Technical
Memo

ECMWF
European Centre for Medium-Range
Weather Forecasts

936

# Data-Driven Emulation of Flow-Dependent Background-Error Variance in Variational Data Assimilation

Wei Pan, Massimo Bonavita, Marcin Chrust, and Elías Hólm

February 24, 2026

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
*http://www.ecmwf.int/en/publications/*

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error or omission or for loss or damage arising from its use.

## Abstract

Accurate representation of flow-dependent background-error covariance is essential for modern ensemble–variational data assimilation systems, but estimating ensemble covariance from sufficiently large ensembles remains computationally expensive. In this work, we propose a machine-learning approach to emulate ensemble-derived background-error variance fields from severely under-sampled ensembles. Conditional generative machine-learning models are trained to map variance estimates derived from 5-member ensembles to those obtained from operational 50-member ECMWF Ensemble Data Assimilations (EDA). The trained models are deployed within a hybrid ensemble 4D-Variational data assimilation system and evaluated over an independent three-month period in an operational numerical weather prediction environment.

Results show that the machine-learning-emulated variance fields reproduce the structure and analysis impact of the full 50-member system, yielding comparable observation-space diagnostics and forecast error characteristics while using only a small fraction of the ensemble size. To our knowledge, this represents the first demonstration in an operational environment of machine-learning emulation of ensemble-derived background-error variance statistics within a fully cycled numerical weather prediction data assimilation framework.

## Plain Language Summary

Scientists who study the Earth system seek to estimate the current state of the atmosphere, oceans, and land surface in order to produce accurate weather and climate forecasts. To do this, they use a method called data assimilation, which combines observations (from satellites and conventional instruments) with predictions from computer models. Because both models and observations are imperfect, it is equally important to estimate their associated uncertainties.

A common way to measure uncertainty is to run many slightly different versions of the model, each representing a possible version of reality. This collection of simulations is called an ensemble. By examining how much the ensemble members differ from one another, scientists can estimate how uncertain the forecast is. However, running many ensemble simulations is computationally very expensive, especially for global weather prediction systems.

In this work, we use machine learning to reproduce the uncertainty information that would normally require a large ensemble of 50 simulations, using only 5 simulations instead. We show that our approach produces uncertainty estimates and forecast performance comparable to the full 50-member operational system, while requiring only a fraction of the computational cost.

# 1 Introduction

Data assimilation is the methodology that combines Earth system observations with model predictions to infer the most accurate estimate of the current state of the system, known as the *analysis*. These analyses can be used as initial conditions for numerical forecasts. In variational data assimilation systems such as four-dimensional variational assimilation (4D-Var), the background-error covariance matrix **B** plays a central role in determining how observational information is propagated spatially and across model variables (Bannister, 2008). Accurate specification of **B** is therefore essential for obtaining high-quality analyses and forecasts, as it governs both the spatial structure and amplitude of analysis increments.

A key requirement of modern data assimilation systems is the ability to represent flow-dependent uncertainty. In operational numerical weather prediction (NWP), flow dependence is introduced through ensemble-based methods, in which an ensemble of perturbed analyses and forecasts provides dynamically consistent realisations of atmospheric uncertainty. At the European Centre for Medium-Range Weather Forecasts (ECMWF), this role is fulfilled by the Ensemble of Data Assimilations (EDA), in which multiple 4D-Var analyses are performed using perturbed observations and model representations (Isaksen, L. and Bonavita, Massimo and Buizza, Roberto and Fisher, Mike and Haseler, J. and Leutbecher, Martin and Raynaud, Laure, 2010). The resulting ensemble provides estimates of flow-dependent background-error variances and correlations that augment static covariance models. Hybrid ensemble–variational formulations combine these ensemble-derived statistics with climatological covariance structures to improve both accuracy and robustness (Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike, 2015).

Despite their demonstrated benefits, ensemble-based covariance estimation remains computationally expensive. Each ensemble member requires an independent data assimilation and forecast integration, making the ensemble component a substantial contributor to operational cost. Furthermore, practical constraints limit ensemble size, resulting in sampling noise that must be mitigated through localisation, filtering, and spectral smoothing (Hamill *et al.*, 2001; Bonavita *et al.*, 2012). Increasing ensemble size would improve covariance estimation but is often infeasible due to computational constraints. These limitations motivate the development of alternative approaches capable of reproducing flow-dependent covariance statistics at substantially reduced cost.

Recent advances in machine learning have demonstrated the ability to learn complex atmospheric structure directly from data. Neural forecasting systems such as FourCastNet (Kurth *et al.*, 2023), Pangu-Weather (Bi *et al.*, 2023), GraphCast (Lam *et al.*, 2023), Aurora (Bodnar *et al.*, 2025), and ECMWF's Artificial Intelligence Forecasting System (AIFS) (Lang *et al.*, 2024, 2026) have achieved forecast skill comparable to, and in some cases exceeding, that of state-of-the-art numerical weather prediction models. These results highlight the ability of data-driven methods to capture multiscale spatiotemporal dependencies in high-dimensional geophysical systems.

Beyond forecasting, machine learning has increasingly been applied to approximate components of data assimilation systems, including observation operators, model error representations, and analysis mappings (Rasp and Lerch, 2018; Bocquet *et al.*, 2021; Hatfield *et al.*, 2021; Melinc and Zaplotnik, 2024; Melinc *et al.*, 2026). In particular, recent studies have demonstrated that neural networks can learn data assimilation updates directly from state and observation information, achieving performance comparable to ensemble Kalman filters while implicitly representing flow-dependent uncertainty (Bocquet *et al.*, 2026). Other work has shown that machine learning models can learn relationships between atmospheric state and associated uncertainty statistics, providing a pathway toward data-driven emulation of covariance structures (Melinc and Zaplotnik, 2024; Melinc *et al.*, 2026). These findings suggest that neural networks can extract and reproduce flow-dependent covariance information directly from atmospheric state variables.

Recent advances in deep generative modelling further provide a principled framework for learning conditional probability distributions and high-dimensional uncertainty structures. Such models are capable of representing complex, state-dependent uncertainty and offer a natural approach for emulating flow-dependent covariance statistics. Building on these developments, the present work takes a complementary and operationally focused approach by directly emulating ensemble-derived background-error variance fields within a hybrid ensemble–variational data assimilation system. To our knowledge, this represents the first demonstration of machine-learning emulation of flow-dependent ensemble variance statistics deployed within a full operational -scale numerical weather prediction data assimilation framework.

# 2 Data assimilation

## 2.1 4D-Var overview

We consider the underlying domain to be the 2-sphere, denoted by $\mathbb{S}^2$. Let $L^2(\mathbb{S}^2;\mathbb{R}^d)$ denote the space of square-integrable $\mathbb{R}^d$-valued functions on $\mathbb{S}^2$, abbreviated by $L^2(\mathbb{S}^2)$. Let $\mathscr{M}_{l\to k}: L^2(\mathbb{S}^2)\to L^2(\mathbb{S}^2)$ denote a dynamical model mapping states from time $t_l$ to $t_k$, satisfying the semi-group property

$$\mathscr{M}_{l\to k} = \mathscr{M}_{k'\to k}\circ\mathscr{M}_{l\to k'}, \quad \text{for } t_l < t_{k'} < t_k. \tag{1}$$

In scientific domains such as meteorology, $\mathscr{M}$ is typically described by a system of partial differential equations.

A model state $\mathbf{x}_k \in L^2(\mathbb{S}^2)$ at time $t_k$ is obtained as

$$\mathbf{x}_k = \mathscr{M}_{0\to k}(\mathbf{x}_0), \tag{2}$$

for some initial condition $\mathbf{x}_0 \in L^2(\mathbb{S}^2)$. We may also introduce stochastic model error in (2), but this is not required given the focus of this work.

In the classic 4D-Var problem, one is given a discrete sequence of observations $(\mathbf{y}_0,\mathbf{y}_1,\ldots,\mathbf{y}_L)$, with $\mathbf{y}_k\in\mathbb{R}^{N_\mathbf{y}}$, where $L\geq 0$ denotes the observation window length. The observations are related to the system states via observation operators $\mathscr{H}_k : L^2(\mathbb{S}^2)\to\mathbb{R}^{N_\mathbf{y}}$ and observation errors $\mathbf{v}_k$,

$$\mathbf{y}_k = \mathscr{H}_k(\mathbf{x}_k)+\mathbf{v}_k. \tag{3}$$

We assume the observation errors are independent and identically distributed Gaussian, with mean zero and covariance matrix $\mathbf{R}$.

We avoid the technical complications of defining probability measures on the infinite-dimensional Hilbert space $L^2(\mathbb{S}^2)$, and adopt the pragmatic viewpoint of treating elements of this space as realisations of Gaussian random fields[1], see (Lang, Annika and Schwab, Christoph, 2015) and references therein. Accordingly, whenever we consider probability measures on $L^2(\mathbb{S}^2)$, it is with respect to a fixed finite-dimensional subspace obtained by discretising $\mathbb{S}^2$ on a grid (e.g. latitude–longitude grid) [2] with $N_\mathbf{x}$ spatial points. The resulting state space is isomorphic to $\mathbb{R}^{N_\mathbf{x}\times d}$. Thus, we can formulate the strong-constraint[3] 4D-Var cost functional as below, on the model trajectory $\mathbf{x}=(\mathbf{x}_0,\ldots,\mathbf{x}_L)$,

$$J(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x}_0-\mathbf{x}^b\right\|^2_\mathbf{B} + \frac{1}{2}\sum_{k=0}^{L}\left(\mathscr{H}_k(\mathbf{x}_k)-\mathbf{y}_k\right)^T\mathbf{R}^{-1}\left(\mathscr{H}_k(\mathbf{x}_k)-\mathbf{y}_k\right), \tag{4}$$

for a given background state $\mathbf{x}^b$, a positive-definite self-adjoint covariance operator $\mathbf{B}$, and weighted norm $\|v\|^2_\mathbf{B} := \langle\mathbf{B}^{-1}v,v\rangle$ on $\mathbb{R}^{N_\mathbf{x}\times d}$. The aim is to find an analysis initial condition $\mathbf{x}_0^a$ such that the associated model trajectory $\mathbf{x}^a=(\mathbf{x}^a_0,\ldots,\mathbf{x}^a_L)$ minimises (4).

In practice, this optimisation is repeated (known as *DA cycling*) over consecutive non-overlapping time windows. In such a cycled DA setup, the model propagates the analysis to the start of the next window, yielding the new background state.

[1] Concretely, a Gaussian random field on $\mathbb{S}^2$ is a collection of random vectors $\{\mathbf{x}(x)\}_{x\in\mathbb{S}^2}$ such that every finite subcollection $\{\mathbf{x}(x_1),\ldots,\mathbf{x}(x_n)\}$ is jointly multivariate Gaussian.

[2] Alternatively, one could consider truncated Karhunen–Loève expansions of elements of $L^2(\mathbb{S}^2)$ (Lang, Annika and Schwab, Christoph, 2015).

[3] When the model is perfect, it is called *strong-constraint* 4D-Var. When the model is imperfect, it is called *weak-constraint* 4D-Var.

## 2.2 Ensemble data assimilation

In addition to the background state $\mathbf{x}^b$, the covariance operator $\mathbf{B}$ is required for the cost function (4) to be well-defined. In meteorological applications, the background state $\mathbf{x}^b$ is an a priori estimate of an unknown true state $\mathbf{x}^{\text{true}}$, and $\mathbf{B}$ encodes the covariance structure of the background errors

$$\delta\mathbf{x}^b := \mathbf{x}^{\text{true}} - \mathbf{x}^b, \tag{5}$$

including spatial and vertical correlations. In cycled data assimilation, $\mathbf{B}$ is expected to be flow-dependent and anisotropic, reflecting the evolving dynamical state of the atmosphere and the spatial inhomogeneity and temporal evolution of the observing system.

Operationally, the Ensemble of Data Assimilations (EDA) run at ECMWF provides flow-dependent estimates of analysis and background-error uncertainty (Isaksen, L. and Bonavita, Massimo and Buizza, Roberto and Fisher, Mike and Haseler, J. and Leutbecher, Martin and Raynaud, Laure, 2010). The EDA consists of an ensemble of perturbed 4D-Var assimilation cycles, in which observations, surface forcing fields, and model are stochastically perturbed[4]. Each ensemble member therefore provides an independent realisation of a dynamically consistent analysis and short-range forecast. Under statistically consistent perturbations, the ensemble spread provides an estimate of background-error uncertainty that captures “errors-of-the-day”, in contrast to purely climatological covariance models. This ensemble-derived information is subsequently filtered to reduce sampling noise before being incorporated into the analysis 4D-Var system in both perturbed and control members.

In practice, $\mathbf{B}$ is constructed empirically from such an ensemble of background states, each obtained from an independent 4D-Var optimisation with perturbed inputs. Additionally, $\mathbf{B}$ is typically represented as an inner block matrix $\mathbf{B}_u$ with one block per model variable and then composed with suitably defined balance operator $\mathbf{K}$, i.e. $\mathbf{B} = \mathbf{K}^{-1}\mathbf{B}_u\mathbf{K}^{-\top}$ (see Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike (2015, Eq (1))). We refer the reader to (Isaksen, L. and Bonavita, Massimo and Buizza, Roberto and Fisher, Mike and Haseler, J. and Leutbecher, Martin and Raynaud, Laure, 2010; Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike, 2015) for detailed exposition of the operational construction.

In this work, as exploratory research, we consider only the diagonal part of $\mathbf{B}$, i.e. the variance of the background errors at each latitude–longitude location $(\lambda,\phi) \in \mathbb{S}^2$. We denote the variance field by $\sigma^2 : \mathbb{S}^2 \to \mathbb{R}^d_+$,

$$\sigma^2(\lambda,\phi) := \operatorname{diag}(\mathbf{B})(\lambda,\phi), \tag{6}$$

and assume $\sigma^2 \in L^2(\mathbb{S}^2;\mathbb{R}^d_+)$.

### 2.2.1 *Ensemble statistics*

Let $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N \in L^2(\mathbb{S}^2)$ denote an ensemble of size $N$ of background states. We estimate $\sigma^2$ using the standard unbiased sample variance estimator and denote the estimate by $\hat{\sigma}^2_N$,

$$\hat{\sigma}^2_N := \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}^n - \bar{\mathbf{x}})^{\otimes 2}, \qquad \bar{\mathbf{x}} := \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^n, \tag{7}$$

where $\otimes$ denotes the pointwise Hadamard product. Equivalently,

$$\hat{\sigma}^2_N = \frac{1}{2N(N-1)}\sum_{n,m=1}^{N}(\mathbf{x}^n - \mathbf{x}^m)^{\otimes 2}. \tag{8}$$

[4] Since IFS cycle 49r1, stochastically perturbed parametrisation (SPP) scheme ((Lang *et al.*, 2021)) replaced stochastically perturbed parametrisation tendencies (SPPT) scheme.

Figure 1 offers a visual interpretation of (8). We also consider the sample standard deviation $\hat{\sigma}_N$ obtained by taking the pointwise square root of $\hat{\sigma}_N^2$. Henceforth, we refer to $\hat{\sigma}_N^2$ and $\hat{\sigma}_N$ as the *ensemble variance* and *ensemble standard deviation*, respectively.

Pointwise properties of $\hat{\sigma}_N^2$ are well known. Under the assumption that $\mathbf{x}^n$, $n = 1,\dots,N$ are i.i.d., for each component $[\cdot]_i$, $i = 1,\dots,d$, and $(\lambda,\phi) \in \mathbb{S}^2$,

$$(N-1)\left([\hat{\sigma}_N^2]_i/[\sigma^2]_i\right)_{(\lambda,\phi)} \sim \chi^2_{N-1}, \tag{9}$$

and

$$\sqrt{N}\left([\hat{\sigma}_N^2 - \sigma^2]_i\right)_{(\lambda,\phi)} \xrightarrow{\mathrm{d}} \sqrt{\left[\mu_4 - (\sigma^2)^2\right]_{i_{(\lambda,\phi)}}}\,\mathcal{N}(0,1), \tag{10}$$

where convergence is in distribution and $\mu_4$ denotes the fourth moment of $\mathbf{x}^n$.

# 3 Hybrid ML–EDA statistics

In this section, we formulate the learning problem and describe the hybrid methodology. As noted in Section 2.2, in the operational EDA each ensemble member is obtained from an independent optimisation of (4). The associated computational cost is a substantial part of the overall budget at operational centres such as ECMWF. The focus of this work is to explore a hybrid data-driven and ensemble 4D-Var approach, where a data-driven model is used to emulate the diagonal part of **B**. This choice is motivated by both relative simplicity and data availability.

From a modelling perspective, restricting attention to the diagonal isolates the impact of flow-dependent amplitude (local uncertainty) while avoiding the substantially more complex task of learning and validating correlation structures. This choice substantially reduces the dimensionality of the problem. From an operational perspective, the variance already plays a critical role in determining the relative weighting of observations and background information in the 4D-Var system, and errors in variance estimates can directly lead to suboptimal assimilation behaviour. By first establishing that machine learning can reliably emulate flow-dependent variance fields from a small ensemble, we aim to provide a foundation for future work extending the approach to correlation structures, which involve additional challenges related to scale, balance relationships, and sampling noise.

## 3.1 ML formulation

Over a DA cycling window, let $(\mathbf{x}^n)_{n=1}^N$ denote a (full) ensemble of background states. Let $(\mathbf{x}^{n_m})_{m=1}^M$ be a subset of size $M$ of the full ensemble, with $M < N$. For notational simplicity, we take the subset to be the first $M$ members of the full ensemble. Let $\hat{\sigma}_N^2$ denote the local variance estimate (8) computed from the full ensemble, and $\hat{\sigma}_M^2$ the corresponding estimate computed from the subset. Then

$$\hat{\sigma}_N^2 = \alpha_{M,N}\,\hat{\sigma}_M^2 + \varepsilon_{M,N}^{\mathrm{exact}}, \tag{11}$$

where $\alpha_{M,N} = \frac{M(M-1)}{N(N-1)}$ and

$$\varepsilon_{M,N}^{\mathrm{exact}} = \frac{1}{2N(N-1)}\left(2\sum_{n=M+1}^{N}\sum_{m=1}^{M} + \sum_{n,m=M+1}^{N}\right)(\mathbf{x}^n - \mathbf{x}^m)^{\otimes 2}, \tag{12}$$

see Figure 1 for an illustration.

We consider the following supervised learning problem. Given a training dataset consisting of pairs of ensemble standard deviation estimates

$$(\hat{\sigma}_{M,k}, \hat{\sigma}_{N,k}), \qquad k = 1,2,\dots,K, \tag{13}$$

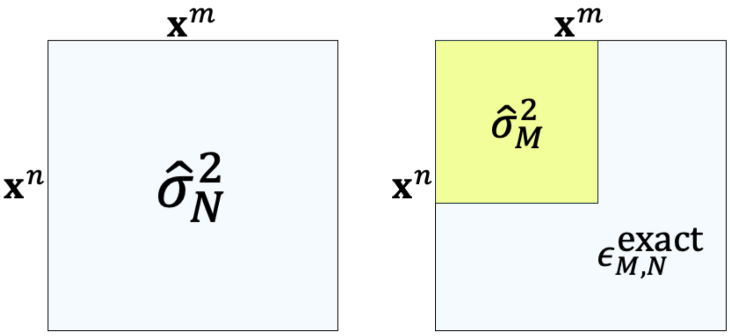


Figure 1: Illustration of the double-summation variance estimator (8), and its small ensemble approximation (11). Each squared-region represents averaging over pairwise squared differences between ensemble members indexed by $m$ and $n$.

collected over $K$ different DA cycling windows, construct a parametric conditional probability distribution

$$\mathrm{p}_\theta\left(\tilde{\sigma}\,\middle|\,\hat{\sigma}_M\right) \tag{14}$$

on $\mathbb{R}^{N_\mathbf{x}\times d}$. We interpret $\hat{\sigma}_N$ as realised values of a random field $\tilde{\sigma}$.

The goal is to find parameters $\theta^* \in \Theta$ that maximise the conditional likelihood over the training data,

$$\theta^* = \underset{\theta\in\Theta}{\arg\max} \prod_{k=1}^{K} \mathrm{p}_\theta(\hat{\sigma}_{N,k}\,|\,\hat{\sigma}_{M,k}). \tag{15}$$

By sampling from $\mathrm{p}_{\theta^*}$, we obtain an estimate of $\hat{\sigma}_N$ from a smaller ensemble of size $M$. In other words, conditioned on the available $M$ ensemble members, the learned distribution emulates the variance estimate from the full-ensemble, including the contribution from the missing $N-M$ members.

A detailed description of the machine learning methodology, including the model architecture, loss function, training procedure, and datasets used, is provided in Appendix A.

### 3.2 Hybrid ML and cycled ensemble 4D-Var

It is well known in the DA literature that an accurate estimate of the background error covariance **B** is crucial in variational data assimilation because it determines how observational information is propagated spatially and across variables, (Bannister, 2008). If **B** is poorly estimated, observational influence may be misrepresented and key features of the analysis may be under- or overemphasised, potentially leading to suboptimal analyses. A well-estimated **B** ensures effective information propagation, especially in regions with sparse observations, and is thus essential for producing accurate and dynamically consistent forecasts.

Therefore, if samples (or suitable point estimators) from the trained distribution (14) provide good approximations to the full-ensemble diagonal of **B**, we expect an $M$-member ensemble 4D-Var system coupled with $\mathrm{p}_{\theta^*}$ to exhibit behaviour close to that of a full $N$-member ensemble. The precise interpretation of “good” and the evaluation criteria are discussed in Section 4.2.

Algorithm 1 describes a hybrid ML–EDA cycling step that replaces the sub-ensemble variance estimate by an ML-based estimate. As written, the correlation component of $\tilde{\mathbf{B}}$ is estimated from the $M$-member ensemble. We emphasise that we do not expect the hybrid setup in Algorithm 1 to match the full-ensemble system, since we target only the diagonal of **B** and do not address sampling error in the correlation structure. Modelling the correlation component using machine learning will be investigated in follow-up work. In our numerical studies, however, and to isolate the impact of ML-estimated variances on analyses, the correlation component is estimated using the full $N$-member ensemble (Section 4.1).

**Algorithm 1** Hybrid ML-for-variance and ensemble 4D-Var over one DA cycle window

```
Require: Ensemble of backgrounds $\{\mathbf{x}^{b,m}\}_{m=1}^{M}$, trained distribution $\mathrm{p}_{\theta^*}$, ensemble of perturbed observations $\{\mathbf{y}_0^m, \dots, \mathbf{y}_L^m\}_{m=1}^{M}$
Ensure: Analysis states $\mathbf{x}^{a,m}$, $m = 1, 2, \dots, M$.
Compute ensemble approximations to $\mathbf{B}$ using $\{\mathbf{x}^{b,m}\}_{m=1}^{M}$ and obtain $\hat{\sigma}_M^2$.
Evaluate the conditional first moment of $\mathrm{p}_{\theta^*}(\cdot \mid \hat{\sigma}_M)$,
    $\bar{\sigma}_M := \mathbb{E}_{\mathrm{p}_{\theta^*}}(\tilde{\sigma} \mid \hat{\sigma}_M).$    (16)
Replace $\hat{\sigma}_M^2$ by $\bar{\sigma}_M^2$ in the variance component. Denote the modified covariance estimate by $\tilde{\mathbf{B}}$.
for $m = 1, 2, \dots, M$ do
    Solve the 4D-Var optimisation problem (4) using $\mathbf{x}^{b,m}$, $\tilde{\mathbf{B}}$, and perturbed observations $\mathbf{y}_0^m, \dots, \mathbf{y}_L^m$, to obtain $\mathbf{x}^{a,m}$.
end for
```

A mathematical analysis of the impact of using an approximate background-error covariance on the resulting 4D-Var analysis is provided in Appendix B. This analysis characterises how errors in the **B** influence the analysis increment and provides theoretical context for interpreting the behaviour of the hybrid ML–ensemble 4D-Var system. Readers primarily interested in the operational and empirical aspects of the method may proceed without reference to these details.

# 4 Numerical setup and experiment results

## 4.1 Hybrid 4DVar-ML experiment setup

For operational medium-range weather forecasting, ECMWF employs an ensemble of 51 members since IFS Cycle 46r1, June 2019: one unperturbed control member and fifty perturbed members (European Centre for Medium-Range Weather Forecasts (ECMWF), 2019). For estimating the background error covariance matrix **B**, only the 50 perturbed members are used. Global analyses are produced twice daily, at 09:00 and 21:00 UTC, using 12-hour assimilation windows.

We assess the effectiveness of the ML sample variance model (14) when deployed in cycled ensemble 4D-Var experiments within the physics-based Integrated Forecasting System (IFS). All experiments are conducted using a standard research configuration of the weak-constraint 4D-Var system, with a 12-hour assimilation window, and the latest available IFS cycle 50r1. The evaluation period comprises the *summer* months of June–August 2022, which lie outside the training and validation data sets.

To isolate and quantify the impact of the ML variance model on 4D-Var analyses, all experimental setups use the full 50-member ensemble to estimate **B**, except that the variance component is modified as described below. Specifically, we retain the full-ensemble estimates for all non-variance components of **B**, while substituting alternative variance estimates. We consider the following setups:

a. Evaluate the ensemble variance (8) using 5 members, and use this 5-member estimate for the variance component of **B**.

b. Use the conditional first moment of the ML model, from (16),

$$\bar{\sigma}_5 := \mathbb{E}_{\mathrm{p}_{\theta^*}}\left(\tilde{\sigma} \,\middle|\, \hat{\sigma}_5\right), \tag{17}$$

Each setup is compared against the default operational control.

In summary, the reference benchmark employs the full 50-member ensemble for variance estimation. We evaluate the impact on the analyses of using reduced sub-ensembles for direct variance estimation (setup a) versus replacing the sub-ensemble variance estimates with ML-based variance estimates (setup b).

## 4.2 4DVar and ML model results

The motivation for learning an emulator of the diagonal of the background error covariance matrix is to obtain an approximation that is as close as possible to the full-ensemble estimate. In principle, as noted in Section 3.2 and Appendix B, the closer this approximation is to the true covariance, the more appropriately observational information is propagated in the variational analysis, and the closer the resulting analysis is expected to be to that obtained using a full $N$-member ensemble.

In this section, we evaluate the trained ML model by examining whether coupling an $M$-member ensemble 4D-Var system with the learned distribution $\mathrm{p}_{\theta^*}$ defined in (14) leads to behaviour that approaches that of the corresponding full-ensemble system. While the overarching goal is to approximate the full-ensemble diagonal of $\mathbf{B}$ as accurately as possible, the specific model architecture and training strategy employed here represent a first exploratory attempt toward this objective, and no claim of optimality is made.

The assessment therefore focuses on diagnosing how the data assimilation system responds to the learned covariance representation. In particular, we evaluate the trained model through (i) a qualitative assessment of the reconstructed variance fields, (ii) consistency with observations as measured using *Iver* (ECMWF 'Verify' package, see Geer (2016)) based observation statistics, and (iii) the impact on subsequent forecast errors.

The numerical results are therefore organised as follows. We first provide a qualitative, field-based assessment of ML-emulated variance estimates. Specifically, we visually compare variance fields obtained from a reduced sub-ensemble, the corresponding ML-emulated variance fields, and the full-ensemble reference. This comparison is used to assess whether the ML model reproduces the large-scale structure, spatial coherence, and magnitude of the full-ensemble variance, and whether it reduces obvious sampling artefacts present in small-ensemble estimates. To this end, spectral diagnostics are also included.

We then evaluate the impact of ML-emulated variances on 4D-Var analyses using Iver diagnostics for selected observing systems. This provides a quantitative measure of how changes in the background-error variance affect the use of observational information within the data assimilation system.

Finally, we assess the downstream impact of ML-emulated variances on forecast quality by comparing forecast errors against appropriate verification datasets. Together, these three perspectives provide a coherent framework for evaluating the potential effectiveness of the proposed hybrid ML–EDA approach.

### *4.2.1 Qualitative assessment of ML-emulated variance fields*

We assess the ability of the trained ML model to emulate full-ensemble variance fields from a severely under-sampled sub-ensemble using both field-based and spectral diagnostics. The objective is to determine whether the ML-emulated ensemble standard deviation fields exhibit spatial structure and scale-dependent behaviour consistent with those used operationally within the variational analysis system.

All diagnostics shown in this subsection are computed using operational archive data valid at 21:00 UTC on 1 June 2022, which lies entirely outside the training and validation datasets. The results therefore reflect genuine generalisation rather than interpolation within the training distribution. We focus on model level 74 (approximately 200 hPa), a dynamically active region characterised by strong jets and sharp gradients.

**Field-based comparison.** Figure 2 shows ensemble standard deviation fields for vorticity (vo) and divergence (div) at model level 74. In each panel, the four subplots display: top-left the 5-member ensemble estimate (5-es, input to the ML model), top-right the raw 50-member ensemble estimate (50-es), bottom-left the ML-emulated estimate (ml-es), and bottom-right the wavelet filtered 50-member estimate (50-ses).

The 50-ses field is obtained by applying wavelet filtering to the raw 50-es estimate, and blended with a small fraction of climatology, in order to suppress non-Gaussian sampling features, see Bonavita *et al.* (2012) for details. This filtered es field is the variance estimate that constitutes the background error covariance **B** in the 4D-Var system and therefore represents the operational target for emulation.

For both vo and div, the 5-es estimates exhibit pronounced sampling noise, particularly at small spatial scales, with fragmented structures and locally exaggerated extrema. In contrast, the ML-emulated fields recover the large-scale organisation and spatial coherence of the ensemble variance while substantially reducing small-scale artefacts. Visually, the ml-es field is much closer to the 50-ses reference than to the raw 5-es input.

The ml-es fields are also smoother than the raw 50-es fields that was used as the training target. This behaviour is expected: the MSE/Huber-type training loss used during training penalises high-frequency discrepancies and implicitly favours wavelet filtered reconstructions. As a result, the learned mapping naturally aligns more closely with the operationally filtered 50-ses field than with the unfiltered 50-es estimate.

**Spectral comparison and ratio diagnostics.** To quantify scale-dependent differences, we analyse the isotropically averaged power spectra of the ensemble standard deviation fields. Figures 3(a–b) show spectra for ml-es, 50-ses, 5-es, and 50-es, together with logarithmic spectral ratios computed relative to 50-ses:

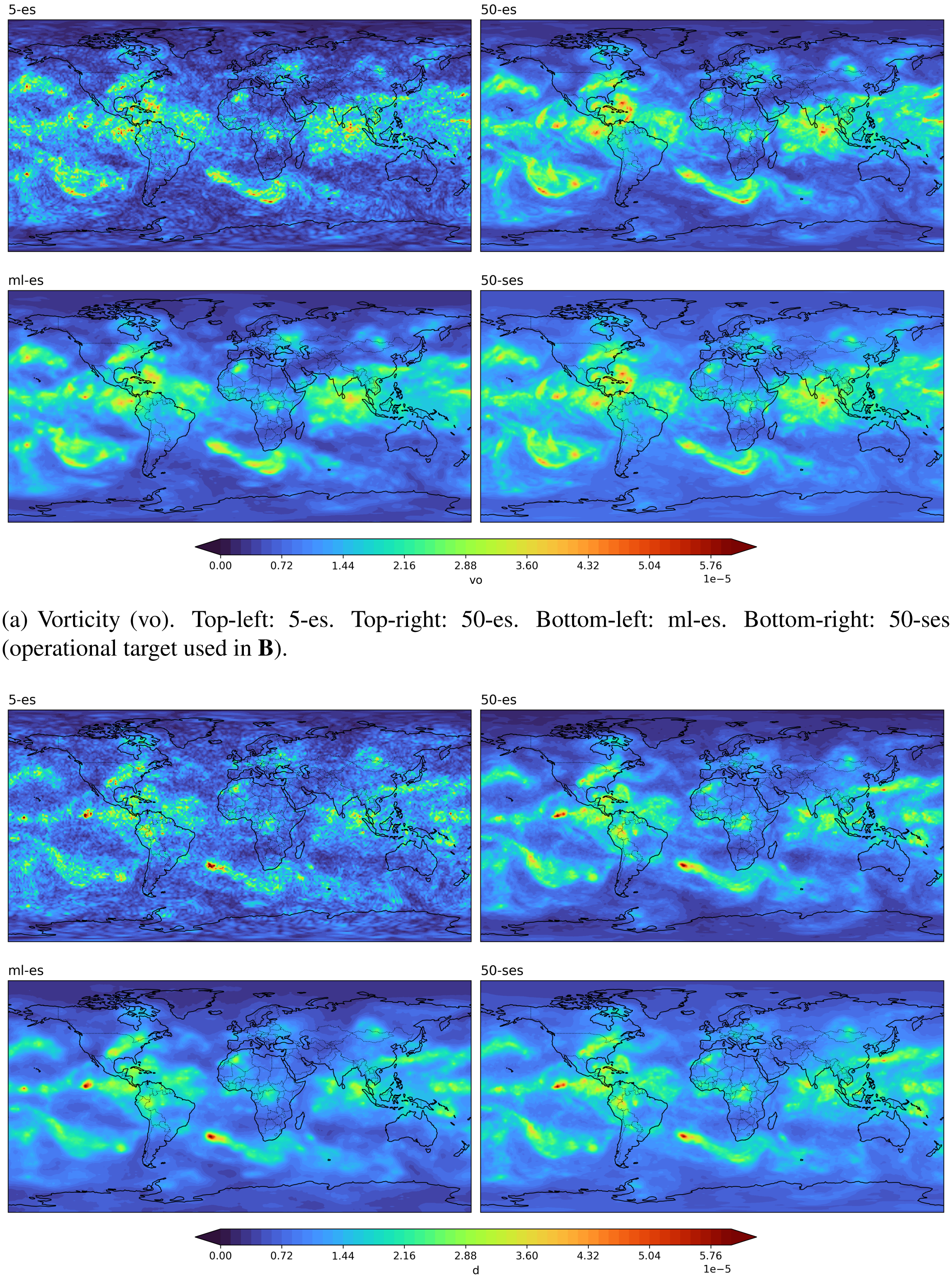
$$R_X(k) = \log_{10}\left(\frac{P_X(k)}{P_{\text{50-ses}}(k)}\right), \tag{18}$$

where $P_X(k)$ denotes the isotropically averaged power spectrum of configuration $X$. For both vo and div, the ml-es spectra closely follow 50-ses across a wide range of wavenumbers, while 5-es exhibits substantial high-wavenumber excess associated with sampling noise. Residual small-scale discrepancies are expected given information lower bounds associated with the severely under-sampled 5-member input (see Appendix B). The proposed ML model cannot reconstruct variance at scales for which insufficient information is present in the 5-member input.

**Limitations.** The close agreement observed for vorticity and divergence does not hold uniformly across all variables. Figure 4 shows the corresponding spectral comparison for ozone (o3) at model level 74. In this case, the 50-es and 50-ses spectra are already relatively close, but the ml-es spectrum exhibits additional smoothing at higher wavenumbers. This suggests that, for ozone, the current *uniform* model architecture and training strategy may introduce excessive small-scale damping, indicating scope for further tailored refinement for different variables.

Nevertheless, even in this case, the ml-es log-ratio remains substantially closer to zero than the 5-es log-ratio over the wavenumber range $10^1$–$10^2$, and remains within approximately $[-1,1]$ beyond $10^2$, whereas the 5-es ratio shows much larger departures. Thus, the ML model still provides a marked improvement over the raw small-ensemble estimate.

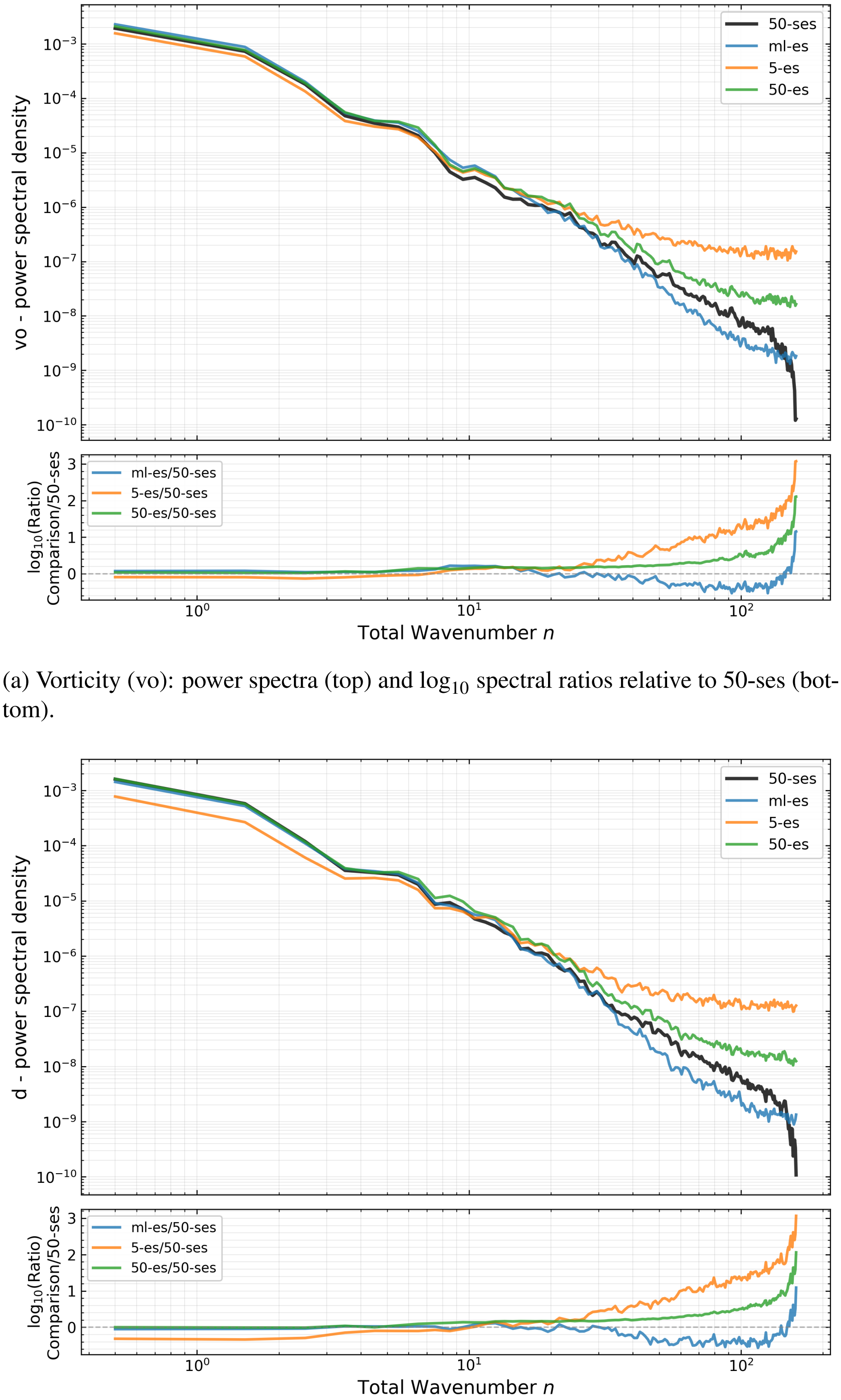
As will be discussed later in the context of Iver satellite observation fit (Section 4.2.2), the additional smoothing introduced by the ML model for ozone is associated with improved fitting of ozone-sensitive channels of satellite instruments, highlighting the interplay between spectral characteristics of **B** and observation-space diagnostics.



(a) Vorticity (vo). Top-left: 5-es. Top-right: 50-es. Bottom-left: ml-es. Bottom-right: 50-ses (operational target used in **B**).

(b) Divergence (div). Same layout as in (a).

Figure 2: Ensemble standard deviation fields at model level 74 (approximately 200 hPa) valid at 21:00 UTC on 1 June 2022 (outside the training/validation period). Each panel contains a 2×2 comparison between the 5-member estimate (5-es), the raw 50-member estimate (50-es), the ML-emulated estimate (ml-es), and the wavelet filtered 50-member estimate (50-ses) used operationally in the 4D-Var background error covariance.



(a) Vorticity (vo): power spectra (top) and $\log_{10}$ spectral ratios relative to 50-ses (bottom).

(b) Divergence (div): same diagnostics and reference as in (a).

Figure 3: Spectral diagnostics for ensemble standard deviation fields at model level 74, valid at 21:00 UTC on 1 June 2022. Each panel shows isotropically averaged power spectra for 50-ses, ml-es, 5-es, and 50-es (top) and logarithmic spectral ratios $\log_{10}(P_X/P_{50\text{-ses}})$ (bottom).

Overall, these qualitative and spectral diagnostics demonstrate that the trained ML model substantially mitigates sampling artefacts present in small-ensemble variance estimates and reproduces the operationally relevant scale-dependent structure of the background error covariance. Residual discrepancies are consistent with information-theoretic limits and suggest targeted directions for future architectural and training improvements.

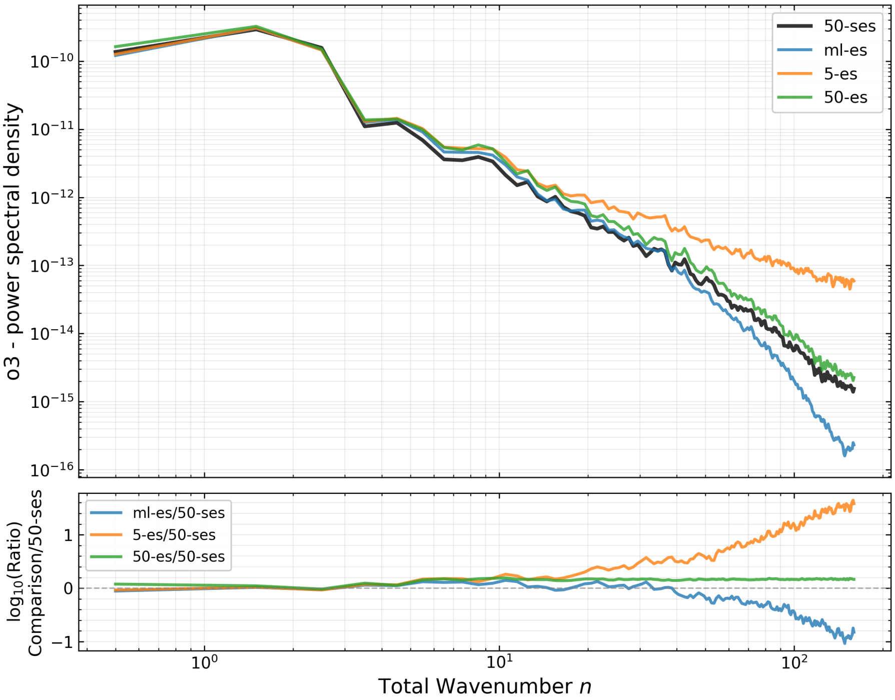


Figure 4: Power spectra and logarithmic spectral ratios for ozone (o3) at model level 74. The ml-es estimate remains closer to the operational 50-ses reference than the raw 5-es estimate across most scales, but exhibits additional high-wavenumber smoothing relative to 50-es, indicating potential scope for refinement of the ML approach.

### 4.2.2 *Impact on observation fit*

We next assess the impact of the ML-emulated variance fields on data assimilation diagnostics in observation-space using the set of observations active in the final analysis trajectory after quality control and screening, referred to as `used`. We do not expect improvements relative to the full 50-member **B**. Instead, the objective is to assess whether using the emulated **B** derived from just 5 members can reproduce the performance of the control experiment.

We focus on the normalised observation statistics (hereafter *obstats*), reported as the analysis (O-A) and background (O-B) departure standard deviations relative to the OD control experiment, with the vertical 100% line indicating the control value, see Figures 5a–5d. The figures show global `used` obstats for three satellite data types (AMSU-A all-sky radiances, GEOS  water-vapour radiances, and IASI radiances) and for conventional radiosonde temperature (TEMPT). All observation-space diagnostics shown in this section are computed over the period 00Z 1 June 2022 to 00Z 31 August 2022. This three-month summer period lies entirely outside the training and validation datasets used for the ML model and therefore provides an independent assessment of generalisation.

The left-hand panel of each plot summarises the normalised standard deviation of analysis departures (O–A), while the right-hand panel summarises the normalised standard deviation of background departures (O–B) over the same used sample. Since modifying the diagonal of **B** alters the assumed background uncertainty, it can change both the relative weighting of observations in the minimisation and the effective screening/downweighting through VarQC[5]; obstats thus provide a compact, observation-space measure of how the altered covariance representation interacts with the assimilation system.

**Balanced and unbalanced estimates.** We compare three configurations against the control: ES-5, ML-UNBAL-ES-5, and ML-BAL-ES-5. Here ES-5 uses the raw 5-member ensemble estimate of the background-error standard deviation. The ML-BAL-ES-5 configuration uses the ML-emulated variance field conditioned on a 5-member estimate computed from *balanced* control variables, consistent with the input representation used during training (see Appendix A.3.1). In contrast, ML-UNBAL-ES-5 uses a 5-member estimate derived from variance fields that include *unbalanced* control variables (in particular unbalanced temperature, divergence, and surface pressure), and is therefore statistically inconsistent with the training distribution. This configuration is included as a stress test of robustness to input inconsistency and to assess the sensitivity of the ML emulator to deviations from the training-consistent control-variable definition. See Appendix A.3.1 for a short discussion on the differences, and (Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike, 2015) for more detailed discussion on balanced and unbalanced control variables.

**Impact of ML-emulated variance on satellite and conventional observations.** Figures 5a–5d show a consistent qualitative pattern: the ML-based configurations generally yield larger improvements relative to the raw ES-5 configuration for satellite observations than for conventional observations. These improvements are particularly pronounced for channels sensitive to upper-tropospheric and stratospheric structure, including the IASI upper water vapour channels (channel indices $\gtrsim 1038$), geostationary water vapour channels, and the AMSU-A stratospheric channels (channel indices $\gtrsim 9$). In these bands, the raw ES-5 configuration exhibits clear departures from the 50-member control, consistent with sampling noise and misrepresentation of background-error variance and vertical correlations in the small ensemble. The ML-emulated variance (particularly ML-BAL-ES-5) moves the obstats closer to the control and, for some channels, yields a tighter fit (reduced normalised standard deviation) in both O–B and O–A.

One plausible explanation is that the ML emulation introduces additional smoothing in the variance fields relative to the small-ensemble estimate, and in some cases even relative to the operationally filtered reference, which can improve the stability and representativeness of observation fits (particularly for channels sensitive to upper-tropospheric/lower-stratospheric structure).

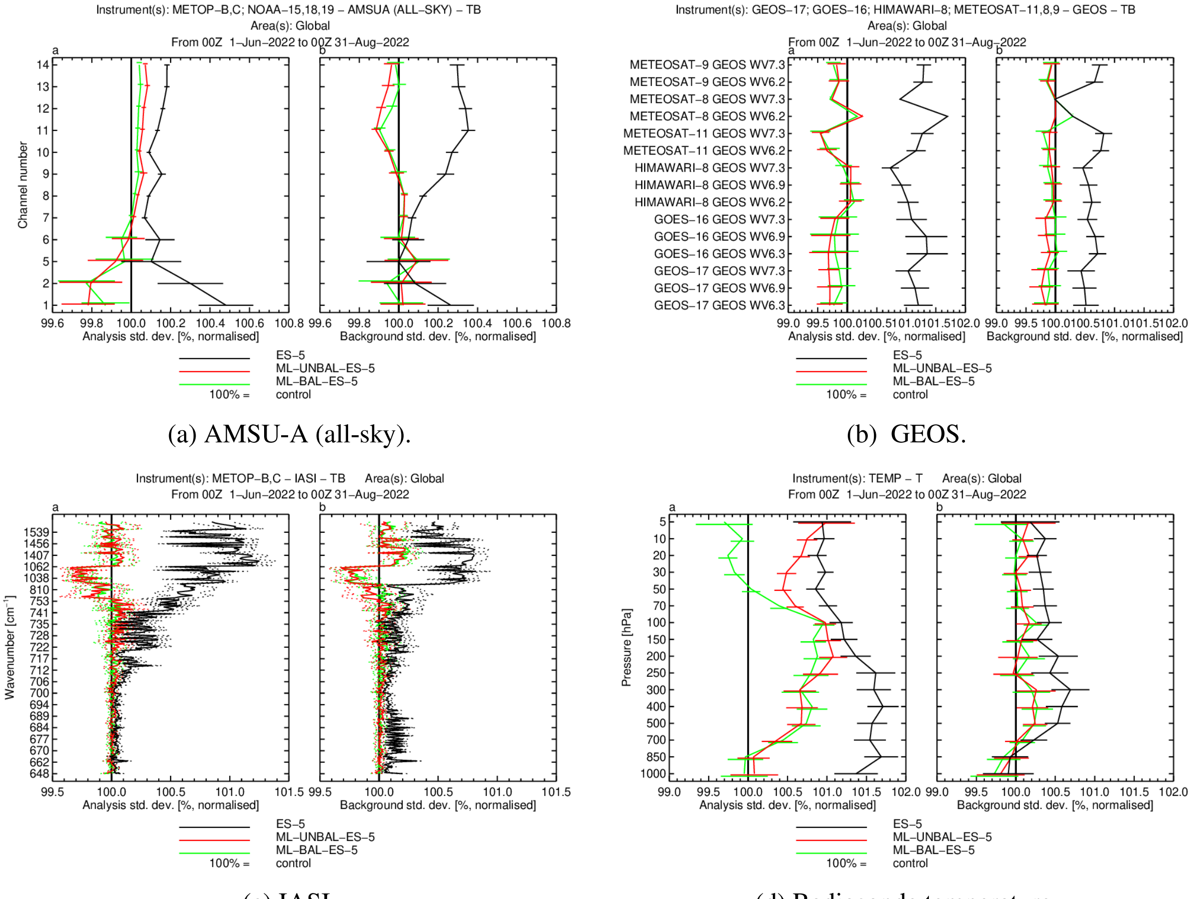
In contrast, conventional radiosonde temperature statistics exhibit smaller and more vertically confined changes. This is consistent with the pointwise and strongly vertically localised nature of radiosonde observations, which makes them more sensitive to modifications in the small-scale structure of the background-error variance compared to satellite radiances.

Differences between ML-BAL-ES-5 and ML-UNBAL-ES-5 are small, but the balanced configuration is generally more consistent across instruments and vertical levels. This suggests that maintaining consistency between training and application is beneficial for stable assimilation behaviour.

### *4.2.3 Impact on forecast errors*

In this section, we show results for assessing the impact of the ML-emulated variance fields on medium-range forecast skill. Verification is performed against the operational analysis (0001), and results are

[5] Variational Quality Control (VarQC) is a built-in mechanism in variational data assimilation that identifies and downweights observations whose analysis departures (O–A) are inconsistent with the expected statistical behaviour of the analysis.



(a) AMSU-A (all-sky).

(b) GEOS.

(c) IASI.

(d) Radiosonde temperature.

Figure 5: Normalised observation statistics (obsstats) for data active in the final analysis trajectory over 00Z 1 Jun 2022 to 00Z 31 Aug 2022, global coverage. Left panels show analysis departure standard deviations (O–A) and right panels show background departure standard deviations (O–B), each normalised by the control (OD 0001). Curves compare ES-5 (raw 5-member estimate), ML-UNBAL-ES-5 (ML emulation driven by an input definition inconsistent with training), and ML-BAL-ES-5 (training-consistent balanced input) against the 100% control reference.

expressed in terms of the normalised change in RMS error (drmse) relative to the control experiment:

$$\text{drmse} = \frac{\text{RMS}_{\text{exp}} - \text{RMS}_{\text{control}}}{\text{RMS}_{\text{control}}}. \tag{19}$$

Negative values (blue shading) therefore indicate an improvement over the control, while positive values (red shading) indicate degradation. Cross-hatching marks regions significant at the 95% confidence level (Sidak correction for multiple testing).

We show two dynamically important fields: vector wind (VW) and geopotential height ($Z$), verified as functions of latitude and pressure for lead times from T+12 to T+72. Longer lead times are excluded due to a lack of statistically significant signals. Forecast verification is performed over the same independent evaluation period as the observation verification discussed in the previous section, from 00Z 1 June 2022 to 00Z 31 August 2022. This period is fully disjoint from the ML training and validation datasets and therefore provides a stringent out-of-sample assessment of forecast impact.

**Vector wind.** Figure 6 compares the change in RMS error for VW in the ML-BAL-ES-5 and ES-5 configurations.

The raw small-ensemble configuration (ES-5) exhibits widespread and statistically significant degradations at short lead times (T+12 and T+24), particularly in the upper troposphere and lower stratosphere across both hemispheres. These degradations are most pronounced in the extra tropics and extend vertically through much of the troposphere. Although the signal becomes more mixed at longer lead times, coherent regions of degradation persist, especially in the Southern Hemisphere mid and high latitudes.

In contrast, the ML-emulated configuration (ML-BAL-ES-5) produces substantially more neutral behaviour. At short lead times, the strong and spatially coherent degradations seen in ES-5 are markedly reduced. The error differences are generally smaller in magnitude and less spatially organised, with fewer statistically significant regions.

This behaviour is consistent with the earlier spectral diagnostics: the raw 5-member ensemble variance contains substantial small-scale sampling noise, which can lead to misrepresentation of background-error amplitudes and therefore suboptimal weighting of observations. The ML-emulated variance, being smoother and closer in scale distribution to the operational reference, mitigates these imbalances and yields a more stable short-range analysis-to-forecast transition.

**Geopotential height.** A similar pattern is evident for geopotential height (Figure 7). The ES-5 configuration again shows notable degradations at short lead times, particularly in the upper troposphere and lower stratosphere of the extratropics. These signals are statistically significant in several latitude bands and are dynamically consistent with the wind degradations, reflecting coupled dynamical error growth.

The ML-BAL-ES-5 configuration, by contrast, exhibits a considerably more neutral response. Short-range degradations are reduced in spatial extent and magnitude, and the overall pattern is closer to zero across most lead times.

**Summary of forecast impact.** Taken together, these results indicate that replacing the raw 5-member variance estimate with its ML emulation substantially stabilises the forecast behaviour of the system. Importantly, the ML-based variance structure is driven solely by a 5-member ensemble estimate, yet its forecast-error characteristics are markedly more neutral and considerably closer to the 50-member control than those obtained using the raw small-ensemble variance.

Whereas the ES-5 configuration introduces coherent and statistically significant short-range degradations—particularly in the upper troposphere and lower stratosphere—the ML-BAL-ES-5 experiment largely suppresses these systematic penalties. The resulting error patterns are more spatially diffuse, smaller in magnitude, and closer to zero across all lead times shown. In this sense, the ML emulation recovers much of the stabilising effect that would otherwise require a substantially larger ensemble.

Although neither configuration yields uniformly positive improvements at all lead times and levels, the ML-emulated variance consistently avoids the pronounced short-range degradations associated with undersampling. This behaviour suggests that the learned mapping effectively regularises the noisy 5-member covariance estimate, producing a background-error representation whose forecast impact is broadly comparable to that of the full 50-member reference, despite being constructed from a severely reduced ensemble.

# 5 Summary

In this study, we investigated a hybrid machine learning–ensemble data assimilation framework for emulating flow-dependent background-error variance fields in a global ensemble 4D-Var system. Specifically, we trained conditional variational autoencoder models to learn a mapping from small-ensemble variance estimates to variance estimates obtained from a much larger ensemble. The learned emulator was then coupled with the ECMWF 4D-Var and Integrated Forecasting systems to assess its ability to reproduce full-ensemble behaviour over an independent evaluation period.

Qualitative and spectral diagnostics demonstrate that the ML emulator substantially reduces sampling noise present in small-ensemble variance estimates and reproduces the large-scale spatial structure and scale-dependent characteristics of operational variance fields. In particular, the ML-emulated variance fields exhibit spectral behaviour that is significantly closer to the operationally used smoothed full-ensemble estimates than to the raw small-ensemble input.

Observation-space diagnostics show that the ML-emulated variance fields restore the consistency between the assimilation system and satellite observations, especially for radiance channels sensitive to upper-tropospheric and stratospheric structure, where raw small-ensemble estimates exhibit clear deficiencies. The ML-emulated configuration produces observation-fit statistics that closely match those of the full operational ensemble. Conventional observations exhibit smaller changes, reflecting their different spatial sampling characteristics and sensitivity.

Forecast verification further demonstrates that replacing the raw small-ensemble variance with its ML-emulated counterpart substantially stabilises forecast behaviour. While the raw small-ensemble configuration introduces significant systematic short-range degradations, the ML-emulated configuration produces forecast errors that are markedly more neutral and substantially closer to those obtained using the full operational ensemble. Importantly, this level of performance is achieved using variance information derived from only 5 ensemble members, highlighting the potential of ML-based emulation to recover much of the benefit of larger ensembles at significantly reduced computational cost.

These results demonstrate that machine learning models can effectively emulate flow-dependent variance fields in operational data assimilation systems, enabling small ensembles to reproduce the performance of much larger ensembles while mitigating sampling noise. At the same time, the present study focuses exclusively on the diagonal component of the background-error covariance matrix. Extending the approach to correlation structures represents a more challenging problem and remains an important direction for future work.

More broadly, this work establishes a proof of concept for hybrid machine learning–ensemble approaches to covariance modelling in operational-scale data assimilation. The results suggest that data-driven emulation can complement traditional ensemble methods and may offer a promising pathway toward more efficient and scalable uncertainty quantification in numerical weather prediction systems. However, we note that the robustness of the emulation under substantial changes to the observing system has not been

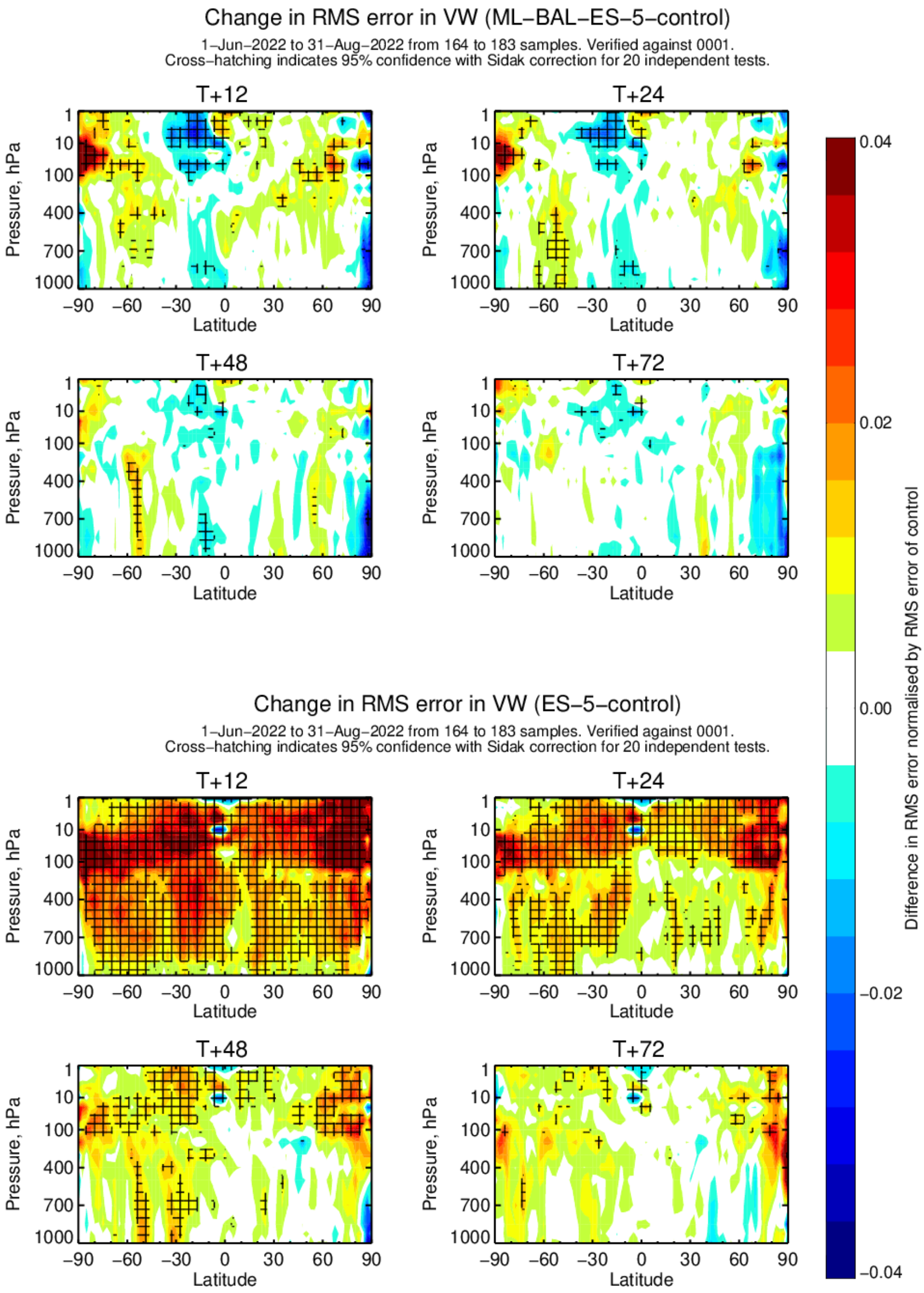


Figure 6: Normalised change in RMS forecast error for vector wind (VW) in ML-BAL-ES-5 (top) and in ES-5 (bottom) relative to the control. Blue (negative) values indicate improvement. Cross-hatching denotes 95% statistical significance.

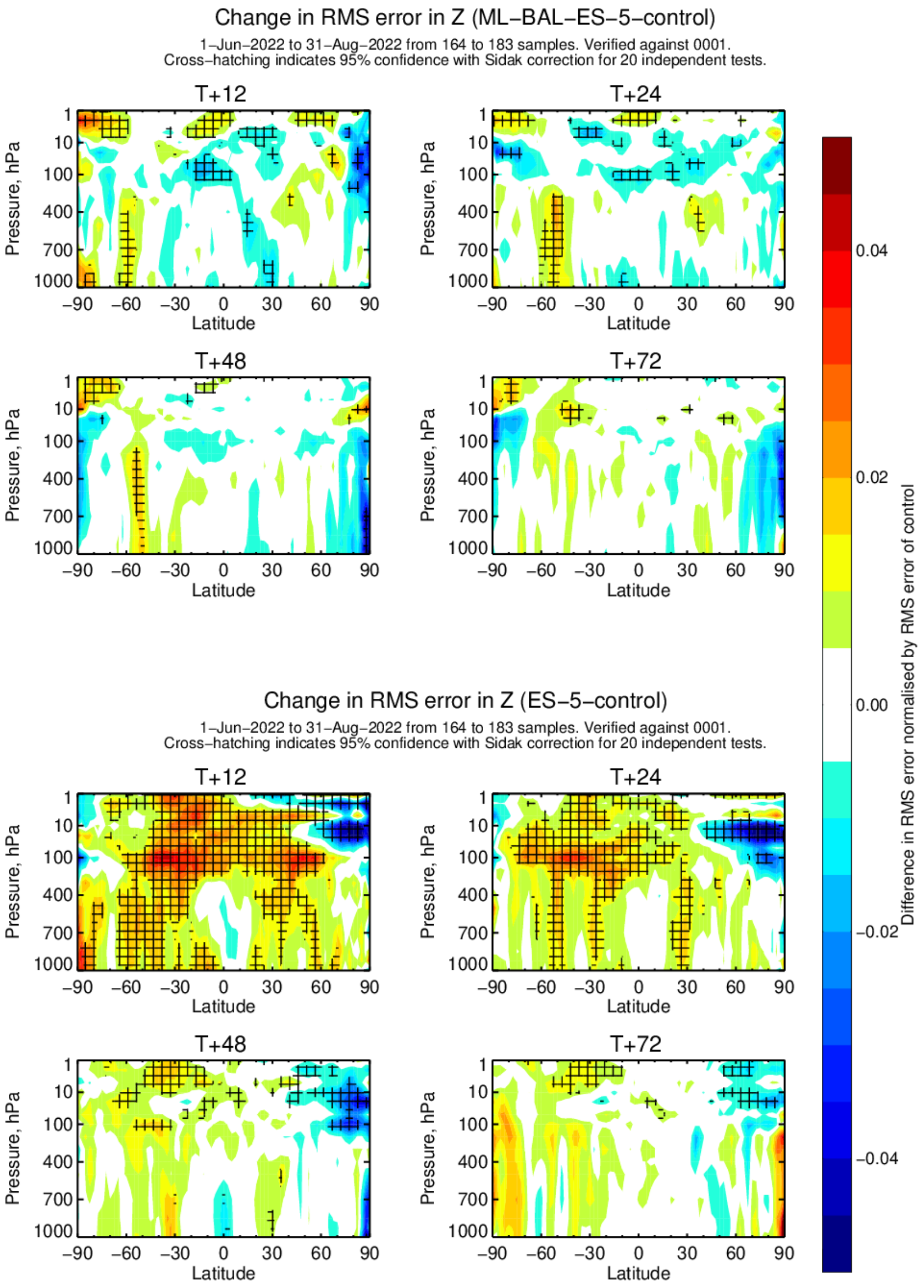


Figure 7: Normalised change in RMS forecast error for geopotential height (Z) in ML-BAL-ES-5 (top) and ES-5 (bottom) relative to the control. Blue (negative) values indicate improvement. Cross-hatching denotes 95% statistical significance.

assessed here, and evaluating its behaviour under such conditions, as well as determining appropriate retraining strategies, remains an important topic for future work.

## Acknowledgements

The work presented in this paper has been produced in the context of the European Union's Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts implementing part of this Initiative with funding by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

The authors would like to thank colleagues at ECMWF for valuable discussions and support throughout this work. In particular, we thank Žiga Zaplotnik, Boštjan Melinc, Patrick Laloyaux, Sebastien Massart, Noureddine Semane, Alban Farchi, Emiliano Orlandi, Ivo Pasmans and Luca Cantarello for insightful discussions. We are grateful to Boštjan Melinc for providing the TensorFlow code base on which this work is based.

We also thank Tony McNally for insightful comments and suggestions that significantly improved the clarity and quality of this manuscript.

## References

Bannister, R. N. (2008). A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, **134**(637), 1951–1970, doi:10.1002/qj.339, URL http://dx.doi.org/10.1002/qj.339.

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

Bocquet, M., Farchi, A. and Malartic, Q. (2021). Online learning of both state and dynamics using ensemble kalman filters. *Foundations of Data Science*, **3**(3), 305–330, doi:10.3934/fods.2020015, URL https://www.aimsciences.org/article/id/72d38af7-58df-41f7-b2de-6b9b9cb9e05a.

Bocquet, M., Finn, T. S., Cheng, S. and Farchi, A. (2026). Elucidating the performance of data assimilation neural networks for chaotic dynamics. *EGUsphere*, **2026**, 1–34, doi:10.5194/egusphere-2026-245, URL https://egusphere.copernicus.org/preprints/2026/egusphere-2026-245/.

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

Bonavita, M., Isaksen, L. and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, **138**(667), 1540–1559, doi:10.1002/qj.1899, URL http://dx.doi.org/10.1002/qj.1899.

Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike (2015). The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **142**(694), 287–303, doi:10.1002/qj.2652, URL http://dx.doi.org/10.1002/qj.2652.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R. and Bengio, S. (2016). Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21.

Capiński, M. and Kopp, P. E. (2004). *Measure, Integral and Probability*. Springer London, ISBN 9781447106456, doi:10.1007/978-1-4471-0645-6, URL http://dx.doi.org/10.1007/978-1-4471-0645-6.

European Centre for Medium-Range Weather Forecasts (ECMWF) (2019). Forecasting system upgrade set to improve global weather forecasts. URL https://www.ecmwf.int/en/about/media-centre/news/2019/forecasting-system-upgrade-set-improve-global-weather-forecasts, accessed: 2026-02-13.

European Centre for Medium-Range Weather Forecasts (ECMWF) (2024). Atlas grid design documentation. URL https://sites.ecmwf.int/docs/atlas/design/grid/, eCMWF Atlas documentation. Accessed: 2026-02-13.

Geer, A. J. (2016). Significance of changes in medium-range forecast scores. *Tellus A: Dynamic Meteorology and Oceanography*, **68**(1), 30229, doi:10.3402/tellusa.v68.30229, URL https://doi.org/10.3402/tellusa.v68.30229.

Hamill, T. M., Whitaker, J. S. and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, **129**(11), 2776–2790.

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A. and Palmer, T. (2021). Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks. *Journal of Advances in Modeling Earth Systems*, **13**(9), e2021MS002521, doi:https://doi.org/10.1029/2021MS002521, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002521, e2021MS002521 2021MS002521.

Isaksen, L. and Bonavita, Massimo and Buizza, Roberto and Fisher, Mike and Haseler, J. and Leutbecher, Martin and Raynaud, Laure (2010). Ensemble of data assimilations at ECMWF. doi:10.21957/OBKE4K60, URL https://www.ecmwf.int/node/10125.

Kaparakis, C. and Mehrkanoon, S. (2023). WF-UNet: Weather Data Fusion using 3D-UNet for Precipitation Nowcasting. *Procedia Computer Science*, **222**, 223–232, doi:https://doi.org/10.1016/j.procs.2023.08.160, URL https://www.sciencedirect.com/science/article/pii/S1877050923009250, international Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023).

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K. and Anandkumar, A. (2023). FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23, Association for Computing Machinery, New York, NY, USA, ISBN 9798400701900, doi:10.1145/3592979.3593412, URL https://doi.org/10.1145/3592979.3593412.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S. and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, **382**(6677), 1416–1421, doi:10.1126/science.adi2336, URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F. and Rabier, F. (2024). AIFS – ECMWF's data-driven forecasting system. doi:10.48550/ARXIV.2406.01465, URL https://arxiv.org/abs/2406.01465.

Lang, S., Alexe, M., Clare, M. C. A., Roberts, C., Adewoyin, R., Ben Bouallègue, Z., Chantry, M., Dramsch, J., Dueben, P. D., Hahner, S., Maciel, P., Prieto-Nemesio, A., O'Brien, C., Pinault, F., Polster, J., Raoult, B., Tietsche, S. and Leutbecher, M. (2026). AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, **2**(1), doi:10.1038/s44387-026-00073-7, URL http://dx.doi.org/10.1038/s44387-026-00073-7.

Lang, S. T. K., Lock, S.-J., Leutbecher, M., Bechtold, P. and Forbes, R. M. (2021). Revision of the Stochastically Perturbed Parametrisations model uncertainty scheme in the Integrated Forecasting System. *Quarterly Journal of the Royal Meteorological Society*, **147**(735), 1364–1381, doi: https://doi.org/10.1002/qj.3978, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3978.

Lang, Annika and Schwab, Christoph (2015). Isotropic Gaussian random fields on the sphere: Regularity, fast simulation and stochastic partial differential equations. *Ann. Appl. Probab.*, **25**(6), 3047–3094.

Lin, H., Tang, J., Wang, S., Wang, S. and Dong, G. (2023). Deep learning downscaled high-resolution daily near surface meteorological datasets over East Asia. *Scientific Data*, **10**(1), doi:10.1038/s41597-023-02805-9, URL http://dx.doi.org/10.1038/s41597-023-02805-9.

Lucas, J., Tucker, G., Grosse, R. and Norouzi, M. (2019). Don't blame the ELBO! A linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems*.

Melinc, B., Perkan, U. and Zaplotnik, Ž. (2026). A Unified Neural Background-Error Covariance Model for Midlatitude and Tropical Atmospheric Data Assimilation. *Journal of Advances in Modeling Earth Systems*, **18**(1), doi:10.1029/2025ms005360, URL http://dx.doi.org/10.1029/2025MS005360.

Melinc, B. and Zaplotnik, Ž. (2024). 3D-Var data assimilation using a variational autoencoder. *Q. J. R. Meteorol. Soc.*, **150**(761), 2273–2295.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, **146**(11), 3885–3900, doi:10.1175/MWR-D-18-0187.1.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Springer, doi:10.1007/978-3-319-24574-4_28.

Ryu, J., Kim, H., Wang, S.-Y. S. and Yoon, J.-H. (2026). Increasing resolution and accuracy in sub-seasonal forecasting through 3D U-Net: the western US. *Geoscientific Model Development*, **19**(1), 27–39, doi:10.5194/gmd-19-27-2026, URL http://dx.doi.org/10.5194/gmd-19-27-2026.

Sohn, K., Yan, X. and Lee, H. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, Vol. 28.

# A ML model

## A.1 Generic formulation

We model the parametric conditional distribution (14) as a *mapping* in the sense of a nonlinear probabilistic regressor,

$$\hat{\sigma}_M \longmapsto \tilde{\sigma}, \qquad \tilde{\sigma} \sim \mathrm{p}_\theta(\cdot|\hat{\sigma}_M). \tag{20}$$

A *conditional variational autoencoder* (CVAE) ((Kingma and Welling, 2013; Sohn *et al.*, 2015)) provides a convenient and flexible model class to represent such conditional distributions.

Let $N_\mathbf{z}$ denote the dimension of a latent space. We choose U-Net convolutional neural network ((Ronneberger *et al.*, 2015; Lin *et al.*, 2023; Kaparakis and Mehrkanoon, 2023; Ryu *et al.*, 2026)) encoder and decoder functions, denoted by $f_{\theta_1} : \mathbb{R}^{N_\mathbf{x}\times d} \mapsto \mathbb{R}^{N_\mathbf{z}\times d}$ and $g_{\theta_2} : \mathbb{R}^{N_\mathbf{z}\times d} \times \mathbb{R}^{N_\mathbf{z}\times d} \mapsto \mathbb{R}^{N_\mathbf{x}\times d}$ respectively. This choice is sensible for field-valued regression because U-Nets (i) capture *multiscale* structure via the encoder–decoder hierarchy (coarse features are extracted in the downsampling path and refined in the upsampling path), (ii) preserve *high-frequency* and spatially localized information through skip connections, which is crucial when $\tilde{\sigma}$ contains sharp gradients or localized variability, and (iii) provide an efficient inductive bias for *translation-equivariant* spatial data through convolutions, leading to good sample efficiency and scalability to high-dimensional grids. Moreover, the symmetric encoder–decoder structure aligns naturally with the CVAE: the encoder $f_{\theta_1}$ constructs a compact latent representation conditioned on $\hat{\sigma}_M$, while the decoder $g_{\theta_2}$ combines this representation with latent randomness to generate coherent samples in physical space, (Melinc and Zaplotnik, 2024).

Thus, we have as the CVAE sampler,

$$\tilde{\sigma} = g_{\theta_2}\big(f_{\theta_1}(\hat{\sigma}_M), \varepsilon\big), \quad \varepsilon \sim \mathrm{q}_{\theta_3}(\cdot|\hat{\sigma}_M) \tag{21}$$

where $\varepsilon$ denotes a latent space random field whose distribution is $\mathrm{q}_{\theta_3}(\cdot|\hat{\sigma}_M)$. Parameters $\theta_1$, $\theta_2$ and $\theta_3$ constitute the parameters of $\mathrm{p}_\theta$. We interpret $\varepsilon$ as the "regression error" term, whose distribution is $\mathrm{q}_{\theta_3}$.

To sample from $\mathrm{p}_\theta(\cdot|\hat{\sigma}_M)$, we generate a sample of $\varepsilon$ from the distribution $\mathrm{q}_{\theta_3}(\cdot|\hat{\sigma}_M)$ and evaluate (21).

**Loss function.** To solve (15), we optimise a surrogate loss—the conditional evidence lower bound (ELBO),

$$\mathscr{L}(\theta;\hat{\sigma}_M) := \mathrm{D}^{\mathrm{KL}}\big(\mathrm{q}_\theta(\mathbf{z} \mid \hat{\sigma}_M) \,\|\, \mathrm{p}(\mathbf{z} \mid \hat{\sigma}_M)\big) - \mathbb{E}_{\mathbf{z}\sim \mathrm{q}_\theta}\big(\log \mathrm{p}_\theta(\tilde{\sigma} \mid \mathbf{z}, \hat{\sigma}_M)\big), \tag{22}$$

where $\mathrm{D}^{\mathrm{KL}}$ means Kullback-Leibler (KL) divergence, $\mathrm{p}(\mathbf{z} \mid \hat{\sigma}_M)$ is a chosen prior distribution on the latent space and $\mathrm{q}_\theta(\mathbf{z} \mid \hat{\sigma}_M)$ is a parameterised variational distribution, cf. (21). See Appendix A.2 for discussion of our implementation.

Moreover,

$$\log \mathrm{p}_\theta(\tilde{\sigma} \mid \hat{\sigma}_M) \geq -\mathscr{L}(\theta;\hat{\sigma}_M), \tag{23}$$

which follows from Bayes' theorem, Jensen's inequality, and the introduction of the variational distribution $\mathrm{q}_\theta(\mathbf{z} \mid \hat{\sigma}_M)$. Thus, we minimise (22) to determine $\theta^*$,

$$\theta^* = \arg\min_\theta \mathscr{L}(\theta;\hat{\sigma}_M). \tag{24}$$

## A.2 Model architecture, loss function implementation and training

The generic probabilistic formulation of the ML model is described in Appendix A.1. Here we describe the specific neural network architecture, the implemented loss function, and the training procedure used in this study.

Following exploratory experiments (not reported), we trained six conditional variational autoencoder (CVAE) models, one for each IFS model variable considered, all sharing the same architecture and hyperparameters. This uniform choice is unlikely to be optimal, given the differing physical and statistical characteristics of the variables. Our goal in this work is to establish a baseline performance using a common modelling framework, which can be refined in future studies through variable-specific model design and training.

At each vertical level, the model maps a small-ensemble estimate of the standard deviation to an estimate of the corresponding full-ensemble standard deviation on the same latitude–longitude grid. A single network architecture is therefore applied uniformly across all 137 model levels, while training is performed independently for each variable.

**Network architecture.** Our CVAE adopts a compact U-Net-style encoder–decoder architecture with a stochastic latent bottleneck. We adopt standard convolutional architectures applied to the full latitude–longitude grid rather than spherical convolutional layers, in order to establish a baseline. The potential benefits of spherical architectures are left for future investigation.

For a given model level, the input is a single-channel field

$$\hat{\sigma}_{M,i} \in \mathbb{R}^{160\times 320\times 1},$$

defined on the $F80$ latitude–longitude grid, and the output is a single-channel reconstruction

$$\tilde{\sigma}_i \in \mathbb{R}^{160\times 320\times 1}.$$

All convolutional layers use $3\times 3$ kernels with ELU activations and `valid` padding. To ensure consistency across the date line and near the poles, each convolution is preceded by a custom periodic padding layer that enforces longitudinal periodicity and appropriate latitudinal wrapping. This allows the use of standard convolution operators while avoiding boundary artefacts on the global grid.

The encoder consists of three convolutional blocks with increasing channel depth (8, 16, and 32 channels), interleaved with max-pooling operations that progressively reduce the spatial resolution from $160\times 320$ to $20\times 40$. A final convolution maps the encoder output to a single-channel feature map, which is flattened and projected into a latent space of dimension $d_z = 128$.

**Latent space and prior.** The latent variable $\mathbf{z} \in \mathbb{R}^{128}$ is endowed with a standard normal prior,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{25}$$

which is independent of the conditioning input. The encoder defines a diagonal Gaussian variational posterior,

$$q_\theta(\mathbf{z} \mid \hat{\sigma}_{M,i}) = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\hat{\sigma}_{M,i}), \mathrm{diag}\left(\exp(\log \boldsymbol{\sigma}^2_\theta(\hat{\sigma}_{M,i}))\right)\right), \tag{26}$$

where the mean and log-variance fields are produced by dense layers applied to the flattened encoder output. Sampling from the latent space is performed using the reparameterisation trick.

We did not explicitly investigate posterior collapse ((Bowman *et al.*, 2016; Lucas *et al.*, 2019)) in this work, although indications consistent with partial posterior collapse were observed. This suggests that the conditional nonlinear mapping was sufficiently expressive to capture the variance structure with limited reliance on the latent variables. Since reconstruction accuracy was the primary objective, we focused on evaluating the impact of the resulting ML-emulated variance within the data assimilation system. A more detailed investigation of this potential issue is left for future work.

**Decoder.** The decoder maps the latent variable back to grid-point space via a dense projection and reshape operation, followed by an expanding path with transposed convolutions. Skip connections are employed between encoder and decoder feature maps at matching resolutions, yielding a U-Net-style architecture that preserves fine-scale spatial information. The final decoder layer is a $1 \times 1$ convolution with ELU activation, producing the reconstructed standard deviation field.

**ELBO implementation.** The model is trained by maximising the evidence lower bound (ELBO). For a single model level, the ELBO takes the form

$$\mathscr{L}(\theta) = \mathbb{E}_{q_\theta(\mathbf{z}|\hat{\sigma}_{M,i})}\left[\log p_\theta(\tilde{\sigma}_i \mid \hat{\sigma}_{M,i}, \mathbf{z})\right] - \mathrm{KL}(q_\theta(\mathbf{z} \mid \hat{\sigma}_{M,i}) \,\|\, p(\mathbf{z}))\,. \tag{27}$$

In practice, following (Melinc and Zaplotnik, 2024), the reconstruction term is implemented using a spatially averaged Huber loss between the decoder output and the target full-ensemble standard deviation field. This choice provides robustness to localised extremes while remaining differentiable and computationally efficient. The KL divergence term is evaluated analytically for diagonal Gaussian distributions and estimated via Monte Carlo sampling using the reparameterised latent variable.

To account for all 137 model levels, we assume that the conditional distribution of the full-ensemble standard deviations factorises across vertical levels. Thus the ELBO is evaluated independently at each level and summed,

$$\log p_\theta(\tilde{\sigma} \mid \hat{\sigma}_M) \;\geq\; -\sum_{i=1}^{137} \mathscr{L}(\theta; \hat{\sigma}_{M,i}). \tag{28}$$

This modelling choice is motivated by both empirical evidence and the structure of the ensemble variance estimator.

From a practical perspective, exploratory experiments (not reported) showed that a model trained on data from a single vertical level generalises well when applied to other levels, with only a slight but systematic degradation in performance. This suggests that the dominant factors governing the mapping from sub-ensemble to full-ensemble variance are local spatial structure and finite-sample effects, while explicit vertical coupling plays a secondary role for the diagonal variance component.

From a statistical perspective, this assumption is consistent with the definition of the ensemble variance estimator (8), which is computed pointwise in physical space and independently at each model level. For a fixed latitude–longitude location, the sample variance at a given level depends only on ensemble values at that level and does not explicitly involve neighbouring vertical levels. Vertical correlations in the model state are instead represented through the correlation structure of **B**, rather than through its diagonal variance component.

Under this assumption, the conditional distribution $\mathrm{p}_\theta(\tilde{\sigma} \mid \hat{\sigma}_M)$ factorises as a product measure over model levels. This enables the use of a single shared network architecture across all levels, allowing the latent space dimension to remain small and the overall model capacity to be tightly controlled. The resulting models are lightweight, with fewer than $2 \times 10^5$ trainable parameters, and can therefore be evaluated efficiently (even on CPUs) within cycled data assimilation experiments.

Training is performed using the Adam optimiser with mini-batches, and gradients are computed by automatic differentiation. Each mini-batch contains samples for all 137 model levels, and the objective is accumulated over levels via a vectorised mapping of the per-level loss.

At each optimisation step, the trainable parameters are updated by minimising the negative ELBO objective,

$$\mathscr{J}(\theta) = -\log p_\theta\big(\tilde{\sigma} \mid \hat{\sigma}_M, \mathbf{z}\big) - \log p(\mathbf{z}) + \log q_\theta(\mathbf{z} \mid \hat{\sigma}_M), \tag{29}$$

where $\mathbf{z}$ denotes the latent variable introduced by the CVAE. In practice, the conditional likelihood $p_\theta(\tilde{\sigma} \mid \hat{\sigma}_M, \mathbf{z})$ is implemented implicitly via the decoder mapping (21), and the negative log-likelihood term is approximated by a spatially averaged Huber loss between the decoder output and the target full-ensemble standard deviation field. The latent regularisation terms correspond to a standard normal prior $p(\mathbf{z})$ and a diagonal Gaussian variational posterior $q_\theta(\mathbf{z} \mid \hat{\sigma}_M)$, evaluated using the reparameterisation trick.

The learning rate is initialised at $10^{-4}$ and subsequently adapted based on the evolution of the validation loss. Specifically, if the validation loss exhibits deterioration relative to recent epochs, the learning rate is reduced by a fixed multiplicative factor, subject to a lower bound of $10^{-5}$, c.f. Melinc and Zaplotnik (2024).

Gaussian–Legendre quadrature weights are incorporated into the spatial norms used in the loss function to ensure that the objective consistently approximates integrals over the sphere and avoids over-representation of high-latitude grid points.

### A.3 Description of the training dataset

For the training and validation data sets, we retrieved 30 background ensemble members[6] together with the archived full-ensemble standard deviation estimates from ECMWF's operational data archive, covering the period from 1 January 2023 to 31 December 2023. Since the ECMWF operational data assimilation cycle produces analyses twice daily, this resulted in a total of 730 temporal samples.

This period spans two IFS cycles (47r1 and 48r1). However, no major system updates occurred during this time, and the data are therefore expected to be broadly homogeneous in a statistical sense across the entire data set.

The variance component of **B** is estimated for six IFS model variables: the logarithm of surface pressure (lnsp), temperature (t), vorticity (vo), divergence (div), specific humidity (q), and ozone (o3). For t, vo, div, q, and o3, each data field consists of 137 discrete vertical model levels; that is, at each latitude–longitude grid point, 137 values are defined. For lnsp, only a single vertical level is present. The archived ensemble standard deviation fields are provided in grid-point space. Consequently, we retrieved the ensemble members in grid-point format at the same grid resolution as the archived ensemble standard deviations.

For ensemble data fields archived in spectral representation, transformation to grid-point space requires application of an inverse Legendre transform followed by an inverse Fourier transform. These transformations were performed automatically during the data retrieval process using default ECMWF numerical packages. The resulting grid-point fields are defined on reduced Gaussian grids (European Centre for Medium-Range Weather Forecasts (ECMWF), 2024).

For practical reasons related to the ML architecture (see Appendix A.2), we chose to work with regular latitude–longitude grids. An additional interpolation step was therefore required to project the reduced Gaussian grid data onto regular latitude–longitude grids. This interpolation was again carried out automatically using default ECMWF numerical packages during data retrieval.

The reduced Gaussian grid associated with the archived ensemble standard deviation data is $N80$, corresponding to a horizontal resolution of $1.125^\circ$, or approximately 125 km at the equator (European Centre for Medium-Range Weather Forecasts (ECMWF), 2024), and comprising 35,718 grid points. We interpolate the $N80$ data to a regular latitude–longitude grid of size $160 \times 320$, corresponding to 51,200 grid points. At ECMWF, this regular grid is referred to as $F80$. All data fields used for ML training are therefore expressed on the $F80$ grid.

To construct a sufficiently large data set (13) for training the ML model, we applied combinatorial sub-

[6] Due to local disk storage constraints and to reduce data retrieval time, we retrieved 30 ensemble members rather than the full set of 50.

sampling to the retrieved 30 ensemble members. However, unconstrained subsampling of 5 members from a pool of 30 introduces unwanted colinearity in the data. For example, the index sets $[1,2,3,4,5]$ and $[1,2,3,4,6]$ differ by only a single member. Without further constraints, the resulting training distribution may therefore inadequately represent the sample space of the target conditional distribution (14).

To mitigate this issue, we introduced constraints based on the cardinality of the intersection between any two subsampled index sets. Let # denote set cardinality. For example,

$$\#\big([3,4,5,7,9] \cap [4,7,8,10,15]\big) = \#[4,7] = 2, \tag{30}$$

corresponding to an overlap of two members. Based on this criterion, we imposed the constraint that, when subsampling 5 members from 30, no two index sets may share more than two common members.

The number of admissible combinations under these constraints is not unique. For subsampling 5 members from 30 with a maximum overlap of two, the total number of combinations is *approximately* 229, compared with 142,506 combinations in the unconstrained case.

As a result, for the 5-member setup the data set could be enlarged by a factor of 229. We randomly shuffled the enlarged dataset and retained the first 50 subsamples. These subsamples were then used to evaluate $\hat{\sigma}_{M,k}$ in (13) for $M = 5$, while the supervised learning targets $\hat{\sigma}_{N,k}$ correspond to the archived operational full-ensemble estimates with $N = 50$.

In summary, for each of the six model variables, the final data set consists of $730 \times 50 = 36{,}500$ supervised learning pairs,

$$\big(\hat{\sigma}_{M,k,e}, \hat{\sigma}_{N,k}\big), \quad k = 1,\ldots,730, \quad e = 1,\ldots,50, \tag{31}$$

where the additional index $e$ denotes the subsampled ensemble member set. For t, vo, div, q, and o3, both $\hat{\sigma}_{M,k,e}$ and $\hat{\sigma}_{N,k}$ have dimensions $137 \times 160 \times 320$, while for lnsp they have dimensions $1 \times 160 \times 320$.

For normalisation, we applied a standard $z$-transform to the data in (31), using the pointwise mean and standard deviation of $\hat{\sigma}_{N,k}$ computed over the time index $k$. The resulting normalised data set was then split into training and validation subsets using an 80:20 ratio. For the out-of-sample evaluations in Section 4.2, the z-transform was performed using the same mean and standard deviation fields computed from the training dataset, ensuring consistency with the normalisation applied during training.

### *A.3.1 Balanced and unbalanced control variables*

In the ECMWF hybrid 4D-Var system, the control vector is constructed so that key dynamical balance relationships are represented explicitly through linear balance operators. In particular, vorticity and ozone are treated as primary variables, while divergence, temperature, and surface pressure are decomposed into balanced and unbalanced components. The balanced component is diagnosed from regressions on vorticity and other balanced modes, whereas the unbalanced component represents departures from these balance relationships and captures variability not explained by large-scale dynamical constraints (Bonavita, Massimo and Hólm, Elías and Isaksen, Lars and Fisher, Mike, 2015). This formulation allows the background-error covariance matrix to be represented more compactly and realistically in the transformed control space.

In this work, the training targets consist of ensemble variance fields for both balanced variables (vorticity, specific humidity, and ozone) and unbalanced variables (unbalanced temperature, divergence, and surface pressure). However, the input ensemble variance fields used during training were computed exclusively from balanced ensemble members for all variables. As a result, the trained models are expected to operate on inputs that are consistent with this balanced control-variable definition.

In practice, the operational ensemble data assimilation workflow naturally provides variance estimates derived from the full control vector, including unbalanced components. These variance estimates therefore differ in statistical structure from the balanced inputs used during training. To assess the sensitivity of the ML emulator to this distinction, we evaluate configurations using both balanced and unbalanced small-ensemble variance inputs. This comparison provides a practical test of the robustness of the trained models to deviations from the training-consistent input representation.

## B Impact of approximate covariance on 4D-Var analysis

**Proposition 1.** *Assume the model operator $\mathscr{M}_{0\to k}$ and the observation operator $\mathscr{H}_k$ are $\mathscr{C}^2$ for all $k = 0,\dots,L$. Then the strong-constraint 4D-Var cost function* (4) *is locally Lipschitz continuous with respect to* **B**.

*Proof.* Let $\tilde{\mathbf{B}}_\varepsilon$ be an approximation of **B**. Let $\delta\mathbf{x} := \mathbf{x}_0 - \mathbf{x}^b$ denote the optimiser of (4). And similarly let $\delta\tilde{\mathbf{x}}_\varepsilon$ be the optimiser of (4) but using $\tilde{\mathbf{B}}_\varepsilon$ for the background covariance. Then, we have

$$\mathbf{B}^{-1}\delta\mathbf{x} + \nabla\Psi(\delta\mathbf{x}) = 0, \quad \text{and} \quad \tilde{\mathbf{B}}_\varepsilon^{-1}\delta\tilde{\mathbf{x}}_\varepsilon + \nabla\Psi(\delta\tilde{\mathbf{x}}_\varepsilon) = 0, \tag{32}$$

where

$$\Psi(\delta\mathbf{x}) := \frac{1}{2}\sum_{k=0}^{L}\left(\mathscr{H}_k(\mathscr{M}_{0\to k}(\delta\mathbf{x}+\mathbf{x}^b)) - \mathbf{y}_k\right)^T \mathbf{R}^{-1}\left(\mathscr{H}_k(\mathscr{M}_{0\to k}(\delta\mathbf{x}+\mathbf{x}^b)) - \mathbf{y}_k\right). \tag{33}$$

Combining (32) to obtain

$$\mathbf{B}^{-1}\delta\mathbf{x} - \tilde{\mathbf{B}}_\varepsilon^{-1}\delta\tilde{\mathbf{x}}_\varepsilon + \nabla\Psi(\delta\mathbf{x}) - \nabla\Psi(\delta\tilde{\mathbf{x}}_\varepsilon) = 0. \tag{34}$$

By the Fundamental Theorem of Calculus

$$\nabla\Psi(\delta\mathbf{x}) - \nabla\Psi(\delta\tilde{\mathbf{x}}_\varepsilon) = \underbrace{\int_0^1 (D^2\Psi)(t\delta\mathbf{x} + (1-t)\delta\tilde{\mathbf{x}}_\varepsilon)dt}_{\mathbf{A}:=}\,[\delta\mathbf{x} - \delta\tilde{\mathbf{x}}_\varepsilon], \tag{35}$$

Hence,

$$(\delta\tilde{\mathbf{x}}_\varepsilon - \delta\mathbf{x}) = \left(\tilde{\mathbf{B}}_\varepsilon^{-1} + \mathbf{A}\right)^{-1}\left(\mathbf{B}^{-1} - \tilde{\mathbf{B}}_\varepsilon^{-1}\right)\delta\mathbf{x} \tag{36}$$

from which the result follows. □

In light of Proposition 1, if the ML produced approximation is close to the full ensemble estimate of **B**, we can expect the resulting 4D-Var analyses to be correspondingly close. In particular, if $\tilde{\mathbf{B}}_\varepsilon$ is a perturbation of **B**,

$$\tilde{\mathbf{B}}_\varepsilon := \mathbf{B} + \varepsilon\delta\tilde{\mathbf{B}} \tag{37}$$

for some perturbation matrix $\delta\tilde{\mathbf{B}}$, then by Taylor expansion, from (36) we have

$$\begin{aligned}(\delta\tilde{\mathbf{x}}_\varepsilon - \delta\mathbf{x}) = \varepsilon\,\big(&\mathbf{B}^{-1} - \varepsilon\,\mathbf{B}^{-1}\delta\mathbf{B}\,\mathbf{B}^{-1}\\ &+ O(\varepsilon^2) + \mathbf{A}\big)^{-1}\left(\mathbf{B}^{-1}\delta\mathbf{B}\,\mathbf{B}^{-1} + O(\varepsilon)\right)\delta\mathbf{x} \longrightarrow 0\end{aligned} \tag{38}$$

as $\varepsilon \to 0$. However, how close any estimator can get to $\hat{\sigma}_N$ given only $\hat{\sigma}_M$ is constrained by information-theoretic limits.

On the last point, we can take a measure-theoretic view. Given $\hat{\sigma}_M$, the *optimal* estimator for $\hat{\sigma}_N$ with respect to the mean square error is the orthogonal projection of $\hat{\sigma}_N$ onto the space of $\mathscr{F}_M$ measurable random variables, where $\mathscr{F}_M$ is the sigma-algebra generated by information from the $M$ samples, see (Capiński and Kopp, 2004). This orthogonal projection is given by $\mathbb{E}(\hat{\sigma}_N|\mathscr{F}_M)$. Therefore, using the conditional mean of our generative model (16) as an estimator for $\hat{\sigma}_M$, component-wise and pointwise we have

$$\mathbb{E}\big[\,(\bar{\sigma}-\hat{\sigma}_N)^2_{i,(\lambda,\phi)}\,\big] \geq \mathbb{E}\big[\,(\mathbb{E}(\hat{\sigma}_N|\mathscr{F}_M)-\hat{\sigma}_N)^2_{i,(\lambda,\phi)}\,\big], \quad i=1,\ldots,d, \text{ and } (\lambda,\phi)\in\mathbb{S}^2 \tag{39}$$

where $\bar{\sigma}$ is from (16). The lower bound in (39) is *strictly positive* for $N>M$.

In other words, there is a non-zero lower bound on the mean squared error of any model that estimates $\hat{\sigma}_N$ from $\hat{\sigma}_M$, due to loss of information, and the fact that even the best estimator $\mathbb{E}(\hat{\sigma}_N|\mathscr{F}_M)$ does not achieve zero mean square error. But the ML training is designed to approximate the optimal orthogonal projection using available data, and even goes beyond it to a degree by enlarging the information set $\mathscr{F}_M$ with training data. Furthermore, since our aim is to train an ML model that targets $\hat{\sigma}_N^2$ rather than the unknown true $\sigma^2$, theoretical results such as (9) and (10) cannot be readily applied to obtain explicit uncertainty bounds.