-Permitting Experiments and Multistage Error Growth Dynamics. *Journal of the Atmospheric Sciences*, **64 (10)**, 3579–3594, https://doi.org/10.1175/JAS4028.1, URL: https://journals.ametsoc.org/doi/10.1175/JAS4028.1.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141 (687)**, 563–579, https://doi.org/10.1002/qj.2378, URL: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.2378.

Acknowledgment

First and foremost, I would like to thank my supervisor, Julian Quinting. It was a real privilege to write my master's thesis under your supervision. Your advice and guidance during the past year were truly special. Thank you for the opportunity to work on such an interesting topic for my master's thesis. The way in which you open science and give one the opportunity to grow as a student is remarkable. Thank you for sharing your knowledge and wisdom at a decisive point in my professional career. On top of that, you had opened doors early in my studies that I could have only dreamed about. I cannot tell you how grateful I am for these opportunities.

I also express my gratitude to Peter Knippertz. The valuable input after my half-term presentation was important to the writing of this thesis. At the same time, I would like to thank you for all the excellent lectures in both the bachelor's and master's degree programs that were key in deepening my interest and my understanding of meteorology.

In addition, I want to show my gratitude to Jens Winkler and Michael Denhard for their time to introduce me to the concept of the Arnoldi method. The meetings and discussions that we had about my project were very valuable in shaping this thesis. I am very grateful for your feedback on my thesis that gave it its final touch.

I would like to express a special thank you to my mother. I know, I am not always easy. Especially in the last few months. Thank you for being there for me every time and any time. Thank you for all your support during difficult times. Even if it sometimes means saying nothing at all.

I want to thank my friends for being there for me in the last year. But there is one person I want to specifically mention. Thank you, Michael. Thank you for being there for me when I have a problem. Thank you for getting me out of my routine and once in a while also out of my comfort zone. Thank you for the time I can spend with you, the advice you give, and of course our road trips.

Over across the big pond, there are people that I also want to thank for their support. Chris, I cannot tell you how much I value the deep discussions about life's winding roads we had. Thank you for the emails you send, the stories you tell, and the wisdom you share.

To the sound of *Dire Straits*, there are only a few words left to write. *There's so many different worlds, so many different suns, and we have just one world, but we live in different ones.* Blown away in the winds of time, the memories will forever remain. Thank you so very much.