Machine Learning-Driven Background Error Covariances for High-Resolution Data Assimilation

Ravi S Nemani[1] | Ross N Bannister[1,2] | Amos S Lawless[1,2] | Christopher Thomas[3] | Hong Wei[1]

[1]School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, UK | [2]National Centre for Earth Observation, University of Reading, Reading, UK | [3]Met Office@Reading, University of Reading, Reading, UK

Acknowledgement

This research is supported by the **AFESP-DTP** at the **University of Reading**. Additional funding and collaboration have been provided through the **Met Office CASE Award** and **National Centre for Earth Observation (NCEO), UK**.

The Sub-Km Frontier & the DA Imperative

- **The Paradigm Shift:** Operational NWP is pushing into the sub-km regime (100m – 300m scales).
- **The Physics & The DA Breakdown:** At these scales, models explicitly resolve highly localized, rapidly evolving phenomena (e.g., transient convective updrafts, sharp boundary layer inversions).
- **The DA Requirement:** To properly initialize these scales, assimilation increments must respect these localized structures.
- **The Consequence:** Classic $\boldsymbol{B}$ -matrix assumptions (hydrostatic/geostrophic balance, homogeneity, gaussianity) fail completely. The errors are highly non-linear and anisotropic.

Vertical error covariances in operational VarDA

- **Modelling $\boldsymbol{B}$**: Operational VarDA relies on the control variable transform (CVT): $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}}$.
- The chain of transforms: $\boldsymbol{U} = U_p U_h U_v$
  - U_p: Physical/Balance transform
  - U_h: Horizontal transform (contains horizontal correlations)
  - U_v: Vertical transform (contains vertical error covariances)

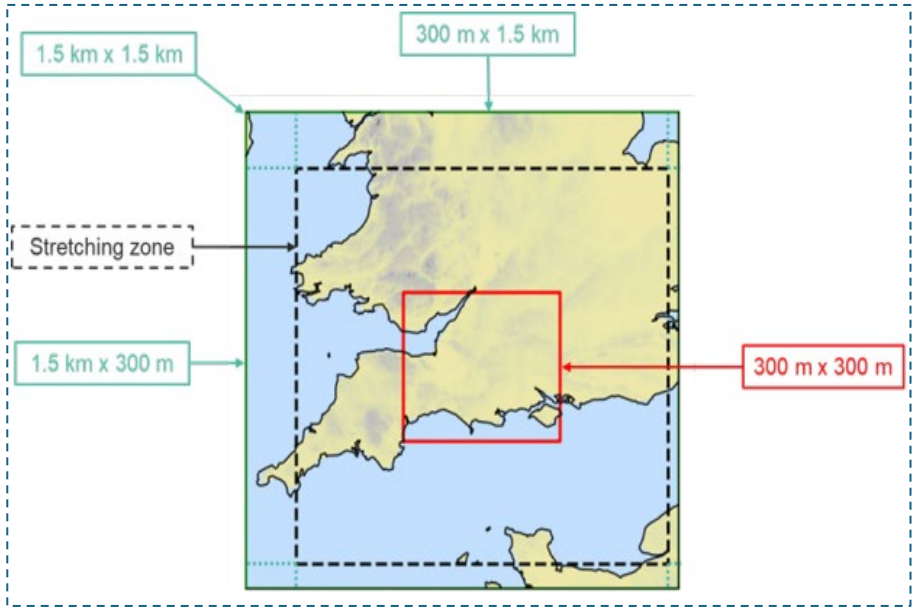
- **The Computational Bottleneck:** A flow-dependent U_v requires a high-resolution Ensemble of Data Assimilations (EDA). Running an EDA at sub-km resolution is computationally prohibitive for real-time operations.
- **The Fatal Compromise:** Reverting to a static, climatological U_v smears increments unphysically (e.g., propagating surface data through a strong capping inversion).

Research Question

“Is it possible to estimate ensemble-based vertical error covariances for sub-kilometre data assimilation using machine learning, without generating an ensemble?”

Met Office High-res ensemble

- **The Dataset:** Met Office 18 mem lagged ensemble Wessex model (WMV) at 300m resolution with 70 model levels.
- Error covariances derived from these expensive ensembles as the ground truth.



The variable resolution 300m Wessex model (WMV) domain.
Kirsty Hanley, Humphrey Lean Technical Report 667
https://doi.org/10.62998/DITE4759

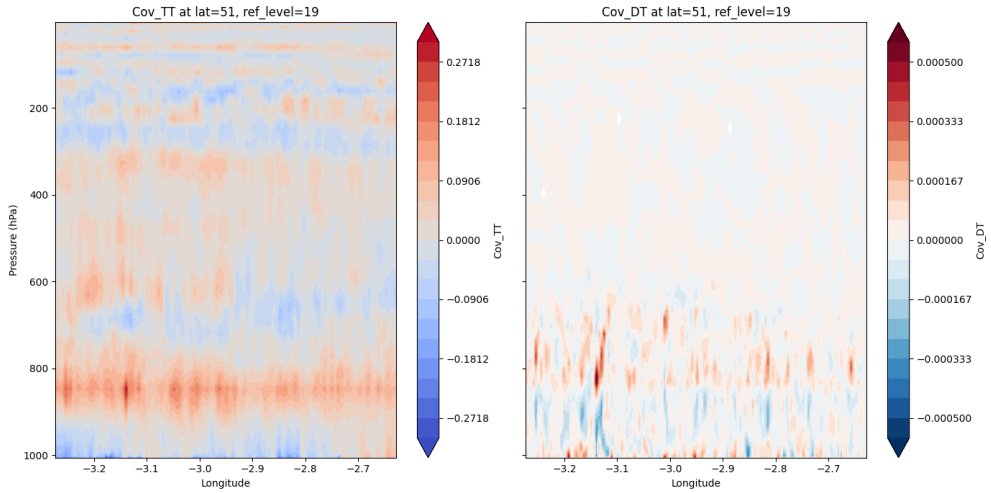
Perturbations:

- $x'_{\mathrm{i}} = x_i - \bar{x}$

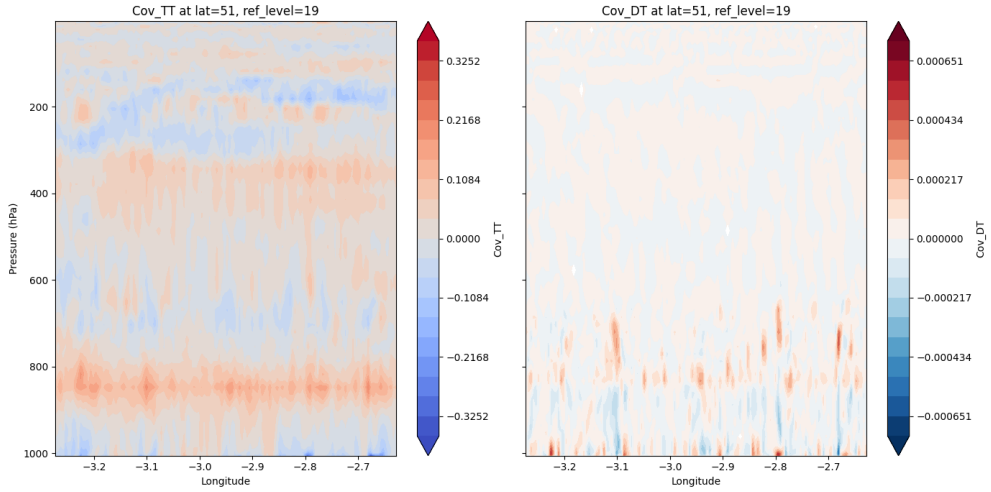
Covariance

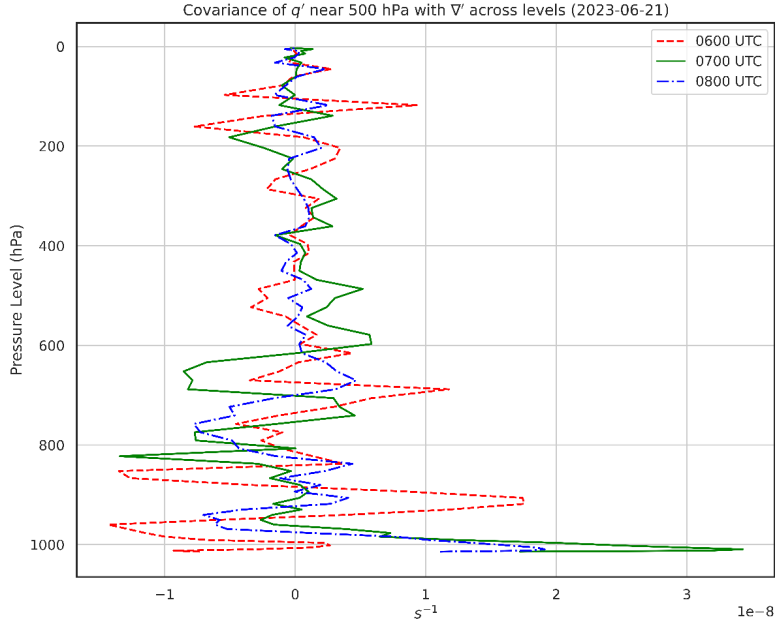
- $B_{ens} = \frac{1}{N-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} x'_{\mathrm{i}} \, x'^{T}_{\mathrm{i}}$

21062023 09:00 UTC



21062023 10:00 UTC



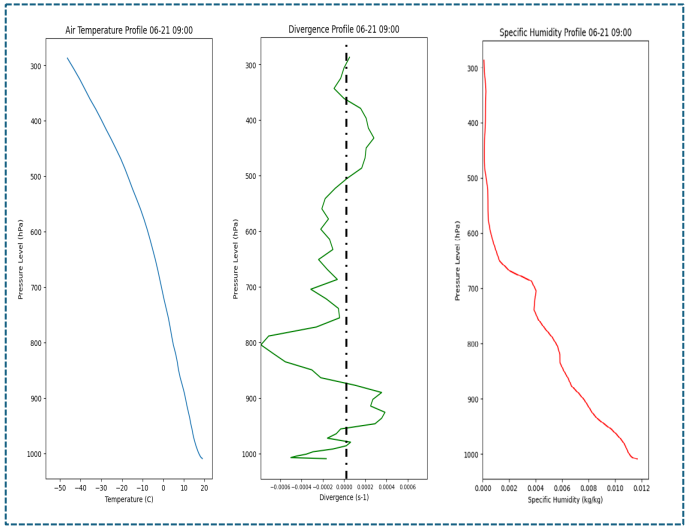


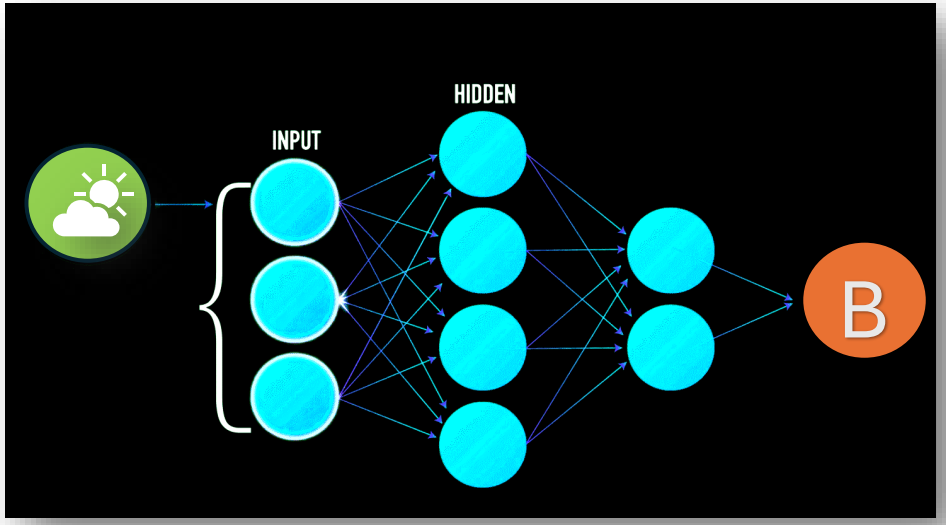
Covariance profiles for specific humidity perturbations (q') near 500 hPa with divergence perturbations (∇') across pressure le levels for 21 June 2023 at 0600, 0700, and 0800 UTC. Each curve shows the covariance between q' at a fixed level and ∇' at all levels from surface to top.

Observed Statistics: The sample covariances exhibit the necessary fine-scale, anisotropic and flow dependent vertical error structures.

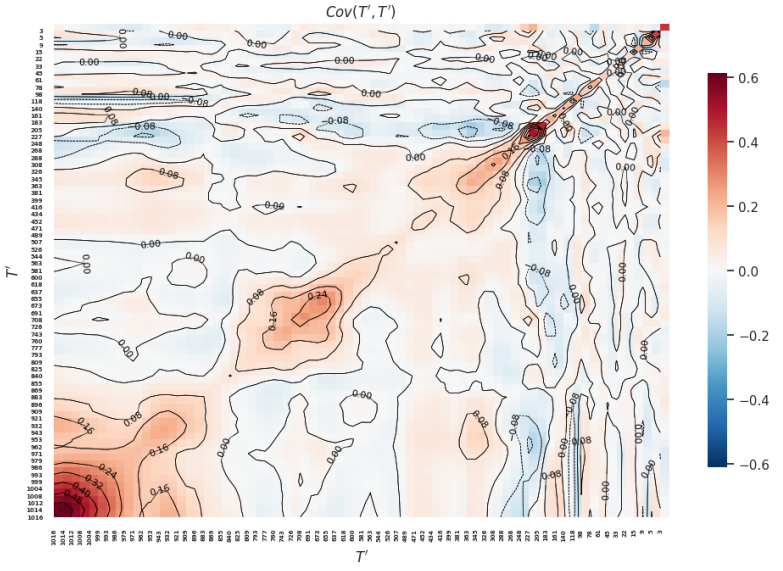
The Machine Learning approach

- **Input:** Vertical profiles(e.g., $T, q, \delta, p, \mathrm{w}, \mathrm{cl}$) from a single deterministic forecast.
- **Architecture:** 3-layer FCN utilizing Swish activation, Batch Normalization, and 10% Dropout.
- **Dataset:** ~90,000 samples derived from vertical profiles and error covariance matrices (collected June 21, 2023, during shallow convection).
- **Training Method:** Offline training across six time steps; evaluated on held-out test data.
- **Optimization:** Constraint-based loss function focused on reconstructed RMSE.





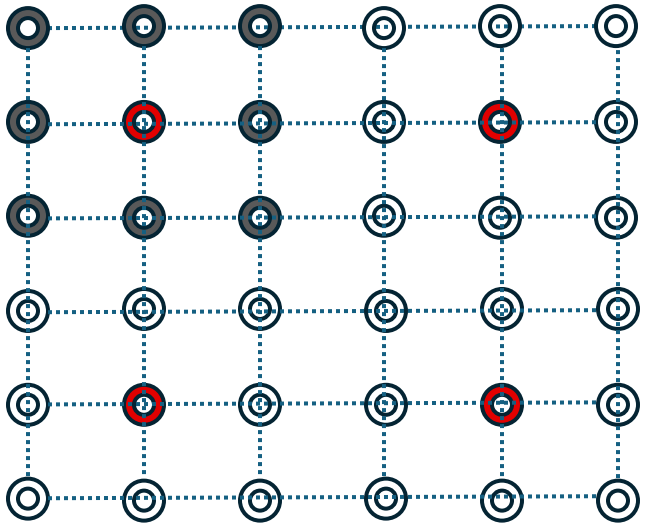
$$Y = f_{FCN}(X, \theta)$$



The Sampling error problem

- The introduction of sampling errors (noisy matrix).
- **Ways to counter the sampling error problem:** More ensemble members & Localisation (preventing spurious long correlations)

We partly addressed this problem by boosting the ensemble members (18 to 162 ens mem).



Rank deficiency & Ill-conditioning

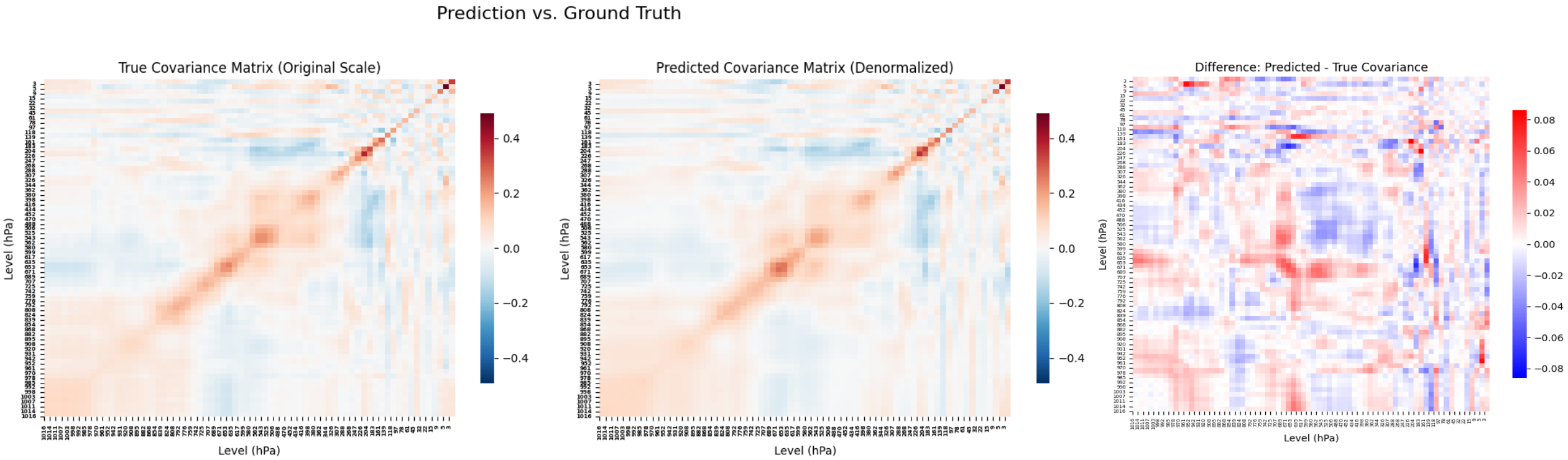
- High vertical resolution + limited members = near-singular target matrices.
- **Solution:** Applied ridge regression to all the vertical error covariance matrices to have strictly positive eigen values

$$\boldsymbol{B}_{reg} = \boldsymbol{B}_{ens} + \epsilon \boldsymbol{I}$$

The Output

The ML model is architecturally constrained to predict the Cholesky factor $(\mathbf{L})$. Since $\mathbf{B_{pred}} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$, SPD is mathematically guaranteed, and $\mathbf{L}$ serves as a direct replacement for U_v.

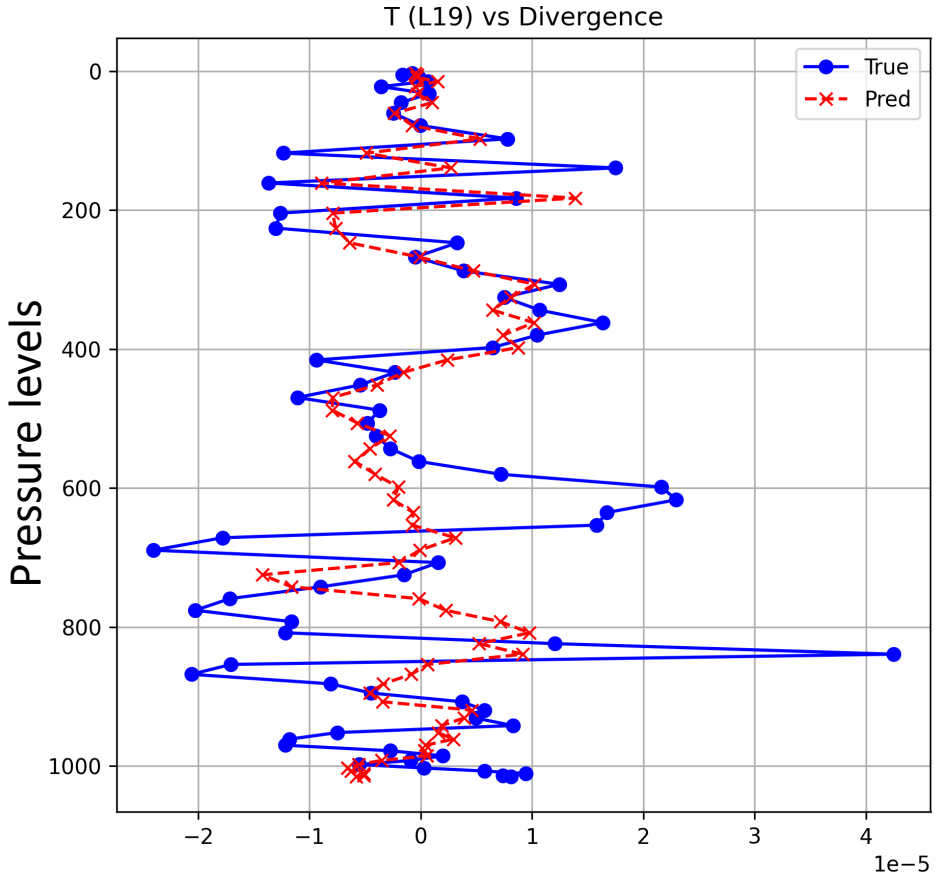
Preliminary Results



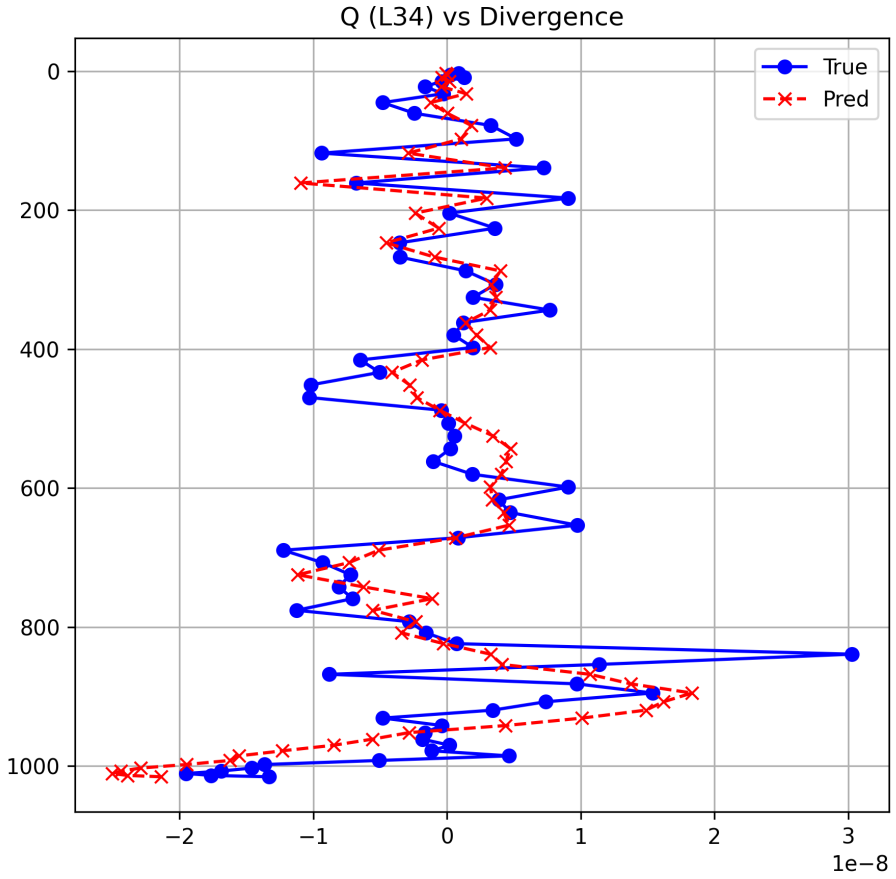
Ensemble estimated Cov(T’,T’)

ML predicted Cov(T’,T’)

R2 Score: 0.9437



Cov(T'@850hPa, D')



Cov(Q'@500hPa, D')

Conclusions & Future work

- Our ensemble-derived error covariances reveal **complex and highly flow-dependent structures**.
- Preliminary results show it can learn the primary structures of vertical error covariances of temperature, divergence and specific humidity.
- Continue training with an expanded dataset and explore a different architecture.

Impact on VarDA :

The predicted vertical error covariances has a potential to replace the static ones in the cost function.

Hence achieving “Ensemble-like” quality at “Deterministic cost” in pure variational and hybrid data assimilation schemes

Contact information: r.nemani@pgr.reading.ac.uk | @itsme-rsn.bsky.social