Emulating background errors for Data Assimilation with ML

Wei Pan, Ivo Pasmans

Contributors: Massimo Bonavita, Elías Hólm, Sebastien Massart, Marcin Chrust, Patrick Laloyaux, Alban Farchi, Marc Bocquet (ENPC), Noureddine Semane, Žiga Zaplotnik, Boštjan Melinc, Emiliano Orlandi, Luca Cantarello, Simon Lang, Ewan Pinnington, Mohamed Dahoui

ECMWF Annual Seminar, 14-18 September 2026

ECMWF

© ECMWF September 17, 2026

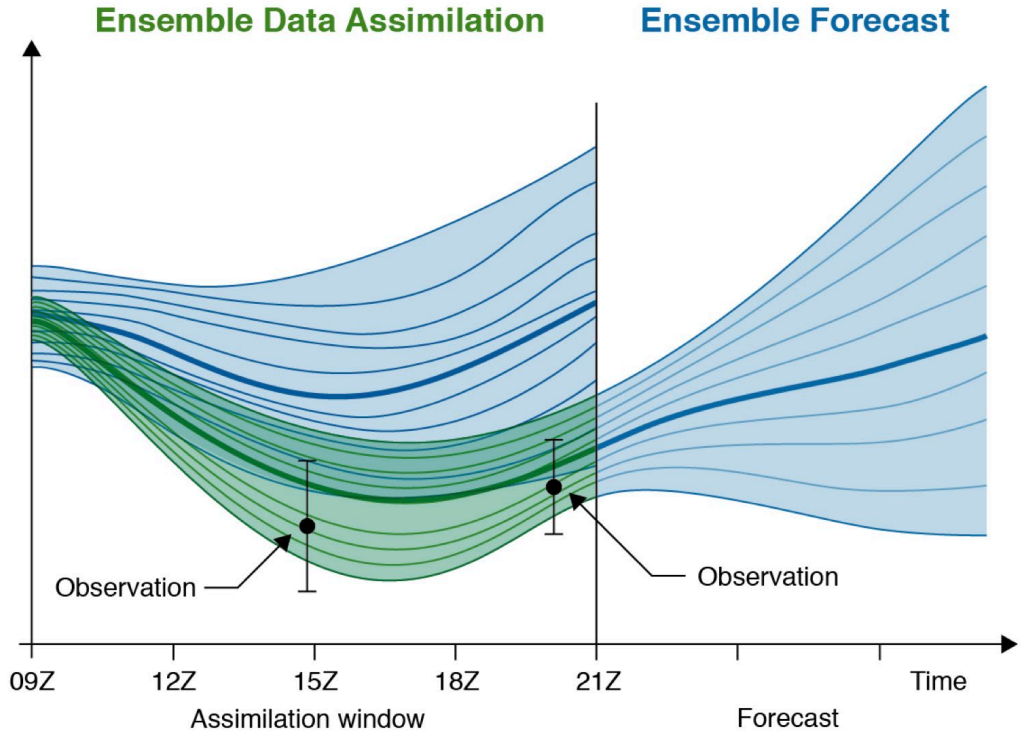
Ensemble 4D-Var

4D-Var in Bayes form:

$$p(\mathbf{x} \mid \mathbf{y}_{\text{obs}}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)\right) p(\mathbf{y}_{\text{obs}} \mid \mathbf{x})$$

Variational DA produces *maximum a posterior* (MAP) estimates

$$\mathbf{x}^a = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}).$$



Ensemble DA (EDA) produces a collection of many perturbed MAP optimisations

$$\mathbf{x}^{a,(i)} = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}^{(i)}, \mathbf{x}^{b,(i)}, \boldsymbol{\eta}^{(i)}), \quad i = 1, \ldots, N.$$

$\boldsymbol{N = 50}$ at ECMWF

As well as providing **ICs** for ensemble forecasting, the EDA provides **flow-dependent estimates of analysis and background error uncertainty** (*Isaksen et al* (2010)),

$$\mathbf{B}(t) \approx \frac{1}{N-1}\sum_{i=1}^{N}[\mathbf{x}^{b,(i)}(t) - \bar{\mathbf{x}}^b(t)]\,[\mathbf{x}^{b,(i)}(t) - \bar{\mathbf{x}}^b(t)]^T.$$

Hybrid 4DVar-ML EDA

Aim – Develop an EDA emulator and integrate it into the operational EDA workflow.

Two complementary directions:

1. EDA statistics
    - Emulate full-EDA statistics from a smaller ensemble.

$$\{\mathbf{x}^{b,1}, \mathbf{x}^{b,2}, \ldots, \mathbf{x}^{b,N}\} \longrightarrow \mathbf{B}$$

approximates

$$\{\mathbf{x}^{b,1}, \mathbf{x}^{b,2}, \ldots, \mathbf{x}^{b,M}\} \longrightarrow \mathrm{ML} \longrightarrow \tilde{\mathbf{B}}, \quad M < N$$

2. EDA members
    - Emulate EDA members that effectively mimic those from the full 4D-Var EDA.
        - Full 4DVar EDA ensemble $\{\mathbf{x}^{a,1}, \mathbf{x}^{a,2}, \ldots, \mathbf{x}^{a,N}\}$
        - Hybrid 4DVar and ML EDA ensemble $\{\mathbf{x}^{a,1}, \mathbf{x}^{a,2}, \ldots, \mathbf{x}^{a,M}\} \cup \{\tilde{\mathbf{x}}^{a,j_1}, \tilde{\mathbf{x}}^{a,j_2}, \ldots, \tilde{\mathbf{x}}^{a,j_{N-M}}\}$

ML emulated members

Hybrid 4DVar-ML EDA

Progress

1. EDA statistics
   - Developed and evaluated a prototype emulator for operational 50-EDA sampled stDev $\boldsymbol{\Sigma}^{\frac{1}{2}}$ (**es**) , using an input ensemble size of **M=5**.

$$\mathbf{B}(t) = \mathbf{T}^{-1}\,\boldsymbol{\Sigma}^{\frac{1}{2}}(t)\,\mathbf{C}(t)\,\boldsymbol{\Sigma}^{\frac{1}{2}}(t)\,\mathbf{T}^{-T}$$

balance — correlations — variance

2. EDA members
   - We now have a working pipeline (RD version):

     model training --> ML inference --> integration into EDA workflow

   - Prototype emulator targets a 50-member EDA at T255 resolution.
   - The test model emulates 38 members, using a 12-member ensemble produced by 4D-Var for ***guidance***.
   - The emulation target comprises the IFS control variables: vorticity (**vo**), divergence (**div**), temperature (**t**), specific humidity (**q**), ozone (**o3**), and logarithm of surface pressure (**lnsp**), at full vertical resolution.

EDA statistics

EDA variance emulator

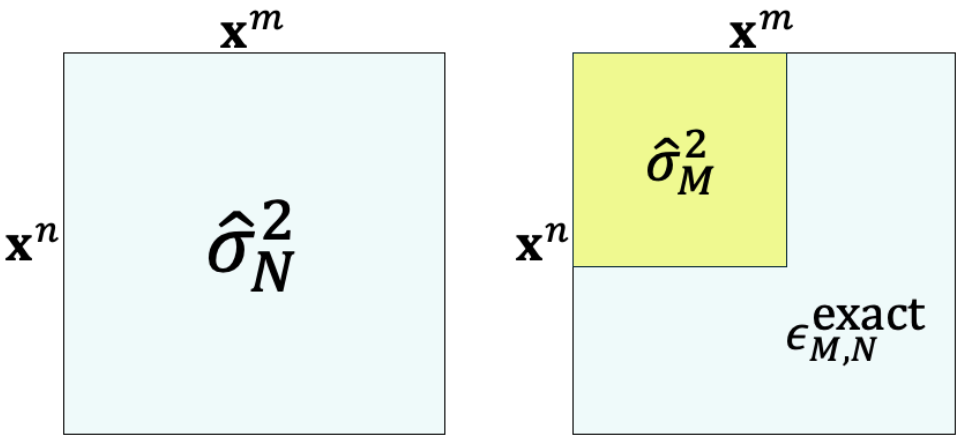
Emulate full-EDA variance from a smaller ensemble *(submitted to QJRMS).*

- The variance **Σ** part of **B** is just the pointwise unbiased sample variance

$$\hat{\sigma}_N^2 := \frac{1}{2N(N-1)} \sum_{n,m=1}^{N} \left(\mathbf{x}^{b,(n)} - \mathbf{x}^{b,(m)}\right)^2$$

- Fix a small ensemble estimate, M < N

$$\hat{\sigma}_N^2 = \alpha_{M,N}\, \hat{\sigma}_M^2 + \epsilon_{M,N}^{\text{exact}},$$



- The proposed model takes as input $\hat{\sigma}_M$. Its output is a probability distribution over $\hat{\sigma}_N$ (a deterministic model could suffice if only a single estimate of $\hat{\sigma}_N$ is required)

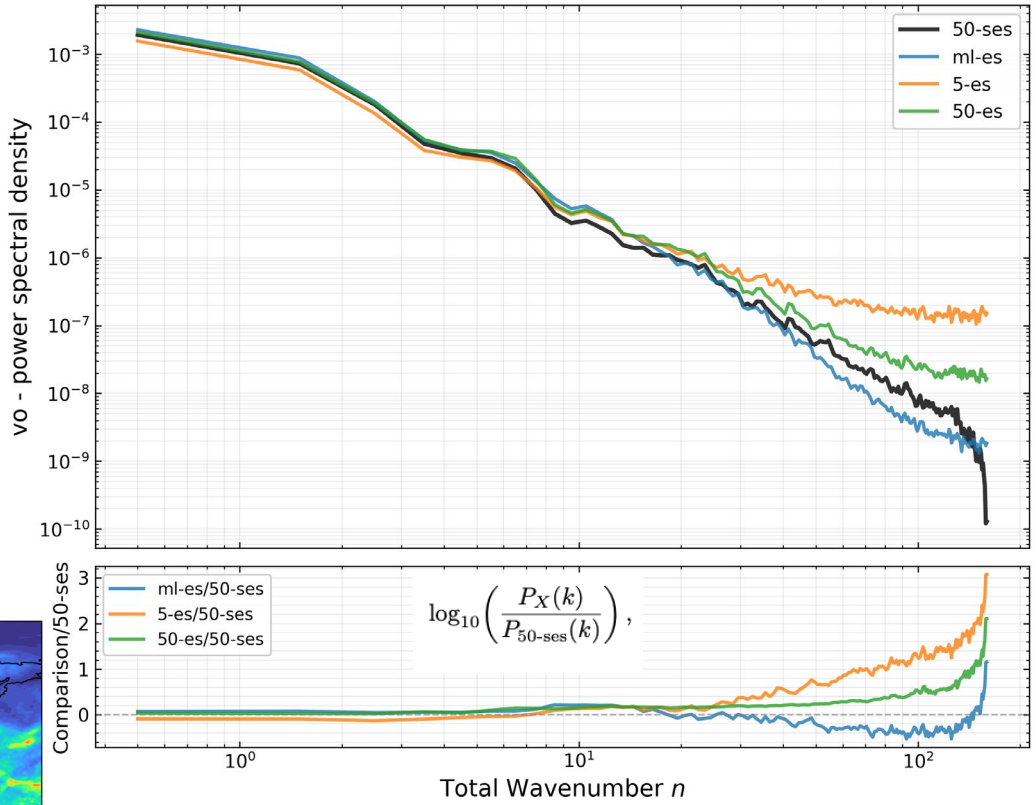
$$\hat{\sigma}_M \mapsto \mathbb{P}_\theta(\sigma_N \mid \hat{\sigma}_M).$$

EDA variance emulator

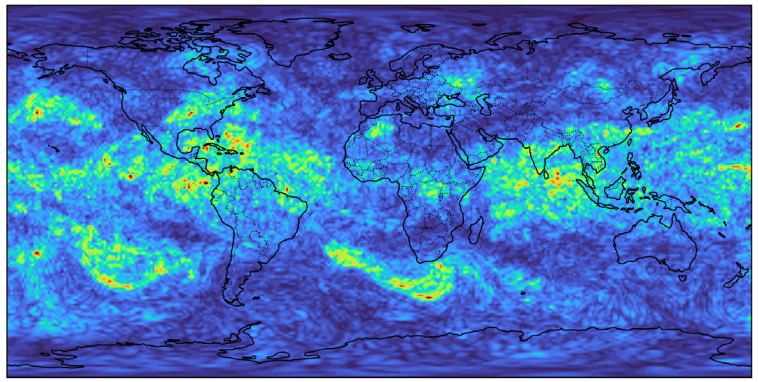
$$\tilde{\sigma} = g_{\theta_2}\left(f_{\theta_1}(\hat{\sigma}_M), \epsilon\right), \quad \epsilon \sim q_{\theta_3}\left(\cdot \mid \hat{\sigma}_M\right)$$

decoder — encoder — latent error distribution

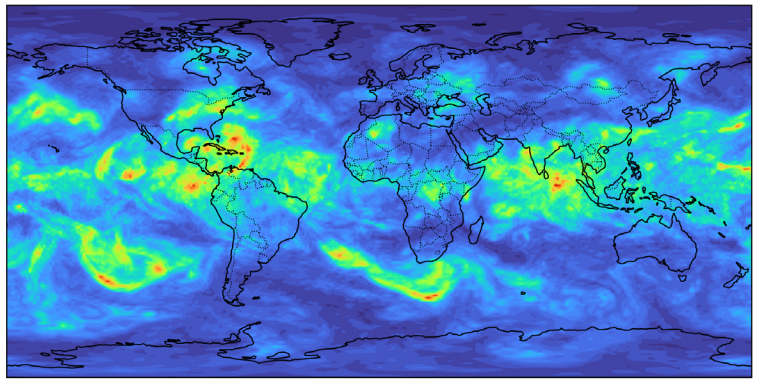
CNN UNet



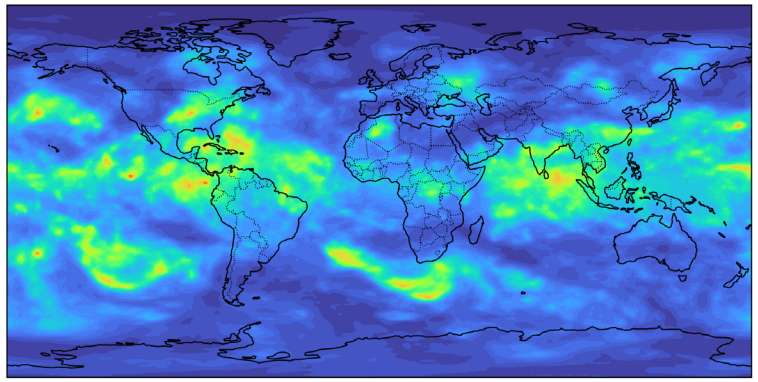
5-es



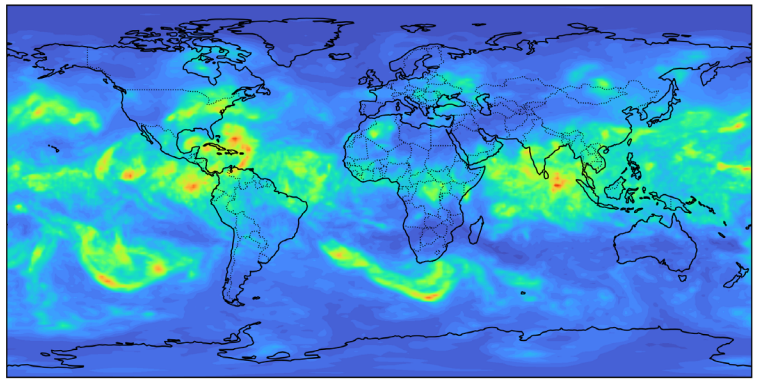
50-es



ml-es



50-ses



M=5, N=50
N80 grid, 137 model levels
6 variables: (vo, div, t, lnsp, q, o3)

The training dataset spans
01/01/2023 – 31/12/2023.

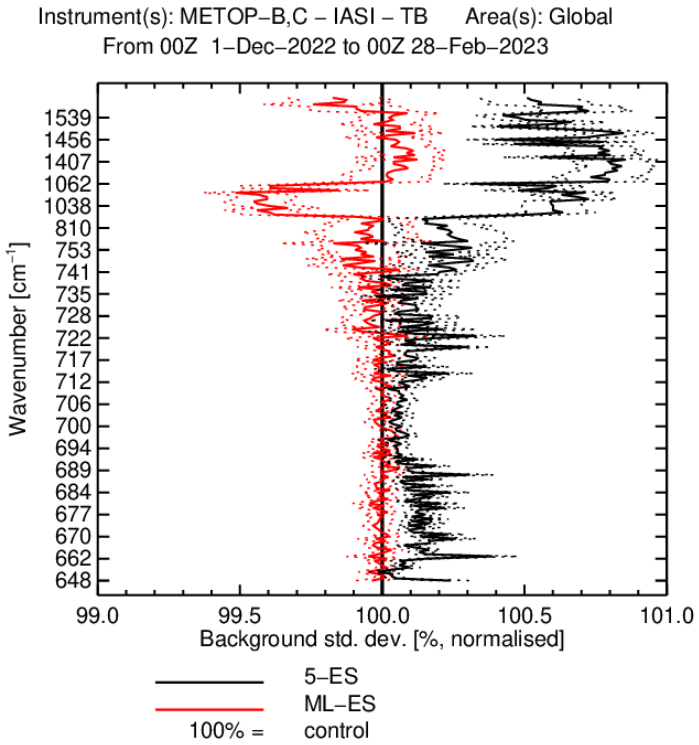
Each model has
< 200,000 learnable parameters

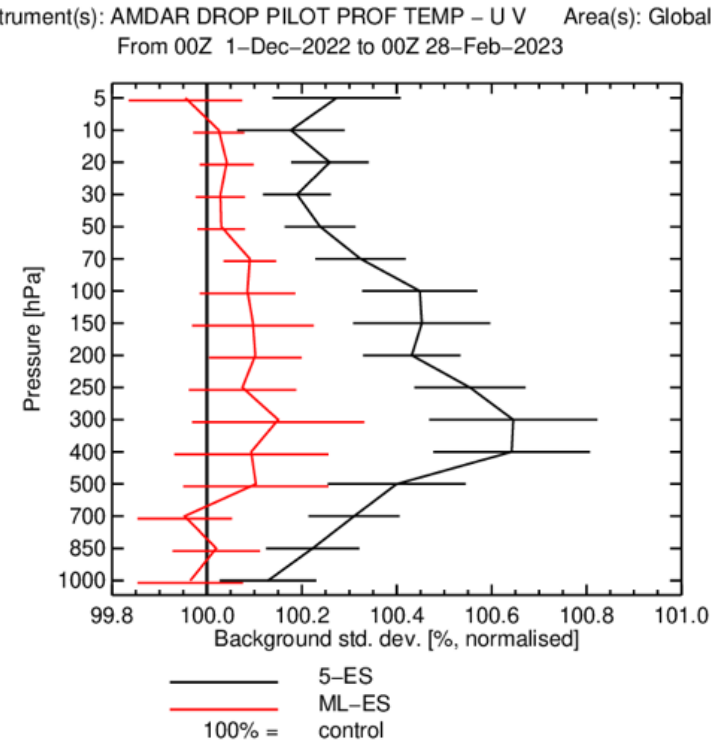
e.g. Vorticity es, ~200hPa
21:00 UTC 01 June 2022

Normalised observation statistics (obstats)

Satellite instruments







Conventional instruments



- Perfect replication is not possible

$$\mathbb{E}\left[\left(\bar{\sigma}-\hat{\sigma}_N\right)^2_{i,(\lambda,\phi)}\right] \geq \mathbb{E}\left[\left(\mathbb{E}(\hat{\sigma}_N|\mathscr{F}_M)-\hat{\sigma}_N\right)^2_{i,(\lambda,\phi)}\right],$$

any estimator from M samples

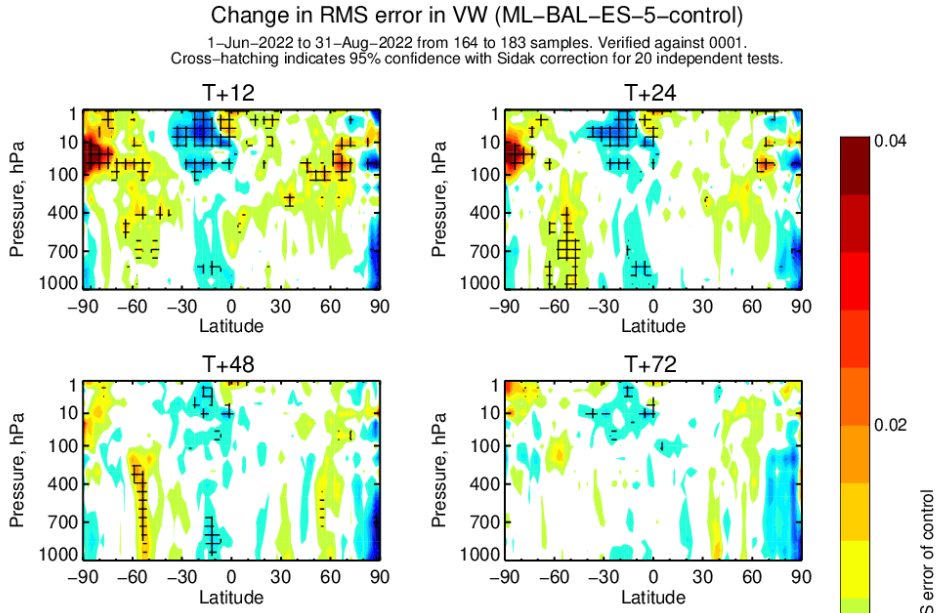
Optimal with respect to mean square error

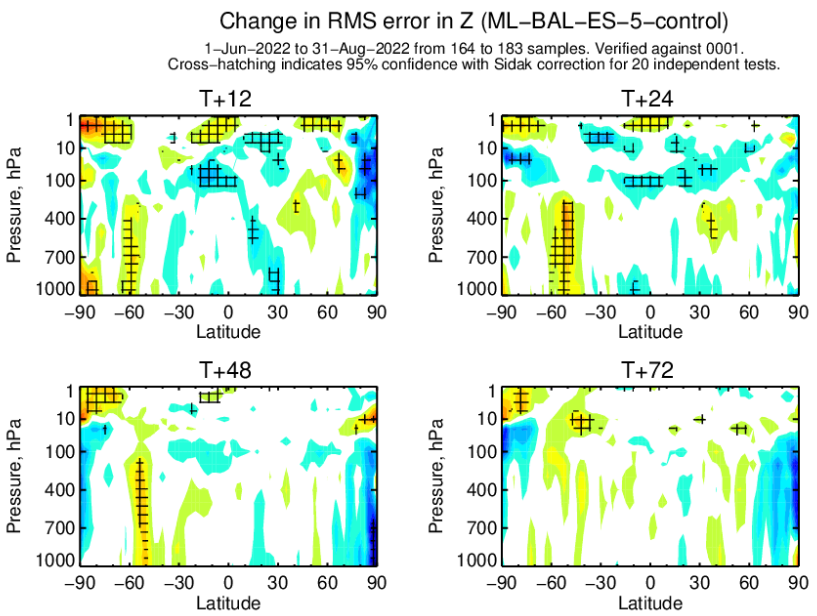
- In the wellposed case, 4DVar optimisation is continuous with respect to **B**, i.e. a small change in **B** means a small change in the resulting analysis increment.

Normalised change in RMS forecast error

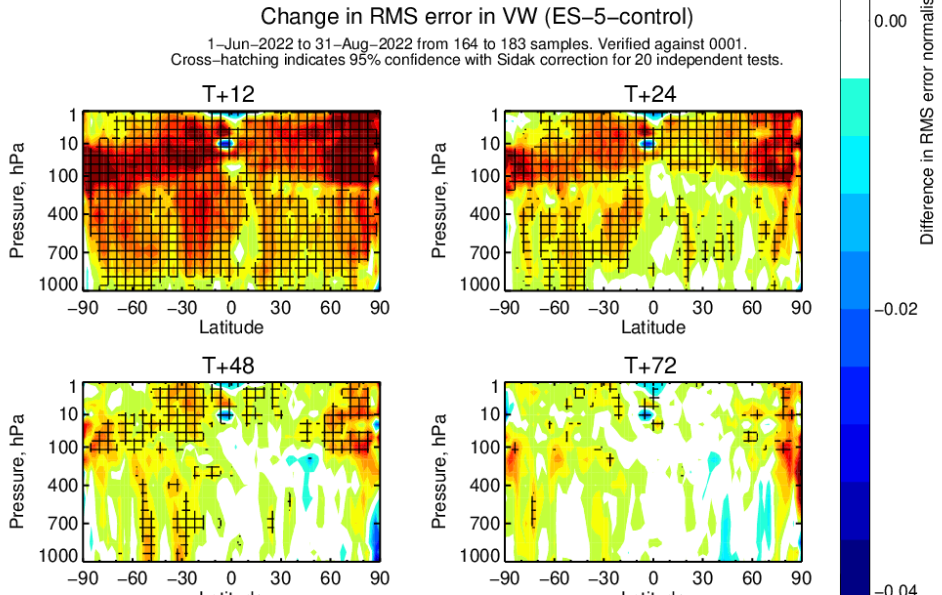
ML-es from 5 members

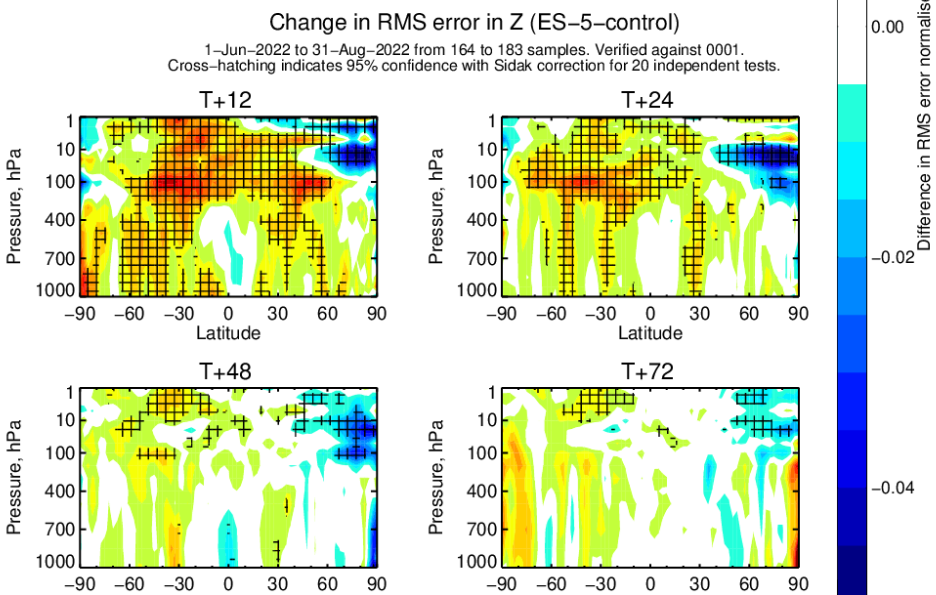
vs 50-es





5-es vs 50-es





EDA members

EDA Emulator

- **Purpose**: use machine learning to reduce the computational cost of producing the EDA ensemble.
- **Idea**: use a small ensemble of 4D-Var analyses to *guide* the ML forecast-to-analysis transition:
  - 4D-Var is responsible for assimilating observations.
  - The small 4D-Var produced ensemble carries observational conditioning and provides a noisier representation of analysis uncertainty.
  - The ML model uses the small ensemble transition to infer how the remaining background states should be transformed. Thus, observation information is incorporated indirectly.

EDA Emulation

The EDA system provides empirical background and analysis distributions

$$\mathbb{P}_N^b := \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{x}^{b,(i)}}, \qquad \mathbb{P}_N^a := \frac{1}{N}\sum_{i=1}^{N}\delta_{\phi(\mathbf{x}^{b,(i)}|\mathbf{y}_{\text{obs}})}$$

where ϕ denotes 4D-Var transformation.

Goal is to learn a **hybrid distribution** that approximates $\mathbb{P}_N^a$

$$\mathbb{P}_N^{a,\text{hybrid}} := \frac{1}{N}\left(\sum_{i=1}^{N_s}\delta_{\phi(\mathbf{x}^{b,(k_i)}\,|\,\mathbf{y}_{\text{obs}})} + \sum_{j=1}^{N-N_s}\delta_{\phi_\theta(\mathbf{x}^{b,(k_j)}\,|\,\mathbb{P}_{N_s}^a)}\right)$$

guidance

Small 4DVar ensemble distribution

ML model

We try out a **flow** model

- **Deterministic** flow for EDA emulation (***we are at this stage***)
- **Stochastic** extension enables cheap ensemble enlargement (*future work*)

$$\phi_t^{(j)} = u_\theta(\phi_t^{(j)}, \mathbb{P}_t, t \mid \mathbb{P}_{N_s}^a)dt + \sigma_\theta(\phi_t^{(j)}, \mathbb{P}_t, t \mid \mathbb{P}_{N_s}^a)\circ dW_t, \qquad \phi_0^{(j)} \in \left\{\mathbf{x}^{b,(k_j)}\right\}_{j=1}^{N-N_s}$$

ECMWF

EDA emulator

- Current prototype uses a GNN-based encoder-processor-decoder architecture, implemented using the Anemoi framework (https://github.com/ecmwf/anemoi).
- The model conditions on probability paths constructed from the small ensemble and maps *forecast perturbations* into *analysis perturbations*:
  - Perturbations can be defined relative to either the ctrl member or the small ensemble mean (to be tested).
- Prototype was trained on a single HPC node using four GH200 GPUs
  - ECMWF operational dataset: 12/11/2024 – 31/12/2025, N128, 50 EDA, fc and an, (vo, d, t, lnsp, q, o3)
  - GPU-memory constraints necessitate a **12/38 configuration**: 12 small-ensemble members, 38 emulated members.
- Training uses a flow-matching loss (*Lipman et al. 2023*).
- Loss function fine-tuning and inference-time constraints are being explored to better constrain the model.

Emulator fidelity

- Training targets one-cycle fidelity of the hybrid distribution

$$W_2^2(\mathbb{P}_N^{a,hybrid}, \mathbb{P}_N^a) \leq K_{L,\alpha} \int_0^1 \mathbb{E}_t(\|u_\theta(x,t \mid C) - u_C(x,t)\|^2) \, \mathrm{d}t$$

Wasserstein distance[2]

flow-matching loss

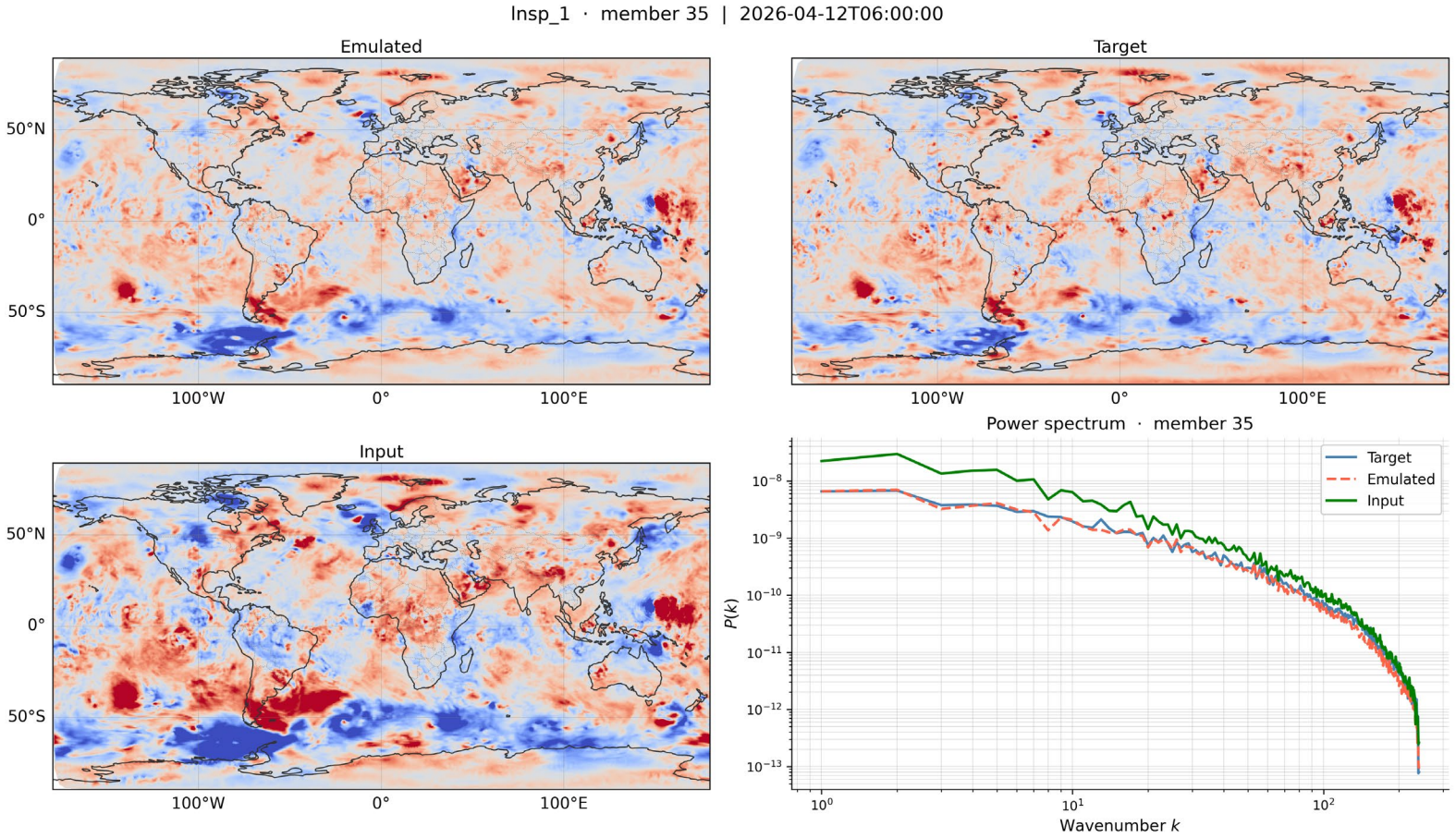
- ML model errors introduce perturbations into the EDA.
  - ***Stability*** means that the "distance" between the cycled hybrid and standard EDA ensembles remains small over successive cycles.

$$E_k^f \longmapsto C_k^f \longmapsto B_k = R_k(C_k^f) \longmapsto E_k^a \longmapsto E_{k+1}^f$$

  - The perturbations need to remain **small enough to be controlled by the EDA regularisation/damping mechanisms**.
  - "Autoregressive" training over the cycled DA system is computationally prohibitive.

Fidelity of emulated fields

- The model targets the ensemble distribution, rather than exact replication of individual EDA members.
- However, **fidelity at the individual-member level remains important**, as these members are used as initial conditions for forecasts.



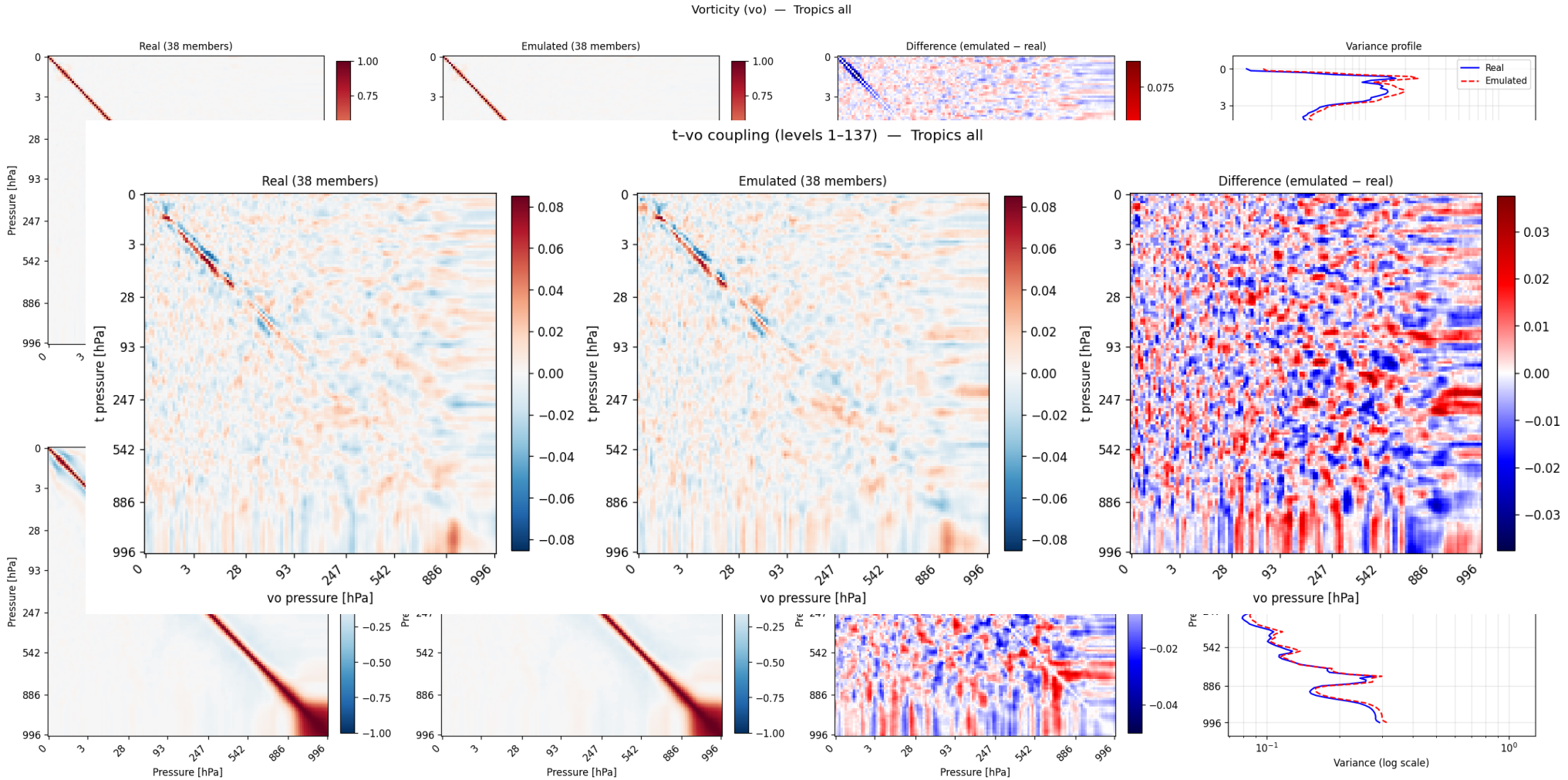
Example of an emulated **Insp analysis perturbation for member 35**, valid at **06 UTC on 12 April 2026** (not in the training dataset).

Power spectra show close agreement between the true and emulated members.

Vertical-structure fidelity

Vertical correlations and cross-correlation for vorticity and temperature perturbations, tropical region, valid at 06UTC 12/04/2026.

Vertical correlations are well reproduced. Emulated ensemble shows a slightly larger variance profile.

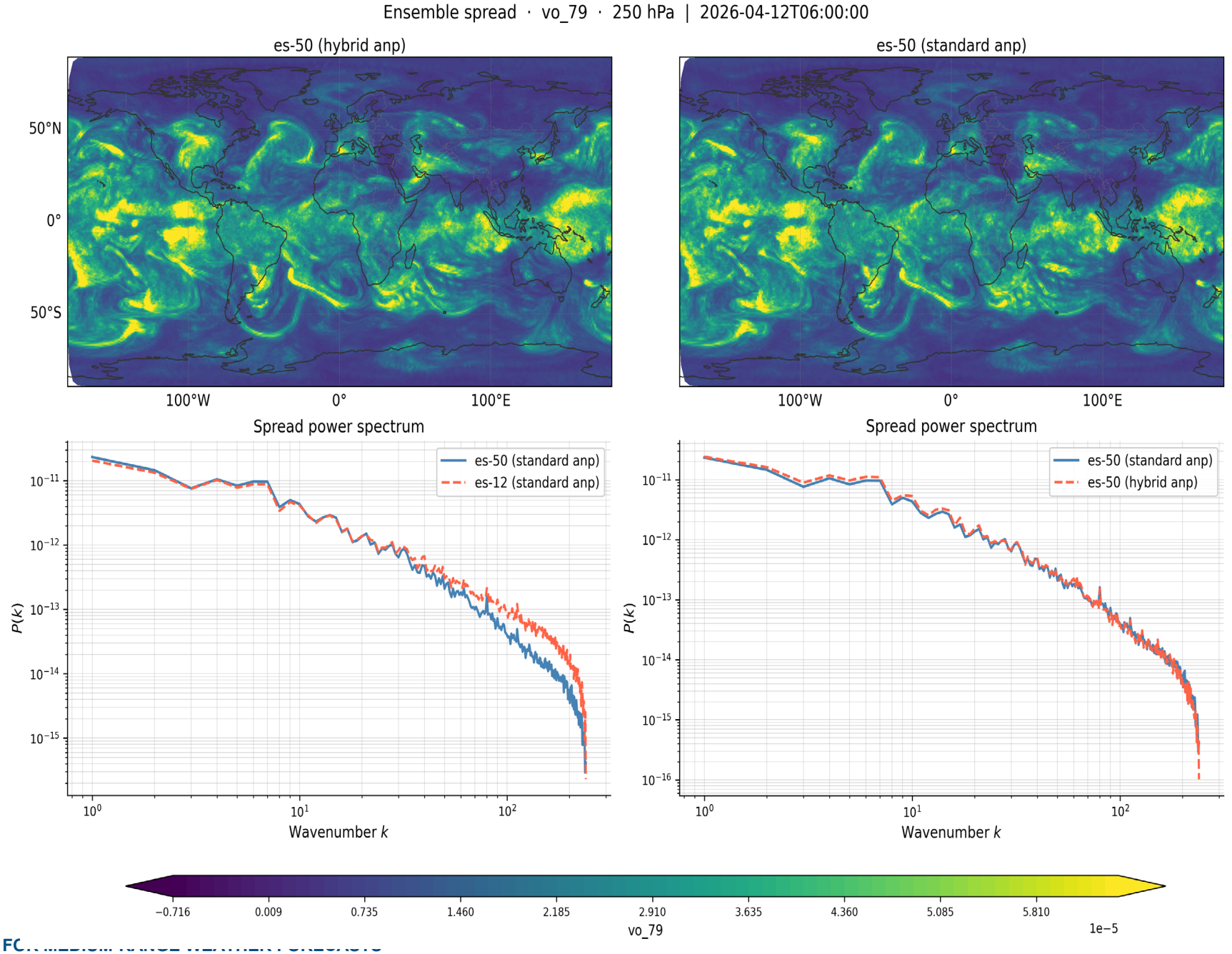


Ensemble spread

- Vorticity ~250 hPa, valid at 06 UTC on 12 April 2026.

- Close agreement between the hybrid and standard ensembles in the spread power spectra.

- Small ensemble spread shows significant excess smale-scale variability



TC example

- Ensemble spread of surface pressure. **Typhoon Kalmaegi**, valid at 18 UTC on 3 November 2025.

- Structure of physical analysis ensemble spread is well reproduced by the hybrid ensemble

- This gives evidence that the emulator is learning an observation-informed forecast-to-analysis transformation, guided by the physically generated 4D-Var small-ensemble.

Implementation in ELDA

Proof of the pudding is in the tasting

- Experiment with partly emulated ensemble is compared with one using full IFS-OOPS ensemble.
- Comparison in the research ELDA setup.



Source: MS Copilot

	REF	EMULATED
FIRSTCYCLE	2024113012	2024113012
NL model	TCo399 (27km)	TCo399 (27km)
TL/ADJ control	TL159-TL196-TL255	TL159-TL196-TL255
TL/ADJ members	T159	T159
Emulator grid	-	N128 (80km)
Non-emulated members	10	5
Emulated members	0	5

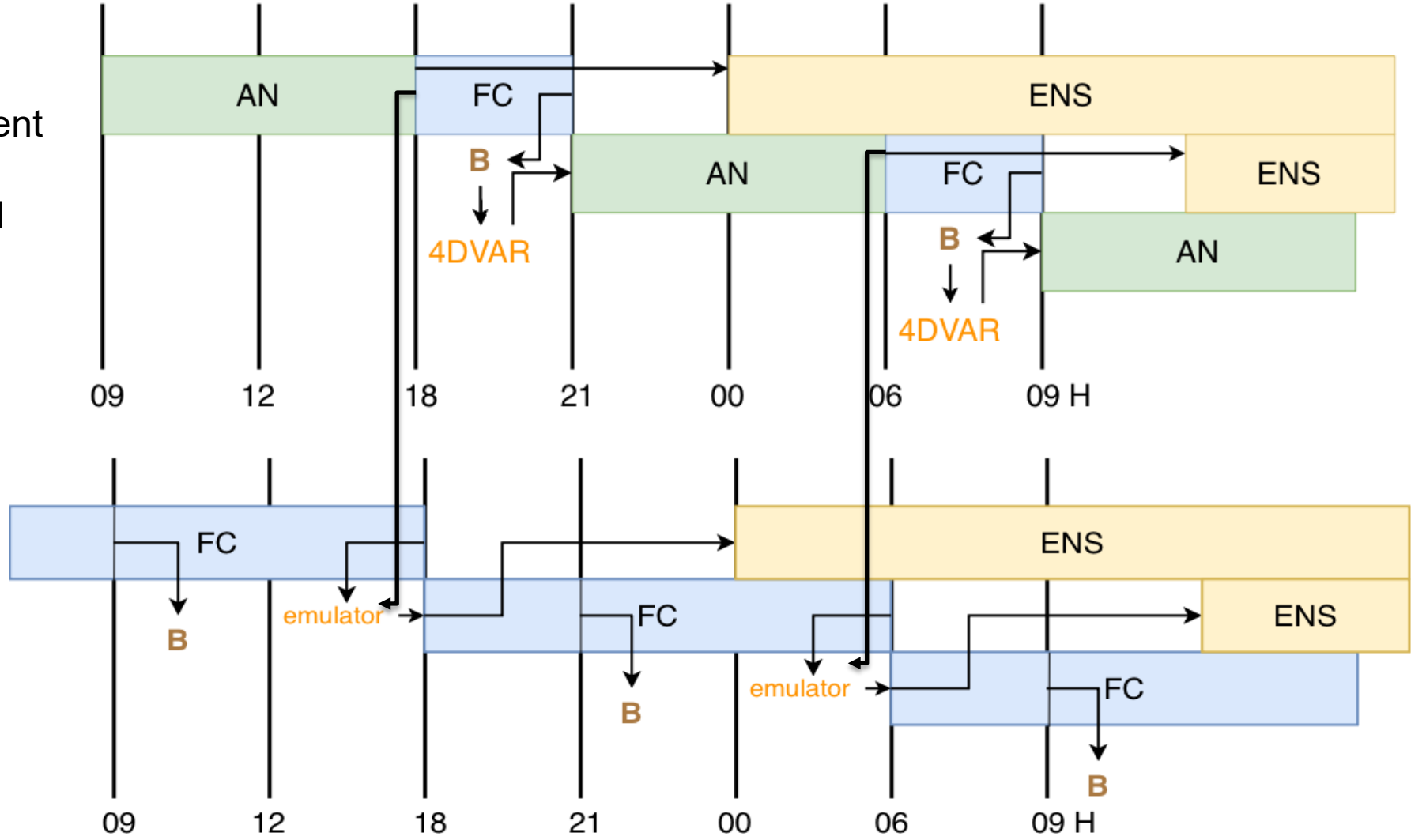
Operational flow emulated vs. non-emulated members

Non-emulated members

- Initial conditions AN are created from FC and increment produced by 4DVAR.
- Final state AN provides initial conditions FC.

Emulated members

- Initial conditions FC are created from previous FC and increment produced by the emulator.
- Emulator replaces final state AN.



Emulated increments

- Increments at 6/18H for the emulated and non-emulated members in experiment EMULATED are
  - qualitatively similar
  - and of the same order of magnitude

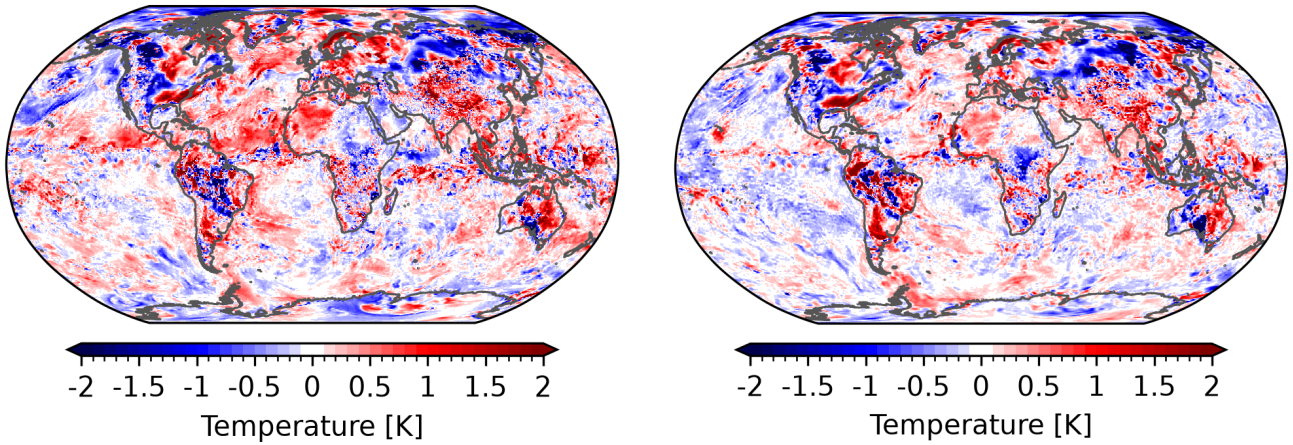
at the beginning of the experiment.

	EMULATED
FIRSTCYCLE	2024113012
NL model	TCo399
TL/ADJ control	TL159-TL196-TL255
TL/ADJ members	T159
Emulator grid	N128
Non-emulated members	5
Emulated members	5

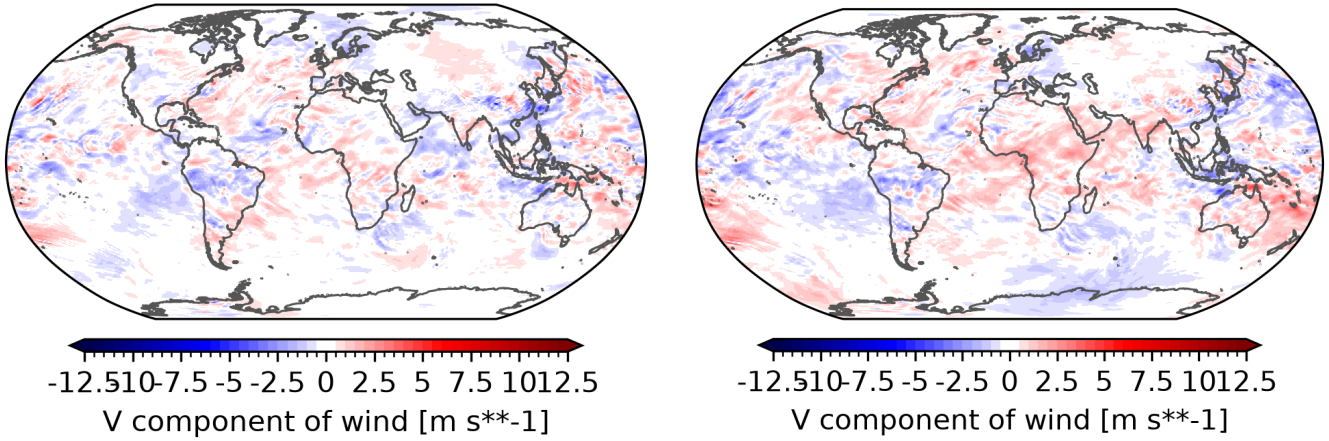
Non-emulated member

Emulated member

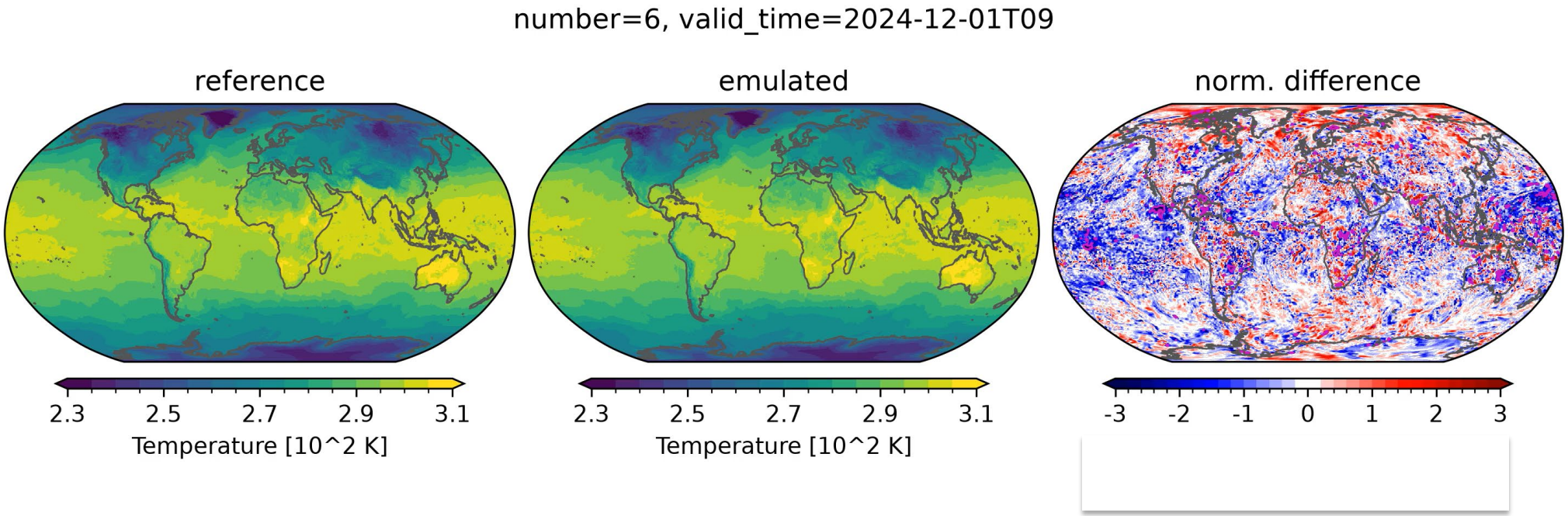
Increment (an-fc) 2024-12-05T18 level 137 (surface)



Increment (an-fc) 2024-12-05T18 level 50 (56 hPa)



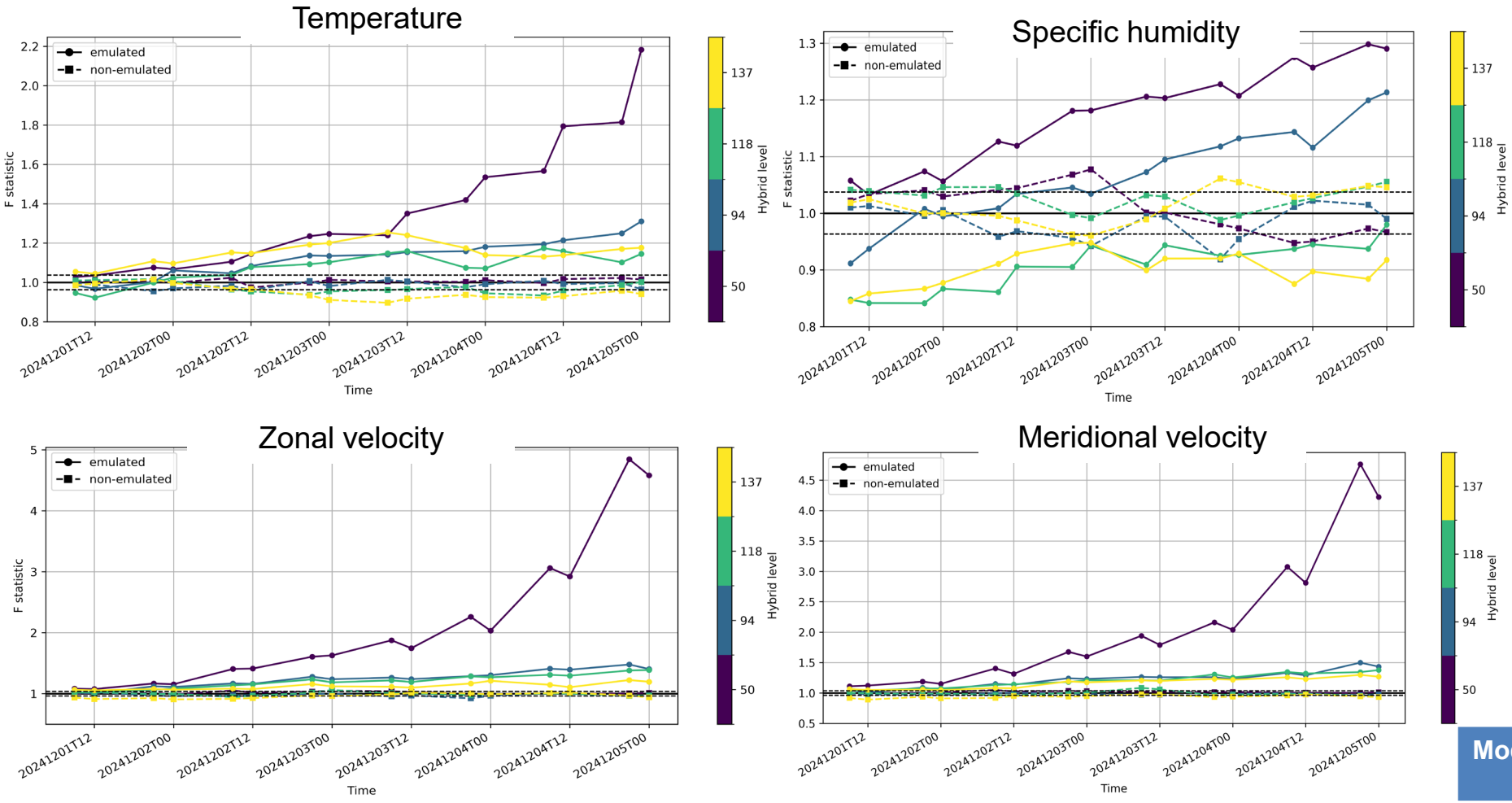
Emulated forecasts fields



Surface level temperature from short forecast for (left) member 6 in experiment REF and (centre) experiment EMULATED. (Right) Difference normalised by the ensemble std. dev. from REF. Areas where difference is significant at 95%-confidence level are contoured in magenta.

- Qualitatively forecasts started from the emulated conditions look similar to the ones from CY50R1 system.
- Differences can mostly be explained by stochasticity ensemble.
- Differences are currently still growing over time.

Emulated variances



F-statistic between the emulated/non-emulated part forecast ensemble in experiment EMULATED versus experiment REF as function of time for different variables and different vertical model levels. Black, dashed lines mark 95%-confidence interval.

- F-statistic: variances summed over all grid points.

$$F_{(\text{non}-)\text{emulated}} = \frac{\sum_i \sigma^2_{i,(\text{non}-)\text{emulated}}}{\sum_{i'} \sigma^2_{i',\text{ref}}}$$

- Non-emulated ensemble variance == REF variance (at 95% confidence level) .
- Emulated variance >> REF variance. Especially higher up (889hPa).
- Refined approach might offer solution here.

Model level	Reference pressure [hPa]
50	56
94	469
118	900
137	1012

Summary

- We presented preliminary results from a first prototype EDA emulator, which has been successfully integrated into the EDA workflow (RD version).

- Emulated analysis increments are comparable qualitatively. Work is on-going to understand the impact on cycled-DA and forecasts.

- Feedback from cycled experiments will guide improvements to model design, training and inference:
  - from fine-tuning to inference-time constraints; with cycling stability as a key requirement.
  - Newer versions of the emulator are already under development.

- The emulator can be developed in parallel with ongoing 4D-Var developments:
  - future updates to 4D-Var and IFS would likely require fine-tuning rather than redesigning the emulation methodology.

- Ensemble enlargement remains a longer-term goal.