# Data Assimilation Methodologies for Future Reanalyses

Ziga Zaplotnik, Massimo Bonavita

Thanks to Hans Hersbach, Paul Poli, Marcin Chrust, Matic Pikovnik, Bostjan Melinc, Lukas Kugler



# Content

1) From ERA5 to ERA6
2) Weaknesses of current reanalyses:
   a) Representation of extreme events
   b) 4D-Var window discontinuities
   c) Mean-state biases (in Hadley circulation)
   d) …
3) DA methodologies to further improve the reanalyses
4) How could a future ECMWF **Earth-system** reanalysis look like?

# ERA6 supersedes ERA5




Paul Poli and Reanalysis Team

# Weaknesses of current reanalysis systems

## Representation of extreme events

# Representation of extreme events

1) Most of the economic/societal value of weather forecasts is concentrated in high-impact extreme events with low prevention costs relative to potential losses (World Bank, 2025)

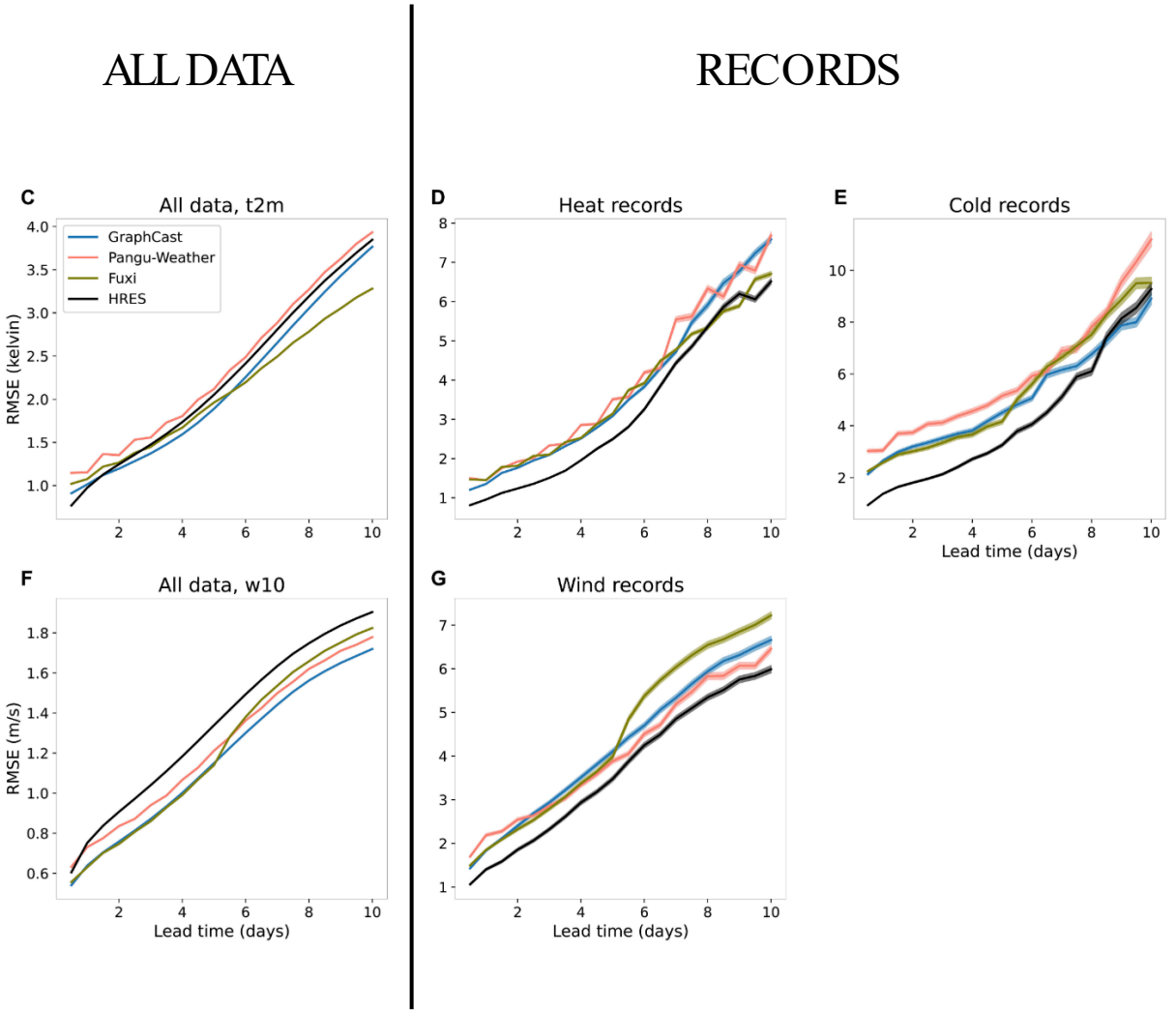
2) Physics-based models still outperform AI weather forecasts of record-breaking extremes (Zhang et al., 2026).

With the current pace of climate change, the attractor is quickly shifting outside of the training sample distribution (NWP models not completely immune either!)

3) Higher-resolution models and DA schemes are needed to better represent:

- Deep tropical cyclones
- Intense precipitation events
- Intense windstorms

→ Physical reanalyses remain crucial!




Zhongwei Zhang et al., 2026. Physics-based models outperform AI weather forecasts of record-breaking extremes. Sci. Adv. 12, eaec1433

# Representation of extreme events

1) Most of the economic/societal value of weather forecasts is concentrated in high-impact extreme events with low prevention costs relative to potential losses (World Bank, 2025)

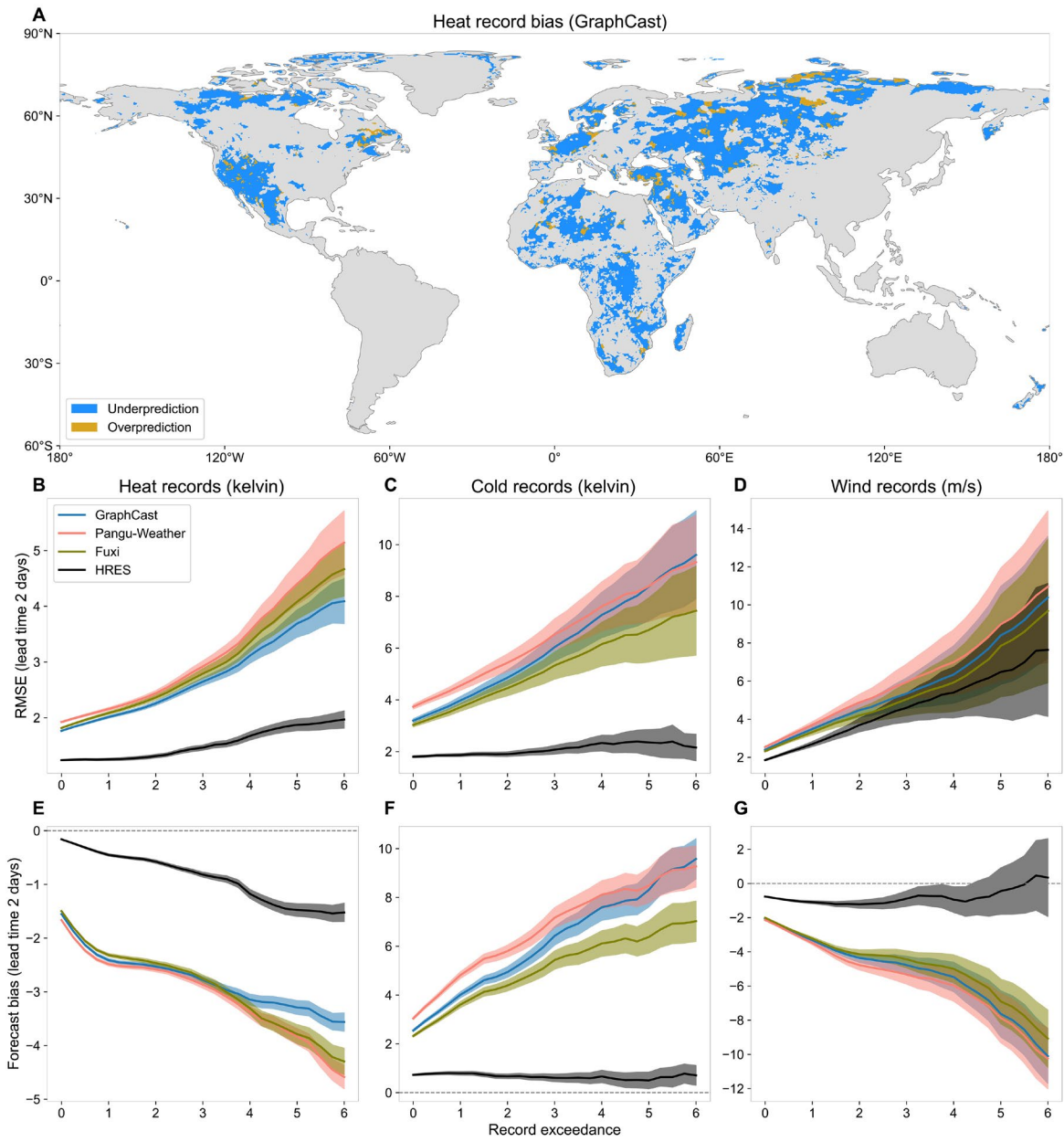
2) Physics-based models still outperform AI weather forecasts of record-breaking extremes (Zhang et al., 2026).

With the current pace of climate change, the attractor is quickly shifting outside of the training sample distribution (NWP models not completely immune either!)

3) Higher-resolution models and DA schemes are needed to better represent:

- Deep tropical cyclones
- Intense precipitation events
- Intense windstorms

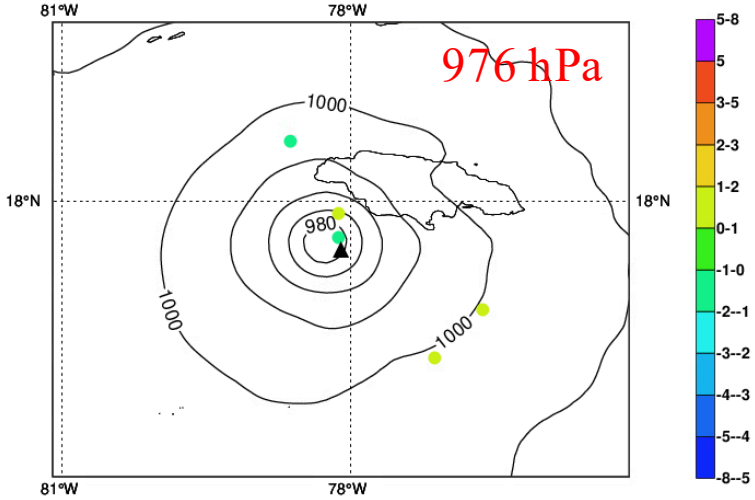
→ Physical reanalyses remain crucial!



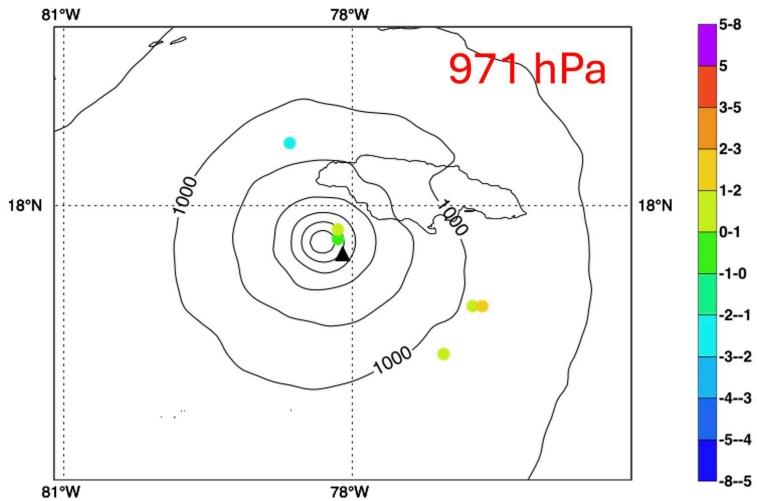
# Representation of extreme events – Hurricane Melissa

- Hit Jamaica on October 28, 2025, at 17:25 UTC as a Category 5 tropical cyclone
- 892 hPa min pressure, max 1-minute sustained winds of 305 km/h
- 95 fatalities, 27 missing (Wiki)
- The plots show the 6-hour bg forecast, initialised at Oct 28, 2025, 06 UTC, valid at Oct 28, 2025, 12 UTC

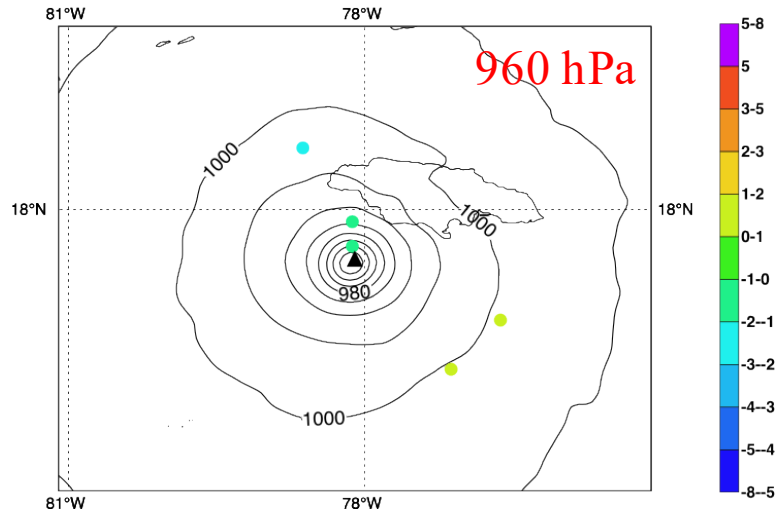
50r1 ERA6-like TCo799 (14 km) + TL399 4D-Var (65 km)



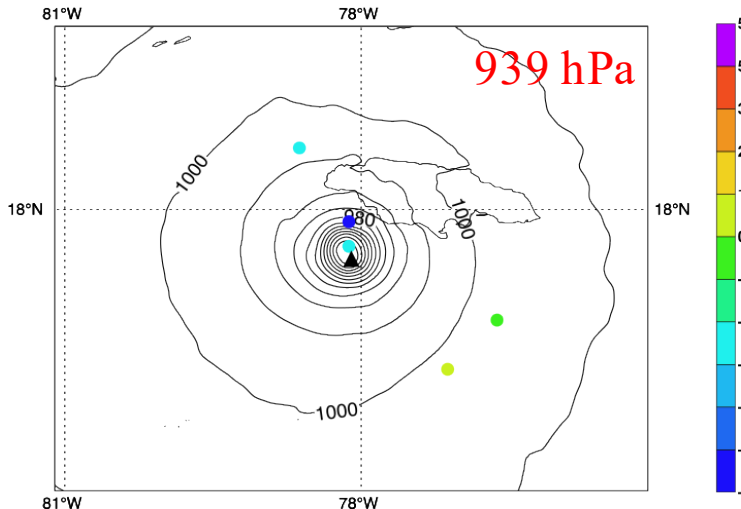

50r1 e-suite TCo1279 (9 km) + TL511 4D-Var (40 km)




50r1 TCo1279 (9 km) + TCo511 4D-Var (20 km)




50r1 TCo2559 (4.4 km) + TCo511 4D-Var (20 km)

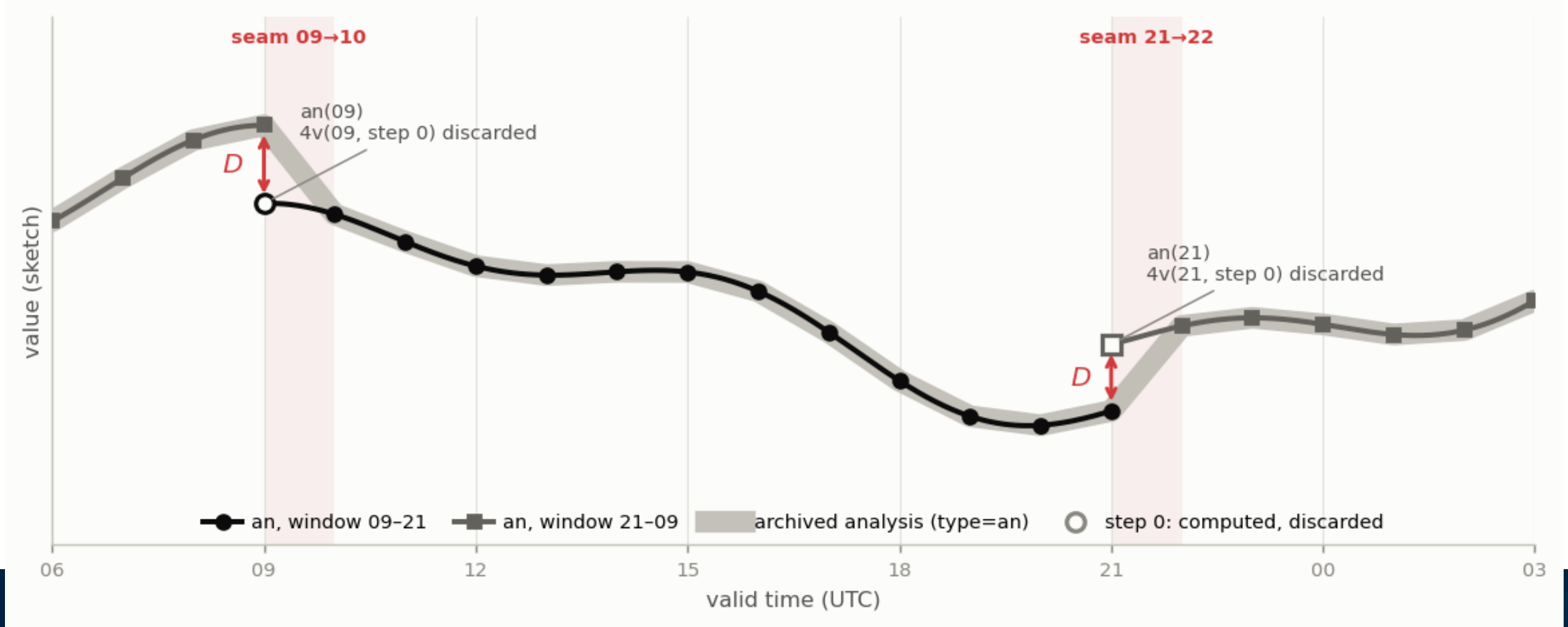


# Weaknesses of current reanalysis systems

## Discontinuities between 4D-Var windows

# Discontinuities between 4D-Var windows

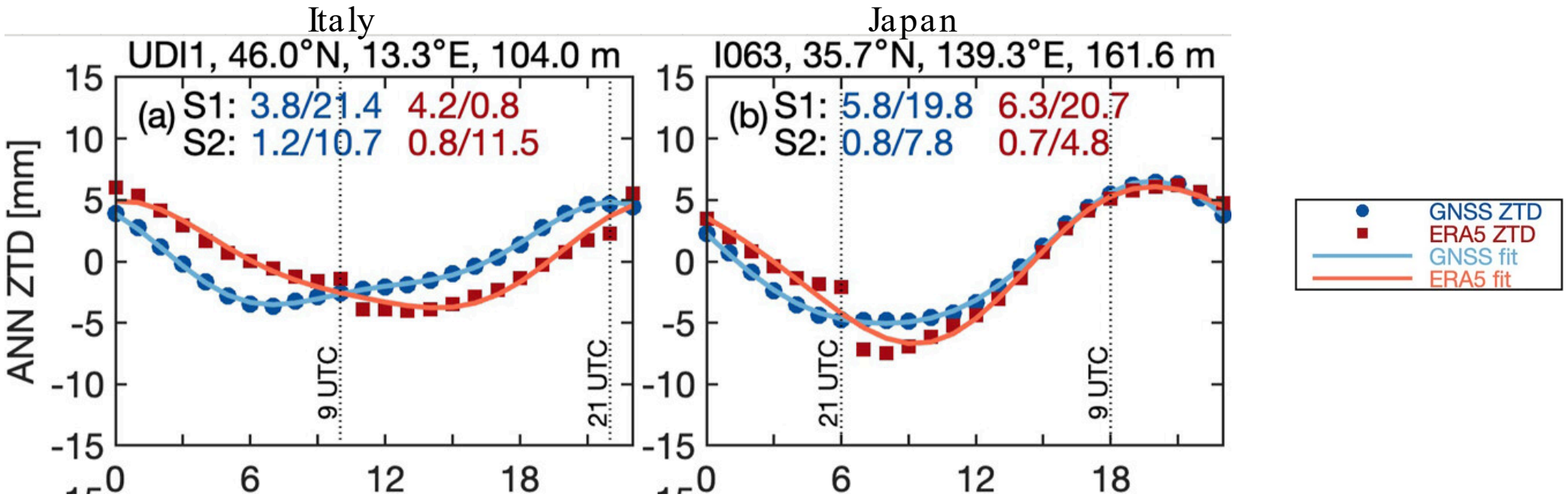
- An example of discontinuity between 4D-Var windows: the last step of the old window is archived as “an”, while the first step of the new window is discarded.



# Discontinuities between 4D-Var windows

- Discontinuities affect the representation of the diurnal cycle
- Numerous recent studies have examined this ERA5 property:
  - Tropospheric zenith total delay discontinuity: Yuan et al., 2025, GRL

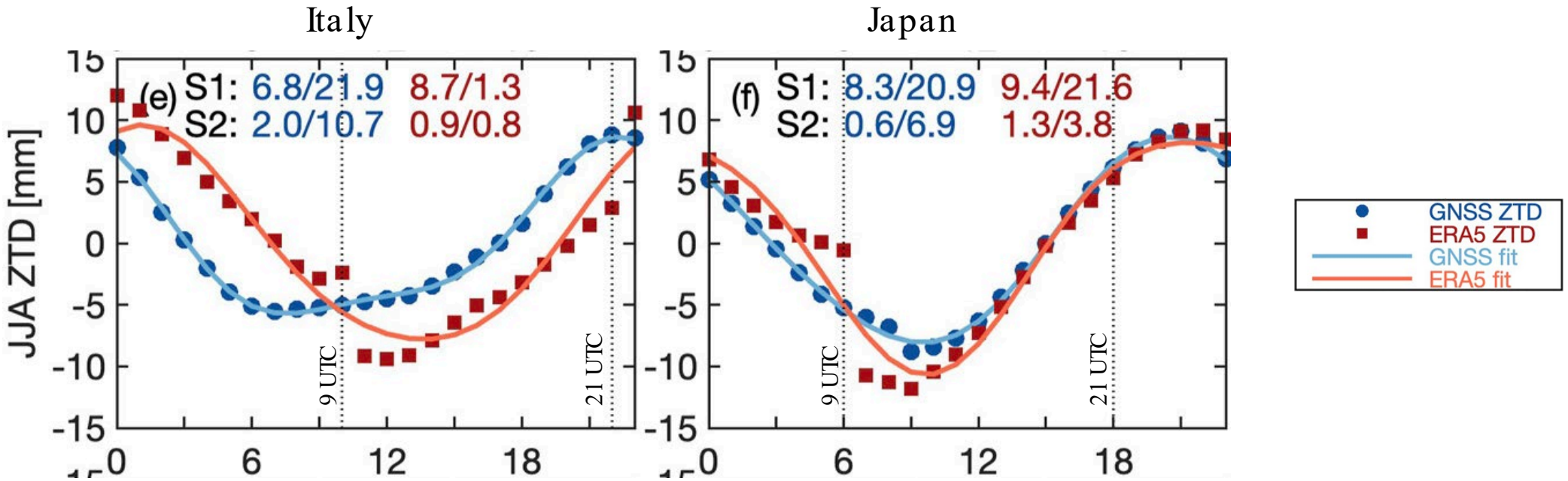
Annual mean



# Discontinuities between 4D-Var windows

- Discontinuities affect the representation of the diurnal cycle
- Numerous recent studies have examined this ERA5 property:
  - Tropospheric zenith total delay discontinuity: Yuan et al., 2025, GRL

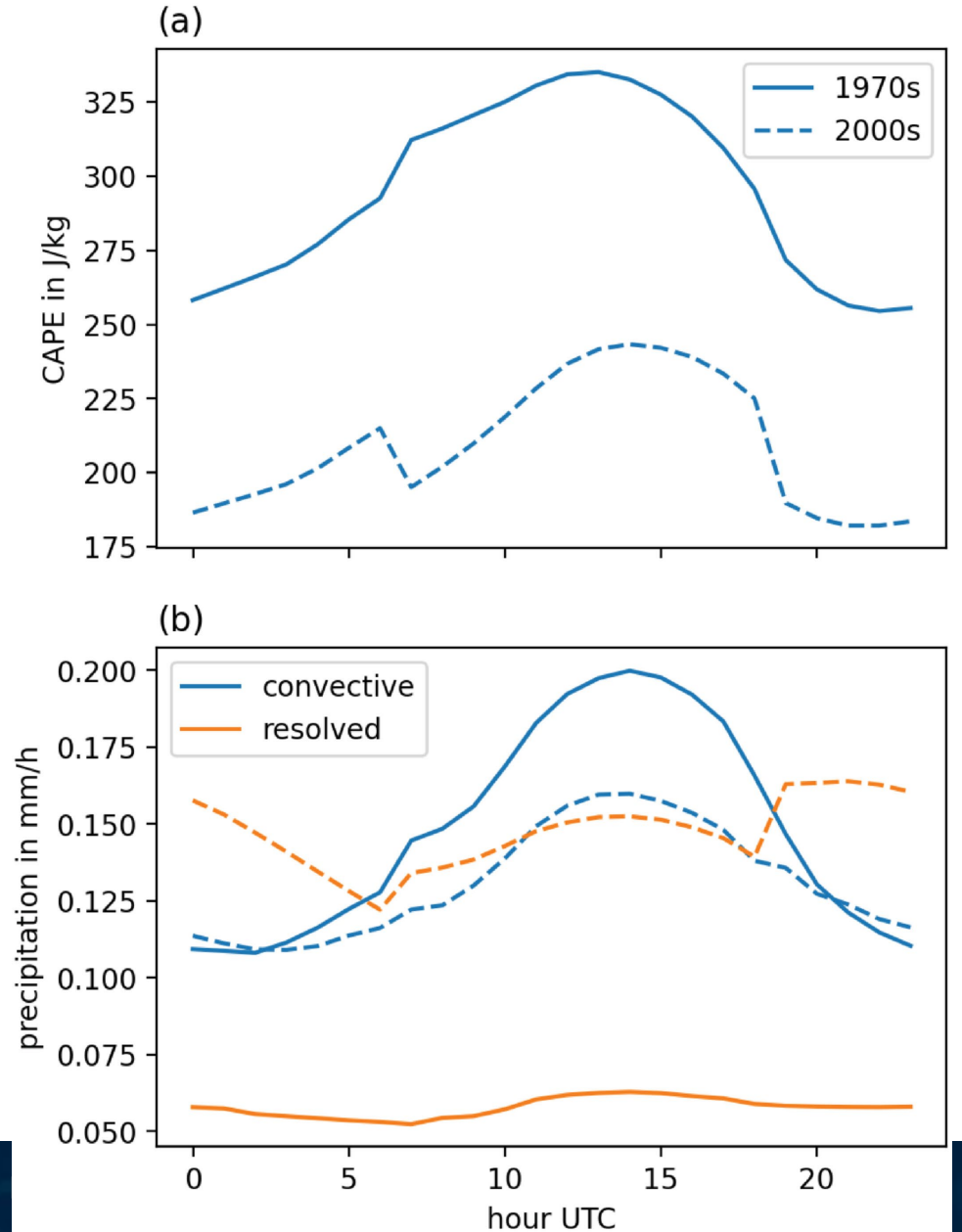
NH summer




The discontinuity is amplified in the summer period, when the amplitude of the daily cycle increases!

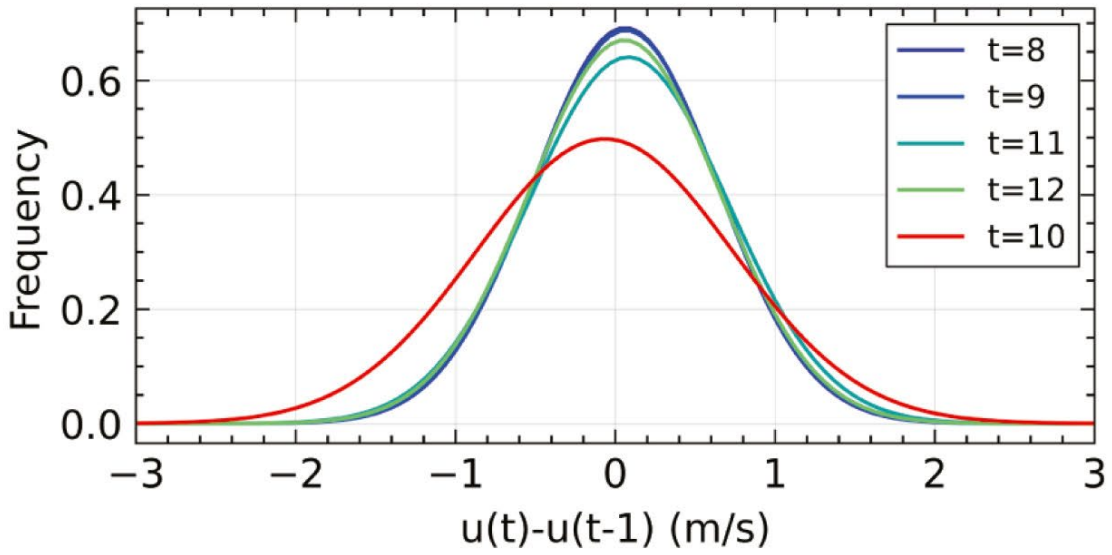
# Discontinuities between 4D-Var windows

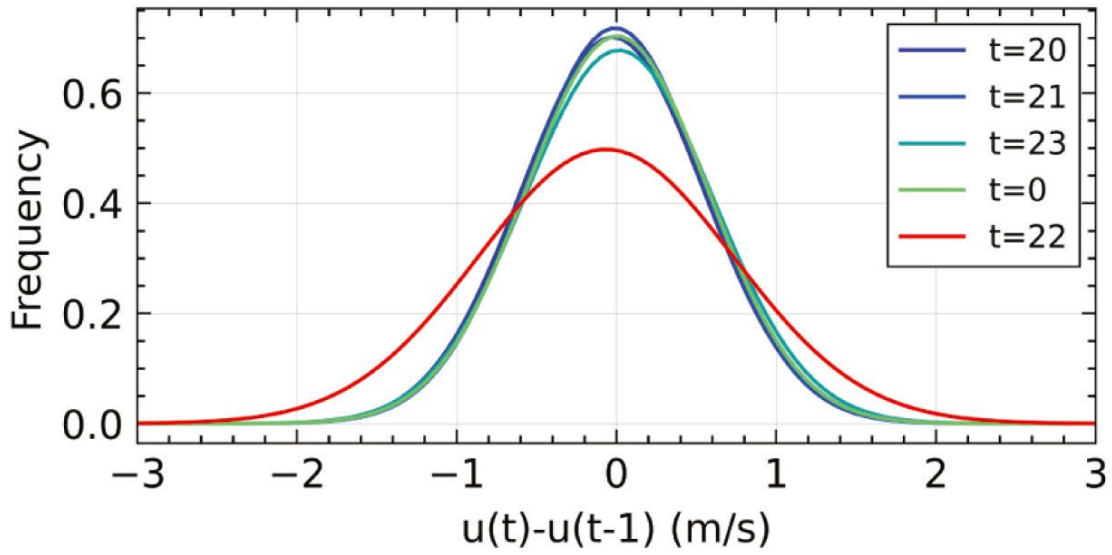
- Discontinuities affect the representation of the diurnal cycle
- Numerous recent studies have examined this ERA5 property:
  - Tropospheric zenith total delay discontinuity: Yuan et al., 2025, GRL
  - Precipitation and CAPE discontinuity: Buschow, 2023, QJRMS



# Discontinuities between 4D-Var windows

- Discontinuities affect the representation of the diurnal cycle
- Numerous recent studies have examined this ERA5 property:
  - Tropospheric zenith total delay discontinuity: Yuan et al., 2025, GRL
  - Precipitation and CAPE discontinuity: Buschow, 2023, QJRMS
  - Surface wind discontinuity: Potisomporn et al., 2023, Energy reports

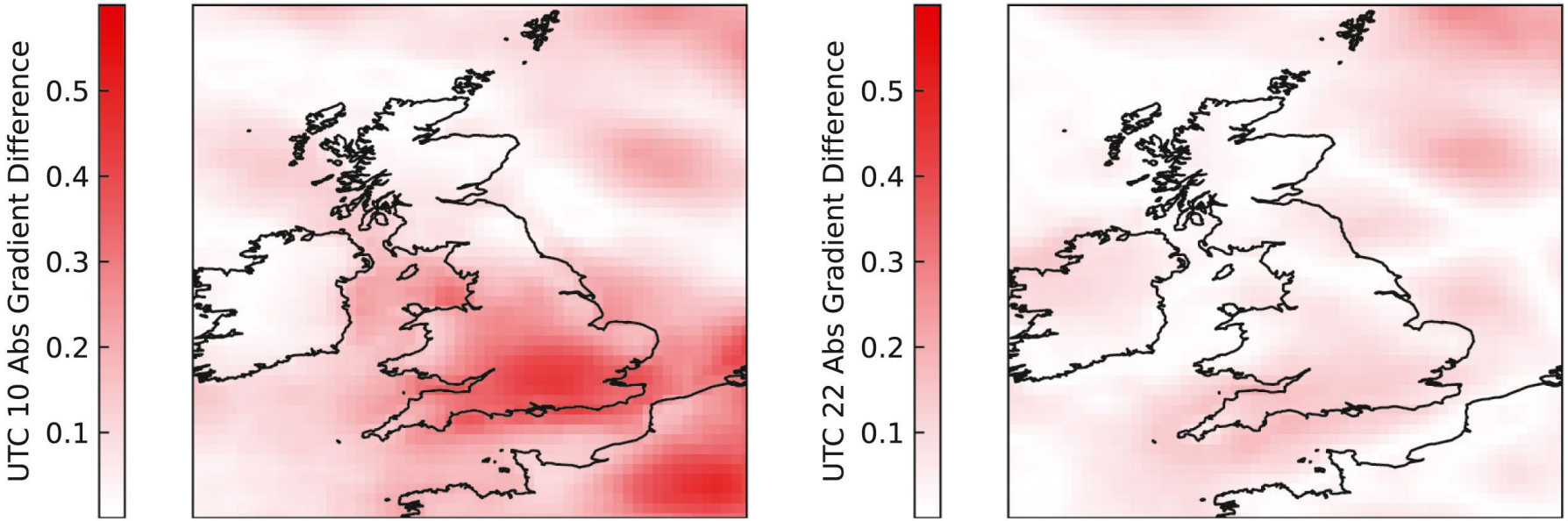


# Discontinuities between 4D-Var windows

- Discontinuities affect the representation of the diurnal cycle
- Numerous recent studies have examined this ERA5 property:
  - Tropospheric zenith total delay discontinuity: Yuan et al., 2025, GRL
  - Precipitation and CAPE discontinuity: Buschow, 2023, QJRMS
  - Surface wind discontinuity: Potisomporn et al., 2023, Energy reports

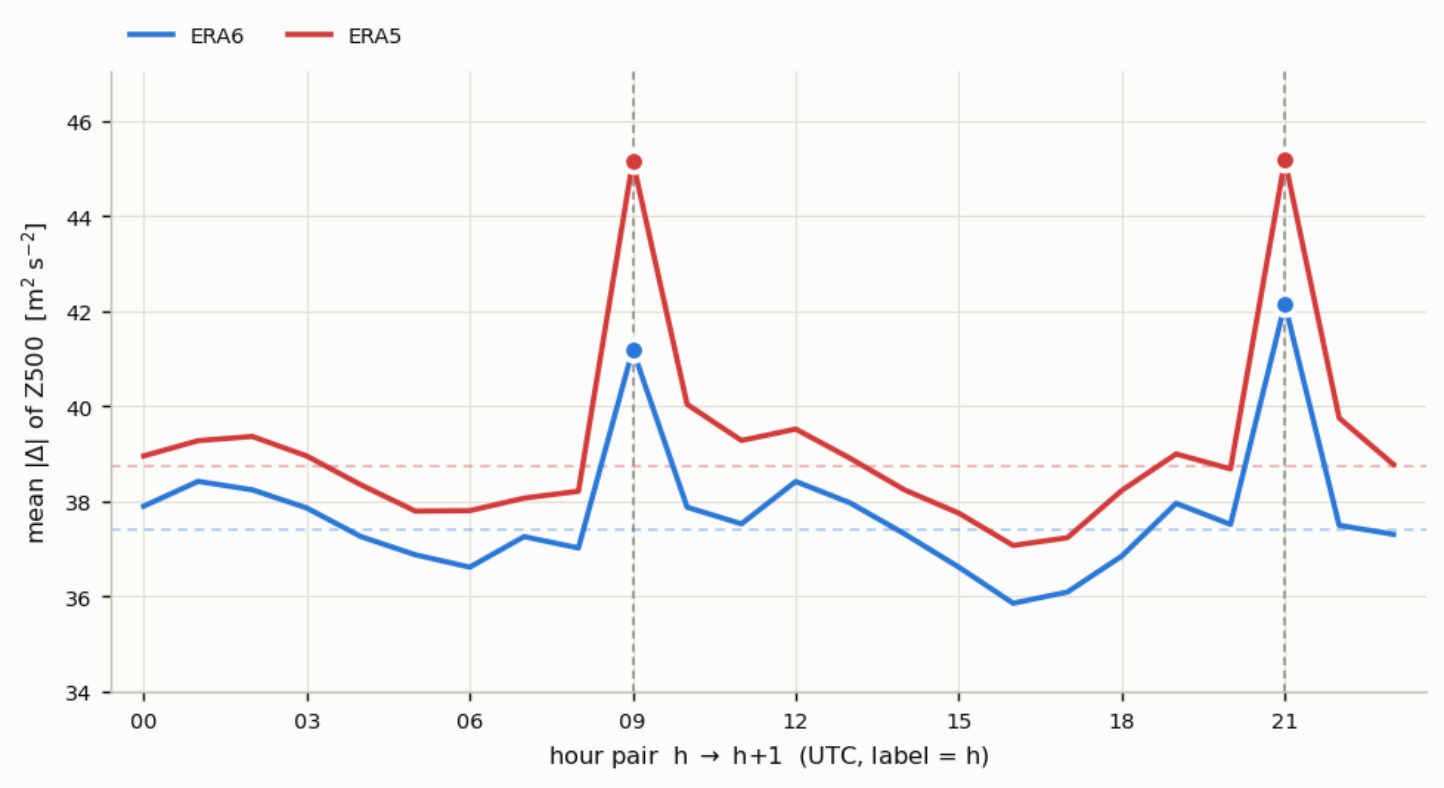
$\left|\frac{\partial u}{\partial t}\right|$



# Discontinuities between 4D-Var windows

How large are the discontinuities?
We can compare the 2 transition steps (9→10 and 21→22) with the 22 non-transition steps in terms of $|\mathbf{x}(t+1) - \mathbf{x}(t)|$

Mean absolute jump in Z500: ERA6 vs ERA5



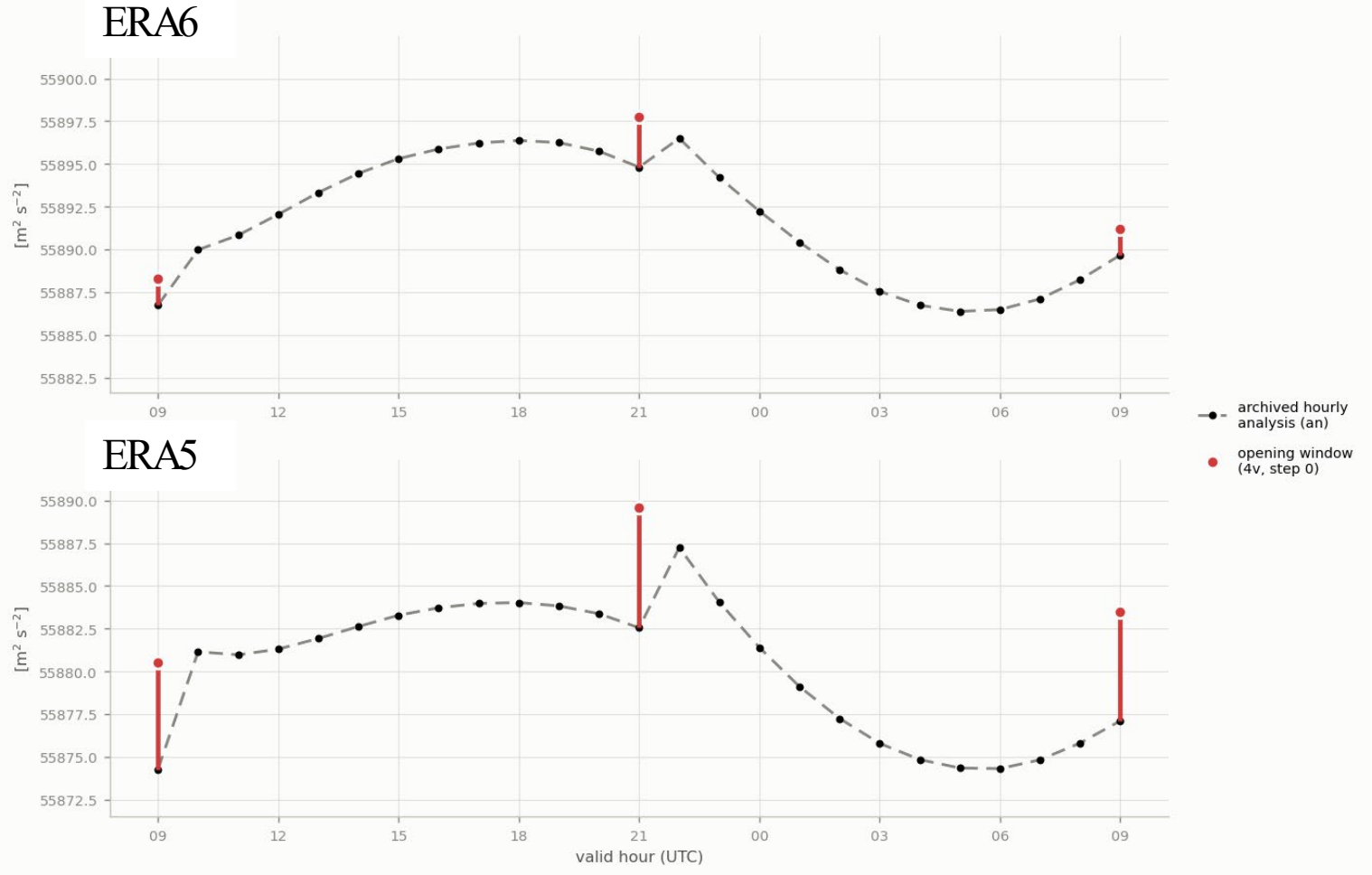

× 1.15

# Discontinuities between 4D-Var windows

How large are the discontinuities?

We can measure $\Delta = |\mathbf{x}(\text{window} = i, \text{step} = 0) - \mathbf{x}(\text{window} = i - 1, \text{step} = 12)|$

Z500, global, 4-17 July 2017, N320

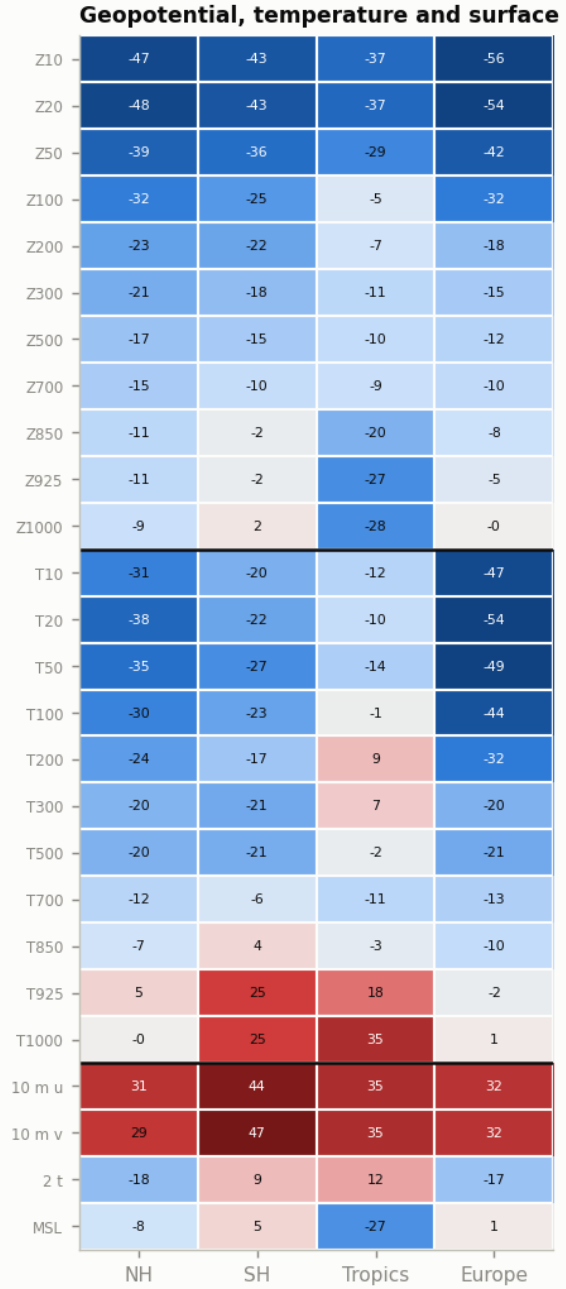


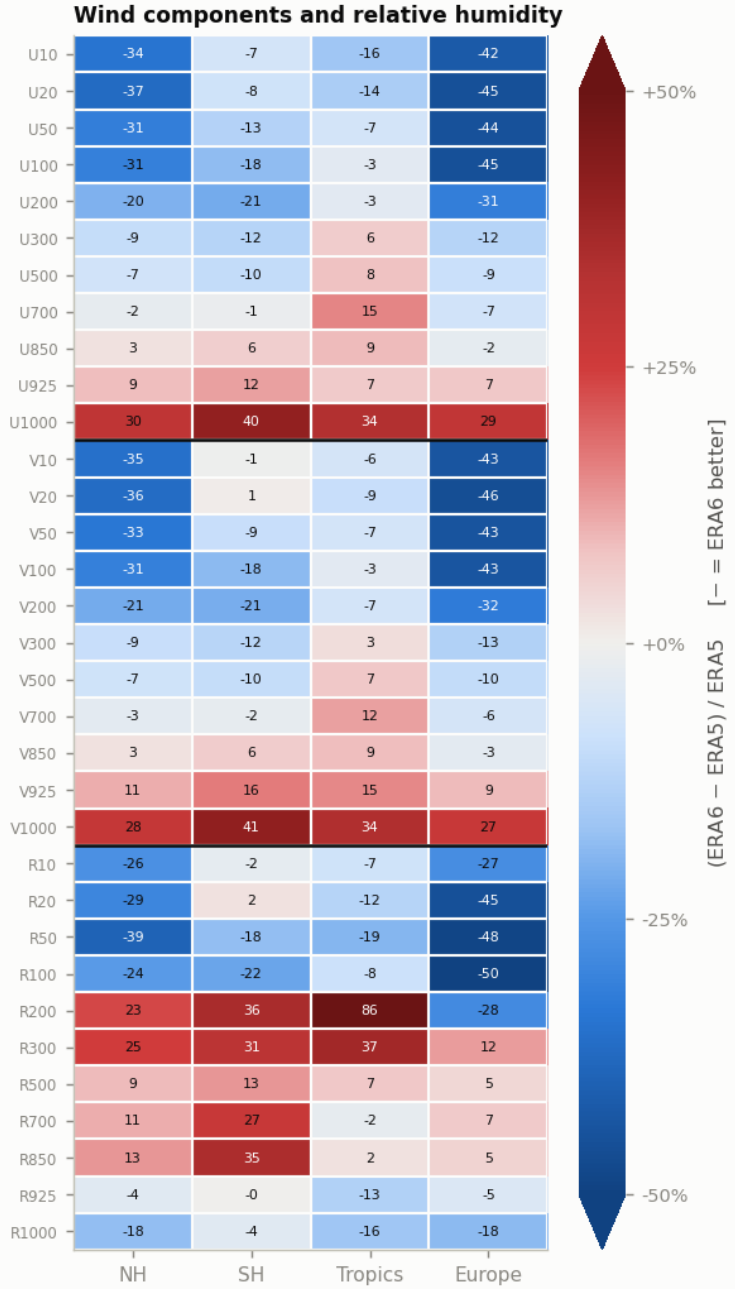
# Discontinuities between 4D-Var windows

$\frac{\Delta(\text{ERA6}) - \Delta(\text{ERA5})}{\Delta(\text{ERA5})}$ at 09/21 UTC,
measured over 4-17 January and 4-17 July, 2017

ERA6 is better
ERA5 is better

Better means lower mean absolute discontinuities.

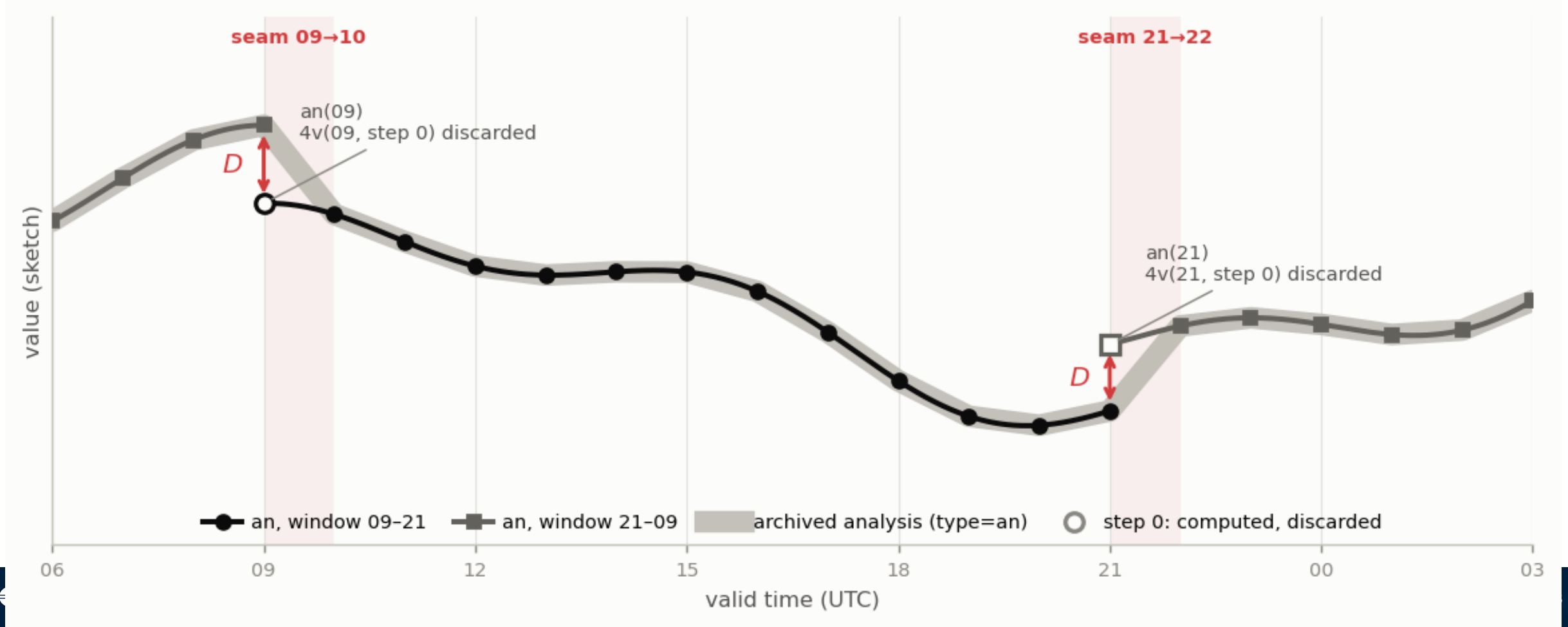



Every row uses 4v(step 0) against the archived analysis. Area-weighted means on a common N320 grid. NH = 20-90N, SH = 20-90S, Tropics = 20S-20N; Europe is a box, 35-70N and 10W-40E.

# Discontinuities between 4D-Var windows

- Potential solutions to reduce the magnitude of the jumps
  - further relax the model as a hard constraint [watch the talk by Patrick, and Alban and Marcin]
  - longer windows (non-overlapping or overlapping) of length up to 24 hours
  - Even longer overlapping windows (Hakim et al., 2026) with pure 4D-Var and its gradient (no TL/AD!)



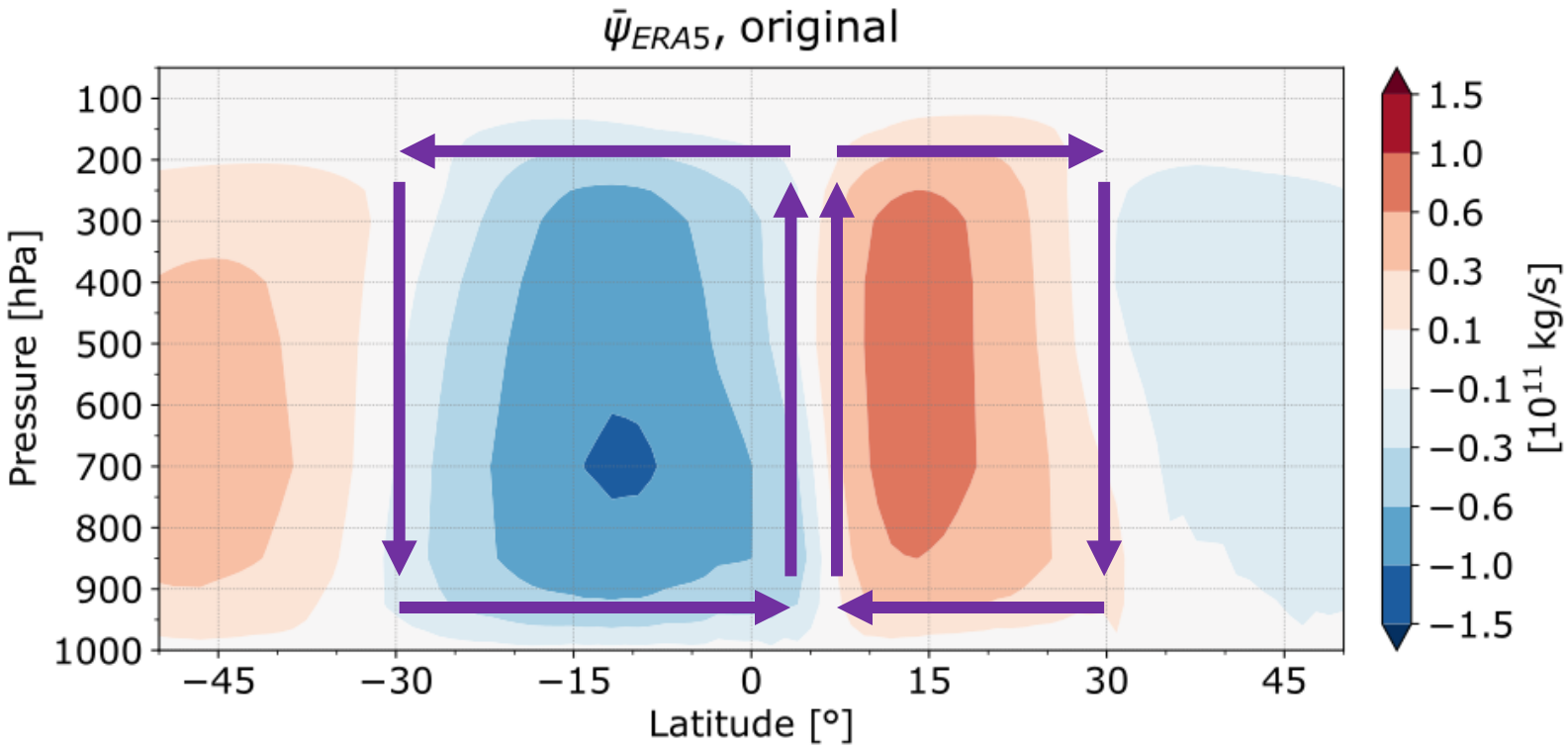
# Weaknesses of current reanalysis systems

## Mean-state biases in Hadley circulation

# Mean state biases in Hadley circulation

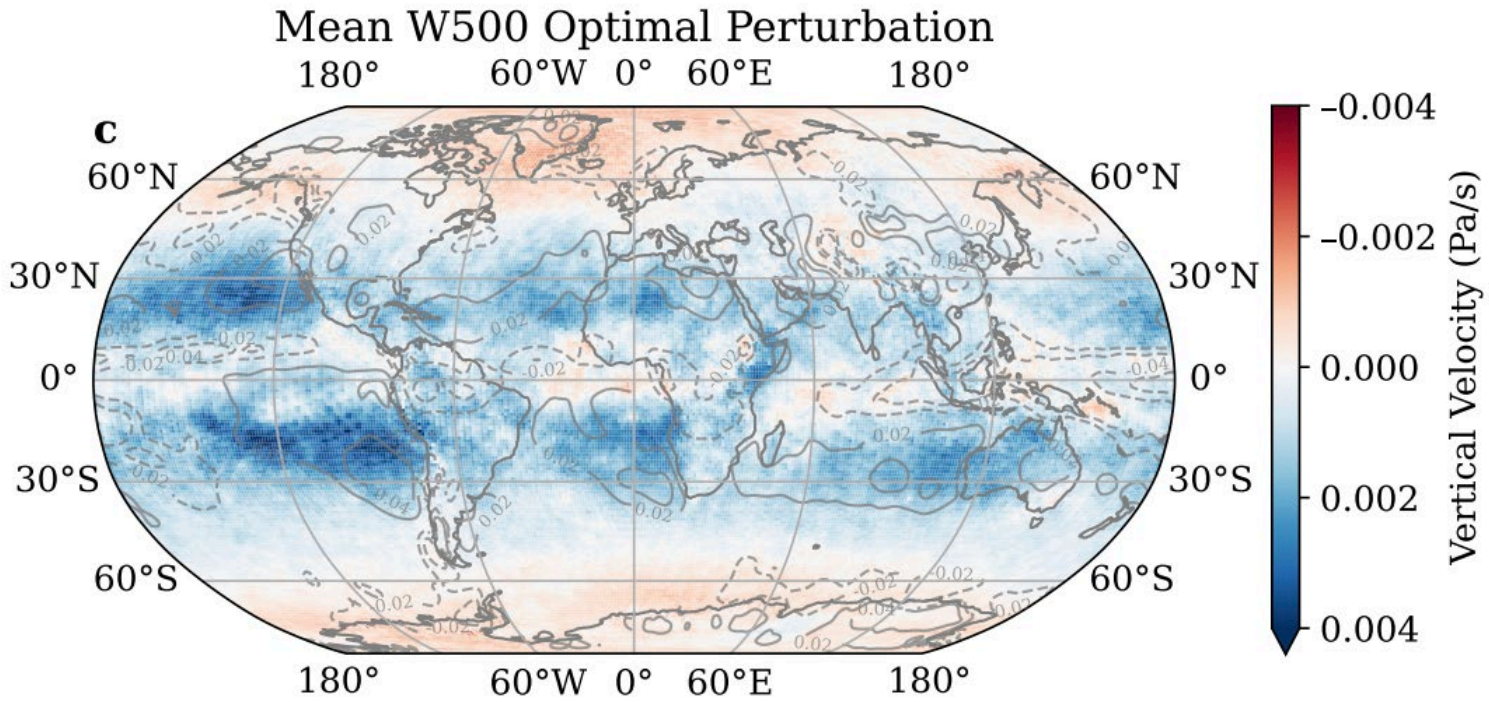
- Hadley circulation encompasses more than half of the Earth's surface.
- The location and magnitude of the ascent/descent area governs the precipitation distribution in the Tropics and Subtropics
- Stream-function representation of the Hadley circulation: $\psi(\varphi, p) = \frac{2\pi R \cos\varphi}{g} \int_0^p [v](\varphi, p')\,\mathrm{d}p'$

1980-2024
annual mean



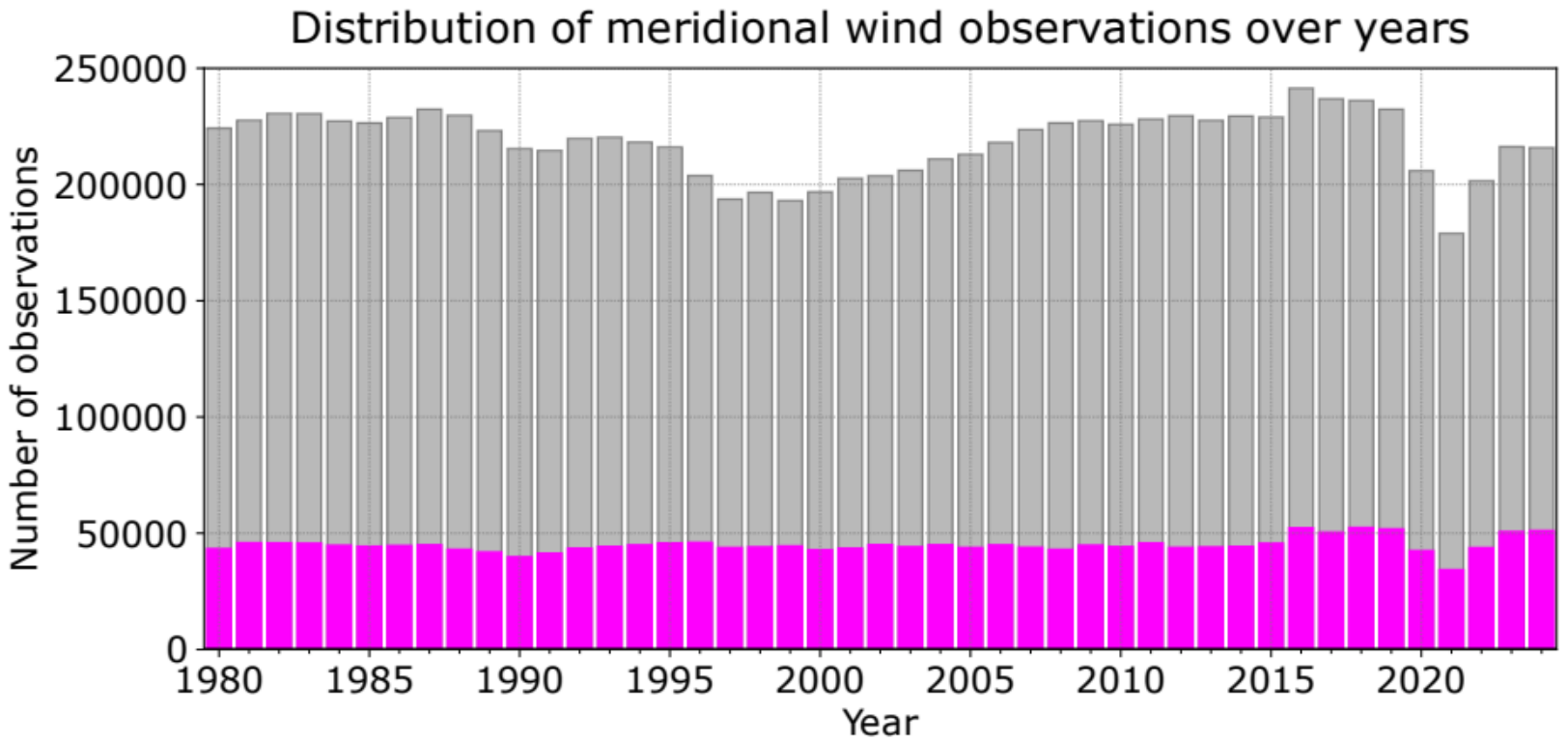
# Mean state biases in Hadley circulation

- Hadley circulation encompasses more than half of the Earth's surface.
- The location and magnitude of the ascent/descent area governs the precipitation distribution in the Tropics and Subtropics
- Enhancing the Hadley circulation is the optimal perturbation to increase NWP prediction skill (Vonich and Hakim, 2026)



# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The radiosonde network has remained approximately constant over the past 45 years, therefore our estimate is not susceptible to the changes in the global observing system, which affect reanalysis mean state/trends



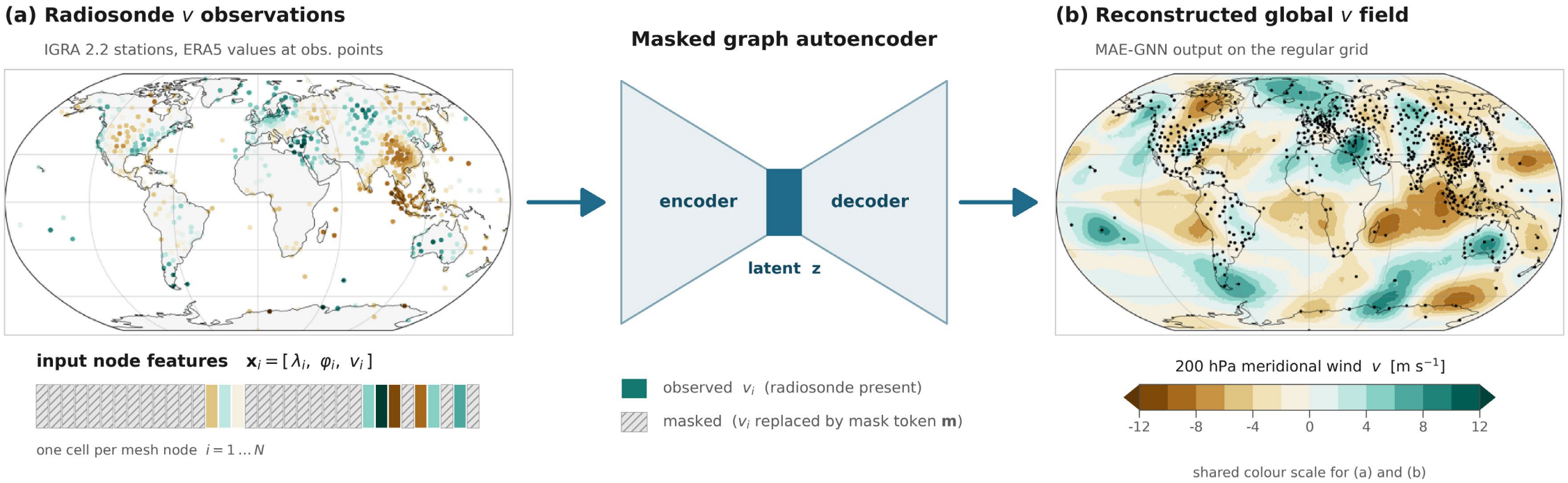

All radiosonde obs

Radiosonde obs. within HC

Pikovnik & Zaplotnik (2026). Radiosonde-constrained reconstruction reveal a weakening Northern Hadley circulation. arXiv:2503.05331 (in review at Nature Communications)

# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.



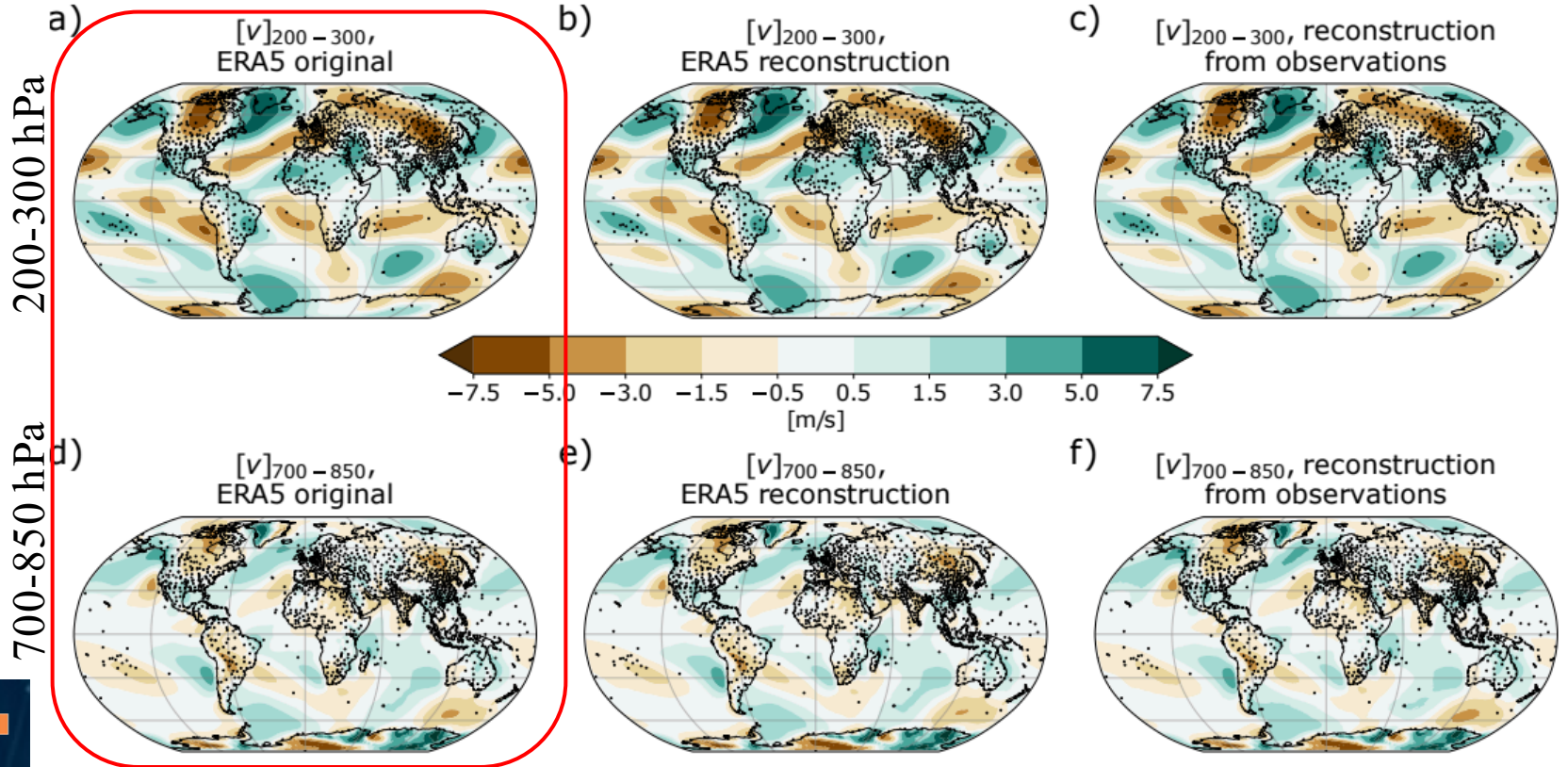

Pikovnik & Zaplotnik (2026). Radiosonde-constrained reconstruction reveal a weakening Northern Hadley circulation. arXiv:2503.05331 (in review at Nature Communications)

# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!

Mean meridional winds in 1980-2024 in

[LEFT] original ERA5 reanalysis

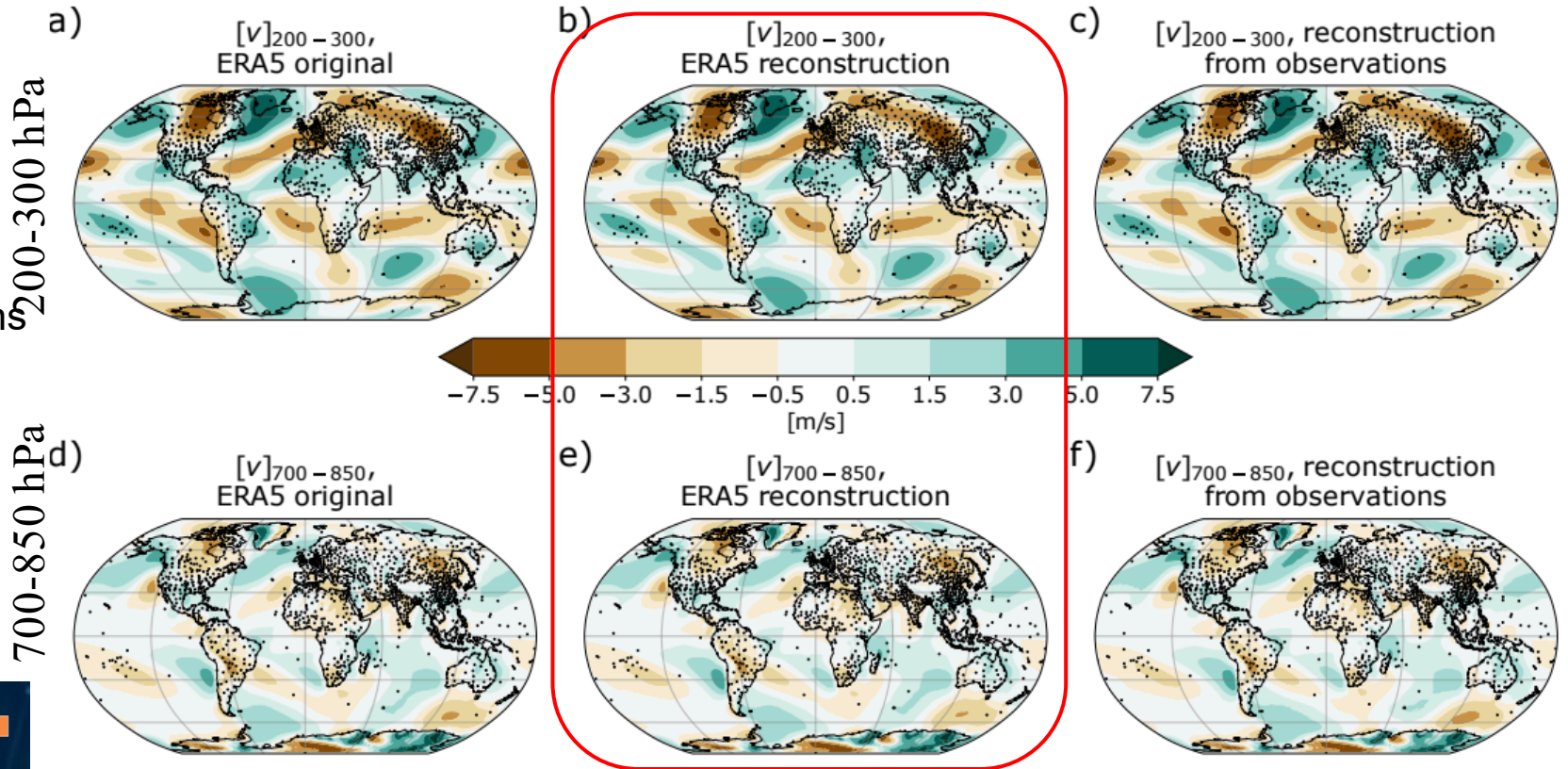


# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!

Mean meridional winds in 1980-2024 in

[CENTER] MAE reconstructions from ERA5 values at the observation locations

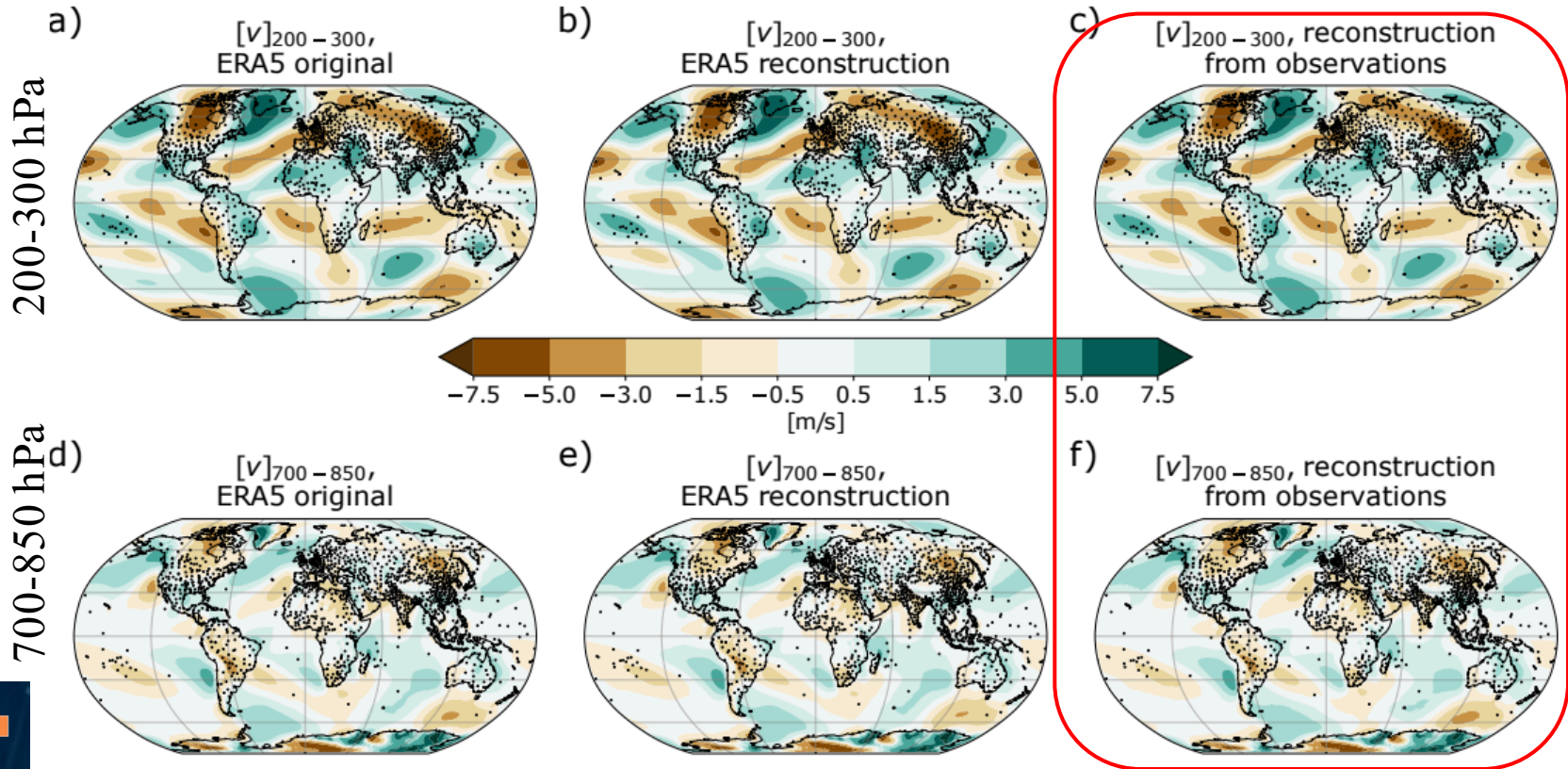


# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!

Mean meridional winds in 1980-2024 in

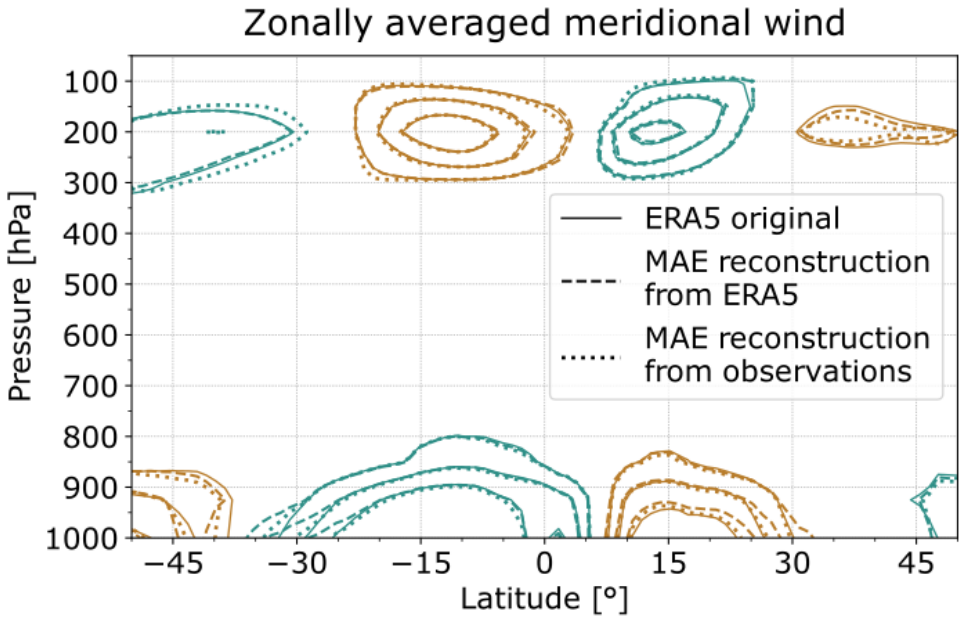
[RIGHT] MAE reconstructions from radiosonde observations

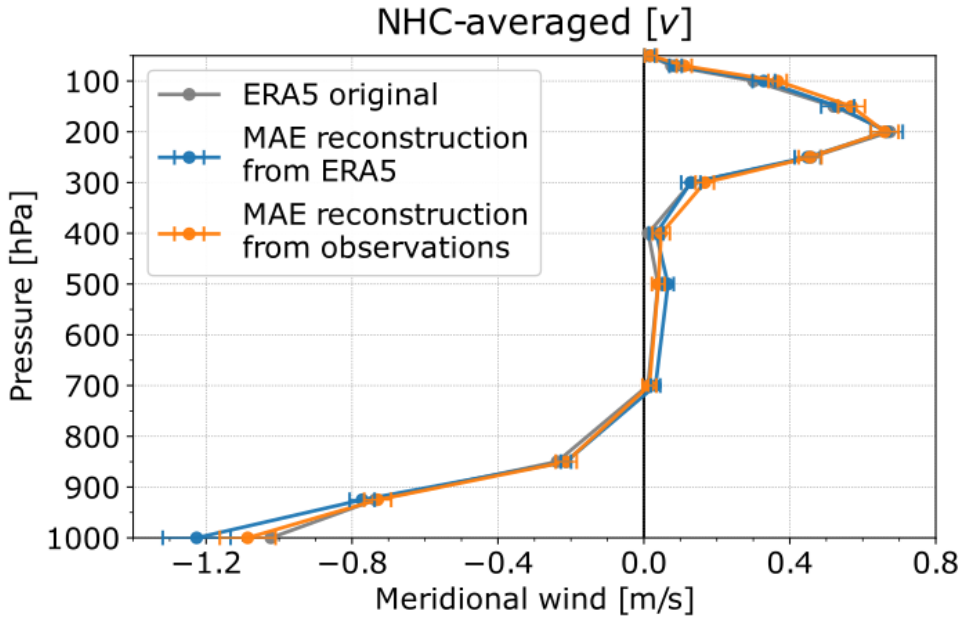


# Mean state biases in Hadley circulation

- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!

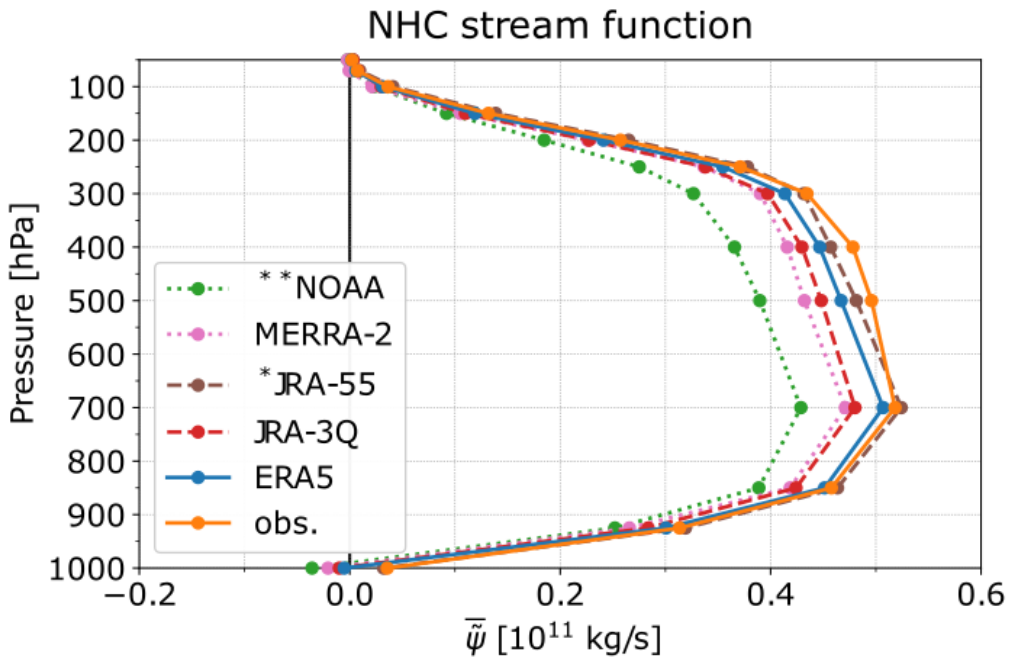
Mean meridional winds in 1980-2024, latitude-pressure cross-section

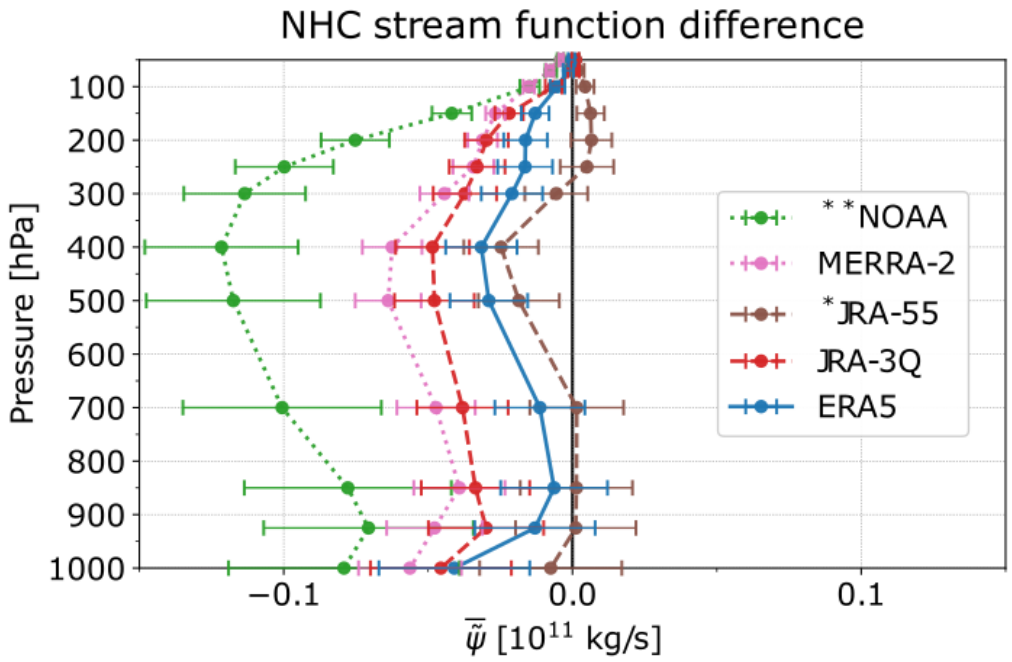


# Mean state biases in Hadley circulation

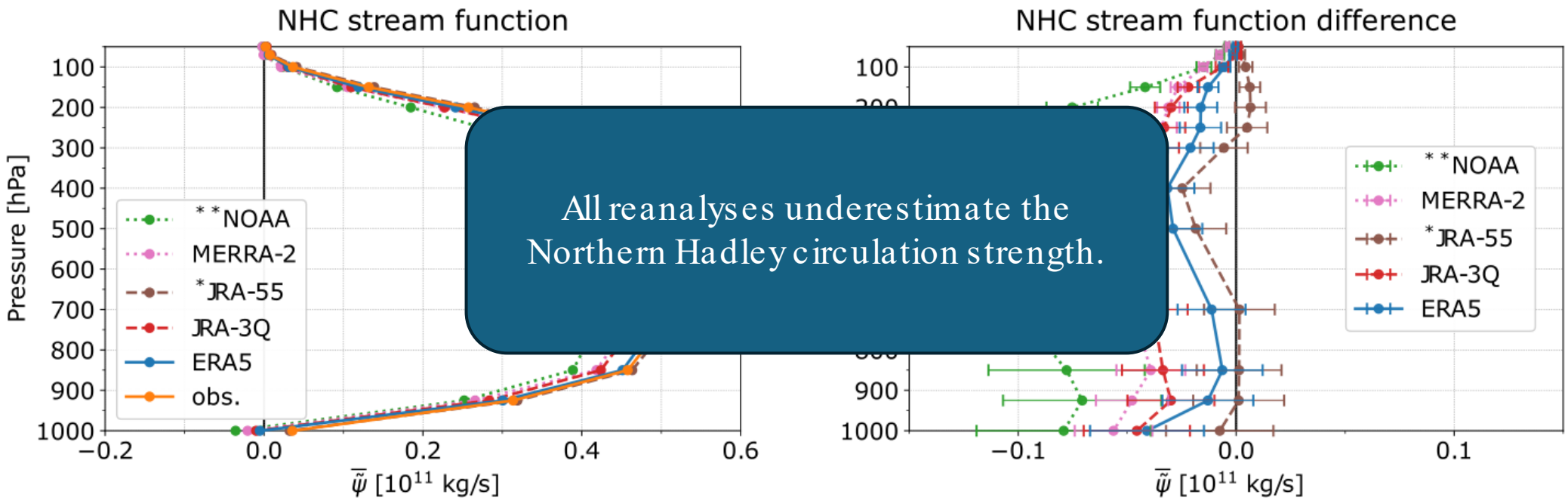
- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!



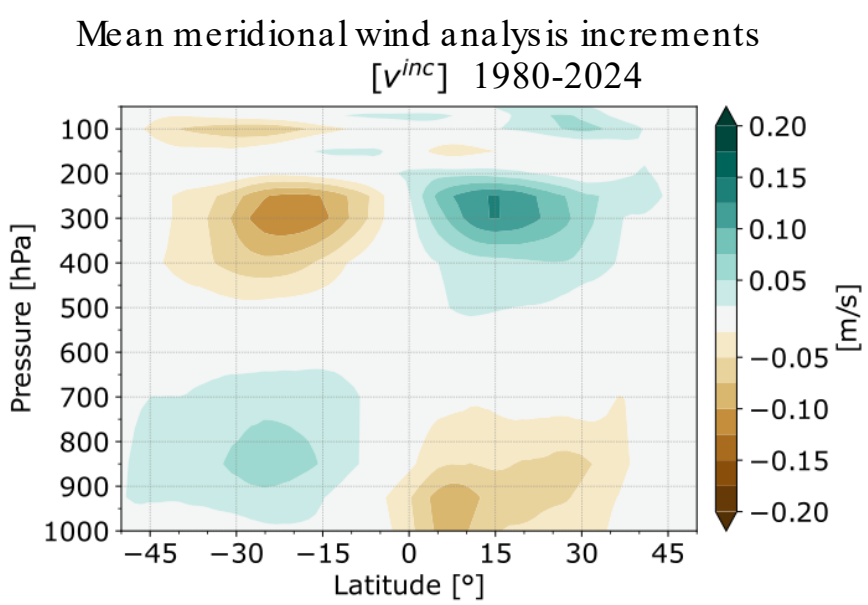
# Mean state biases in Hadley circulation

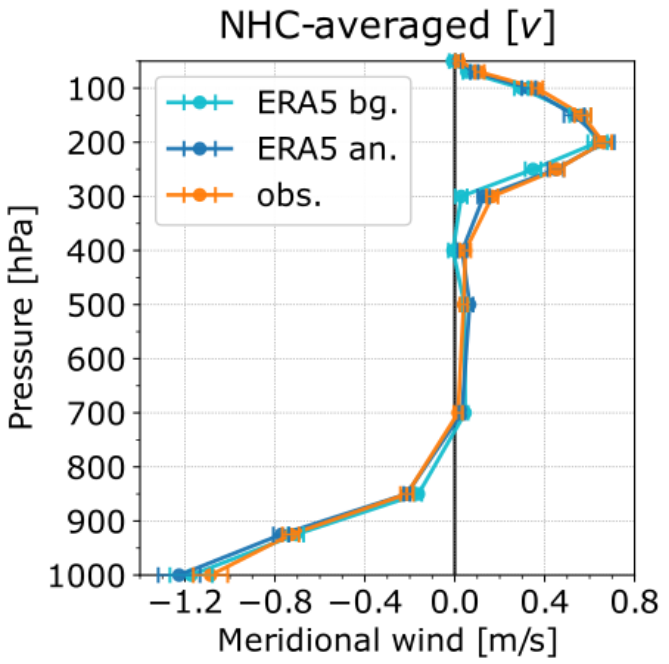
- To estimate HC strength, we reconstruct 3D meridional wind fields from sparse radiosonde observations (IGRAv2)
- The procedure is as follows:
  1) We trained a masked autoencoder (MAE) based on a graph-neural network (GNN) to reconstruct monthly-mean ERA5 fields from ERA5 values at radiosonde observation locations.
  2) Once the MAE-GNN was trained, we input the real monthly-mean radiosonde observations.
- The reconstructions are accurate, but they also reveal some differences!

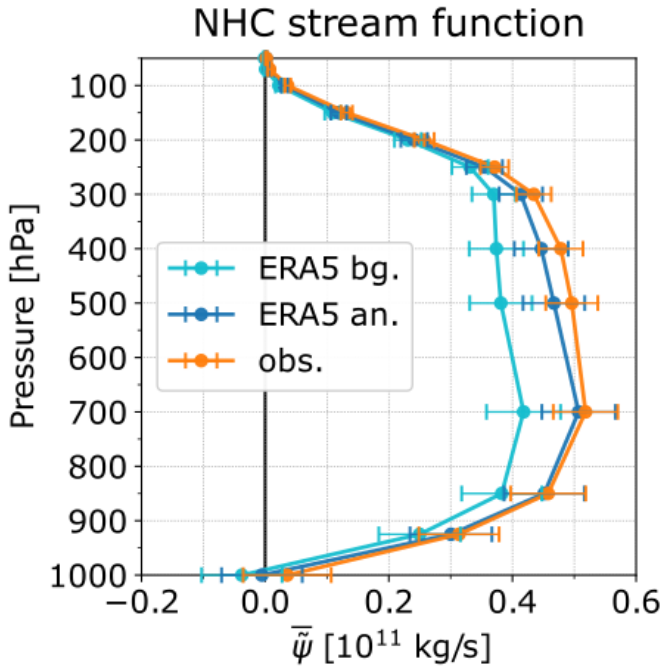


# Mean state biases in Hadley circulation

- Why are we confident in these results?
- Another line of evidence: at each analysis cycle (throughout the whole ERA5 reanalysis), the data assimilation intensifies the Hadley circulation and brings the analysis closer to observations



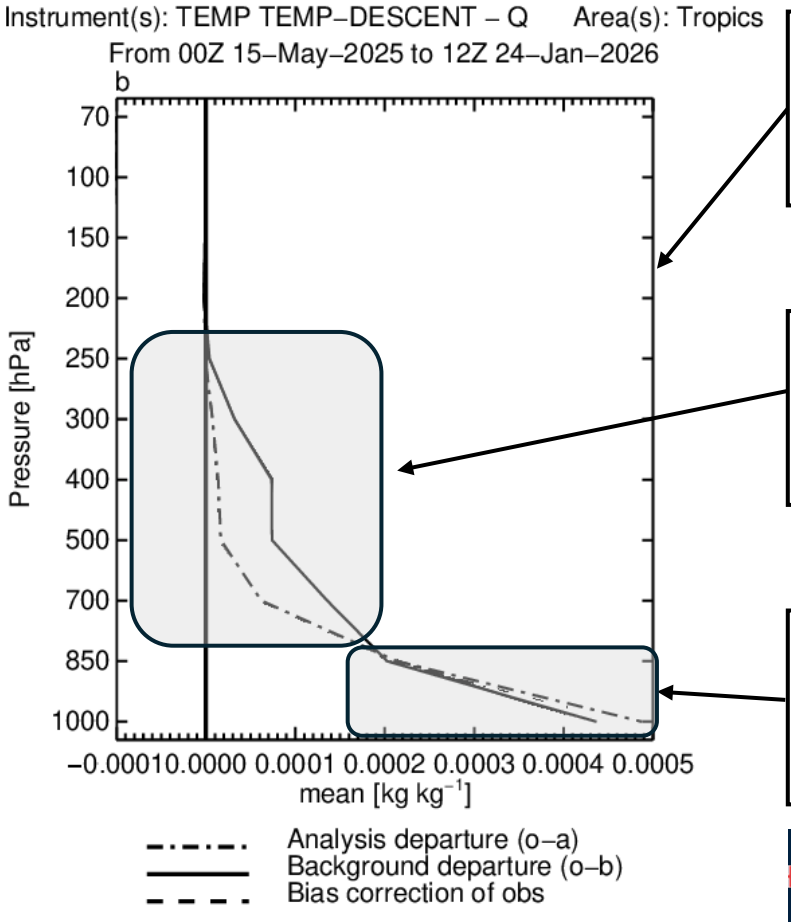




Pikovnik & Zaplotnik (2026). Radiosonde-constrained reconstruction reveal a weakening Northern Hadley circulation. arXiv:2503.05331 (in review at Nature Communications)

# Mean state biases in Hadley circulation

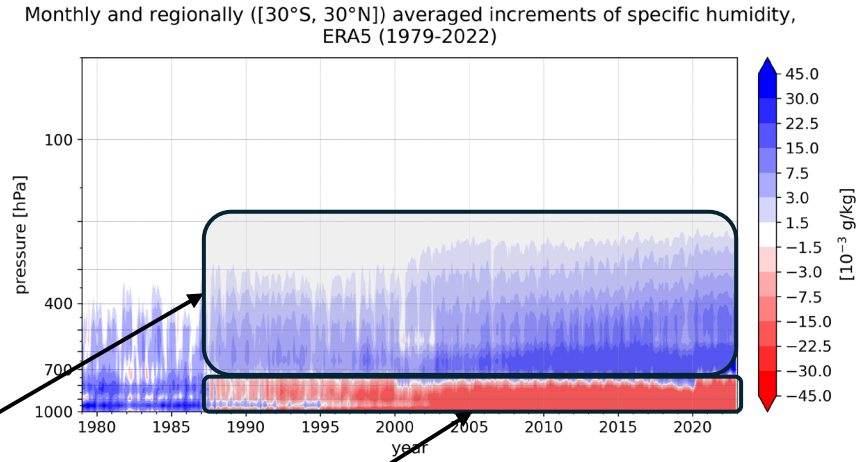
- Why are we confident in these results?
- At each analysis cycle (throughout the whole ERA5 reanalysis), the data assimilation intensifies the Hadley circulation and brings the analysis closer to observations ← another line of evidence
- What is the source of the HC strength bias? Current evidence points to a combination of model and DA bias.

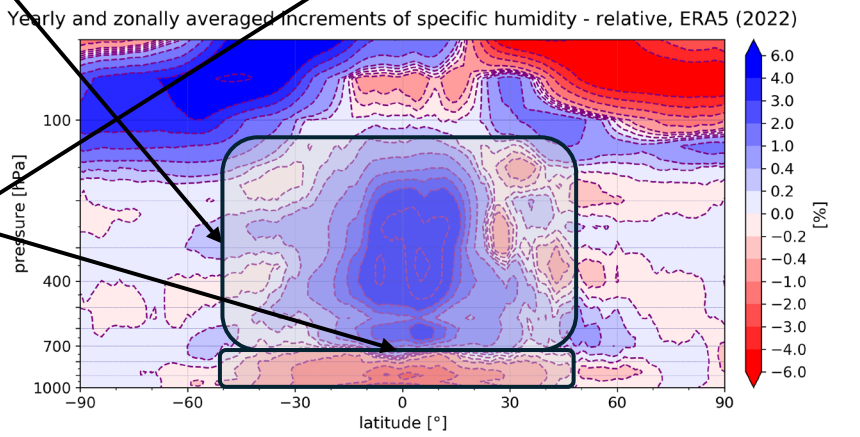



Every IFS experiment has a dry bias in both the background/analysis in the tropical troposphere.

The DAmoistens mid- and upper-troposphere (as it should based on O-B!)

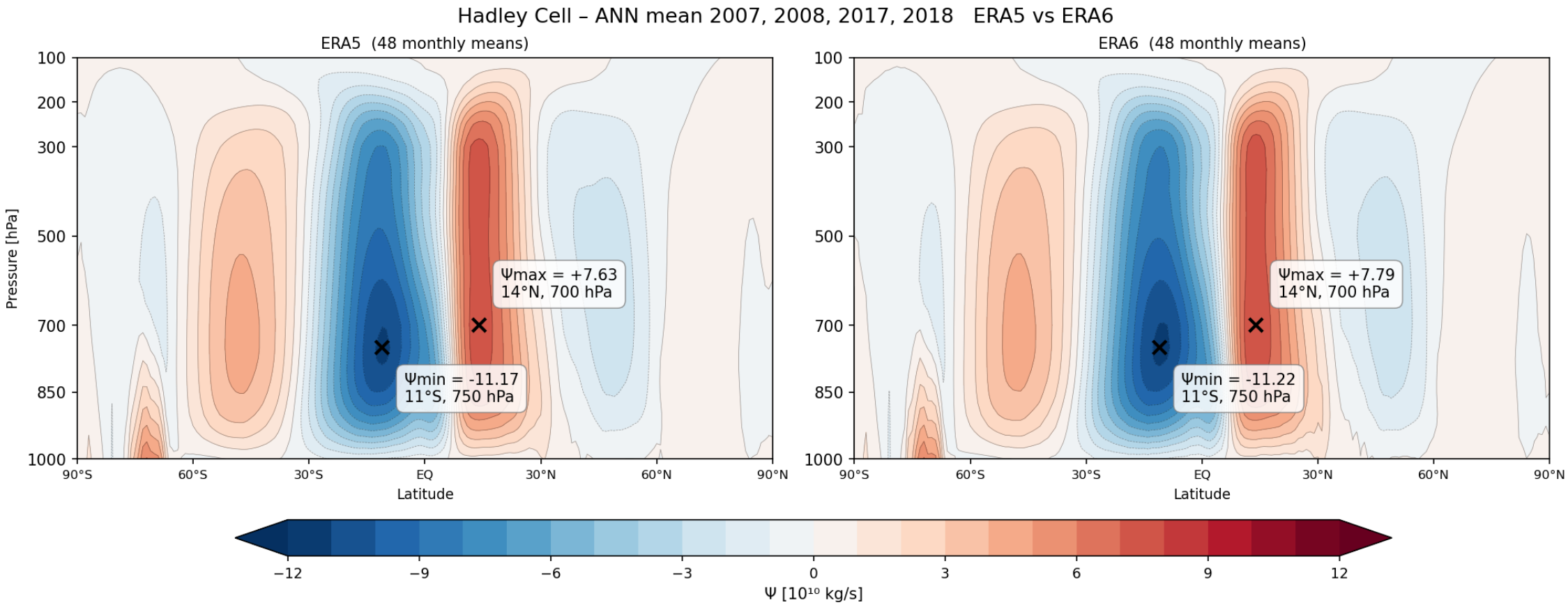
The DAdries the boundary layer. But it should not, it further intensifies the bias!



# Mean state biases in Hadley circulation

According to our study, ERA5 underestimates NHC by around 10%.
ERA6 intensifies it by around 2%.

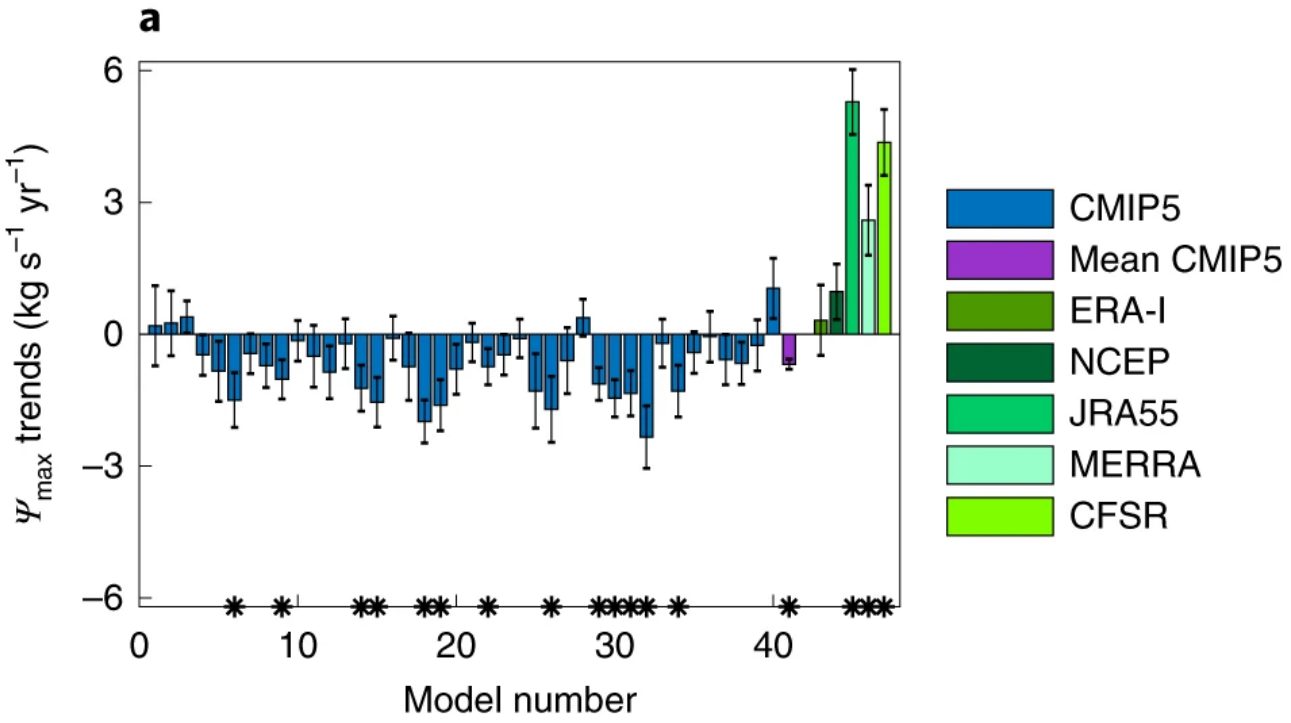


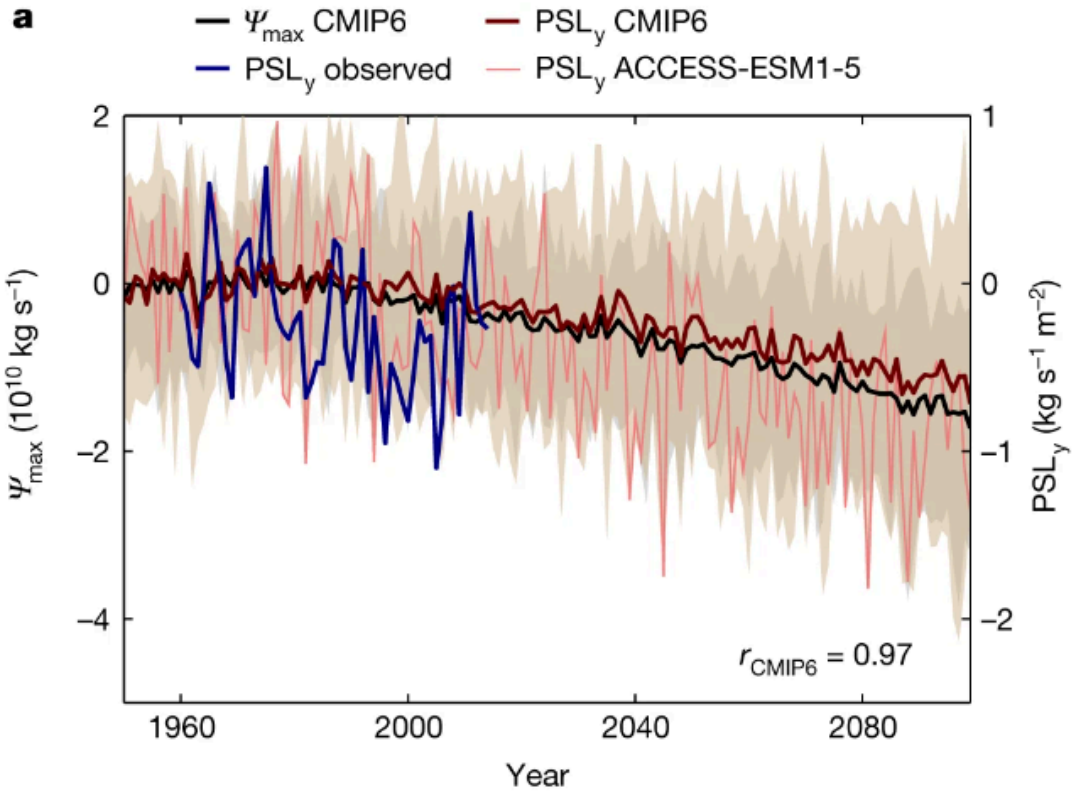
# Weaknesses of current reanalysis systems

## Miscaptured climate trends

# Miscaptured sensitive climate trends

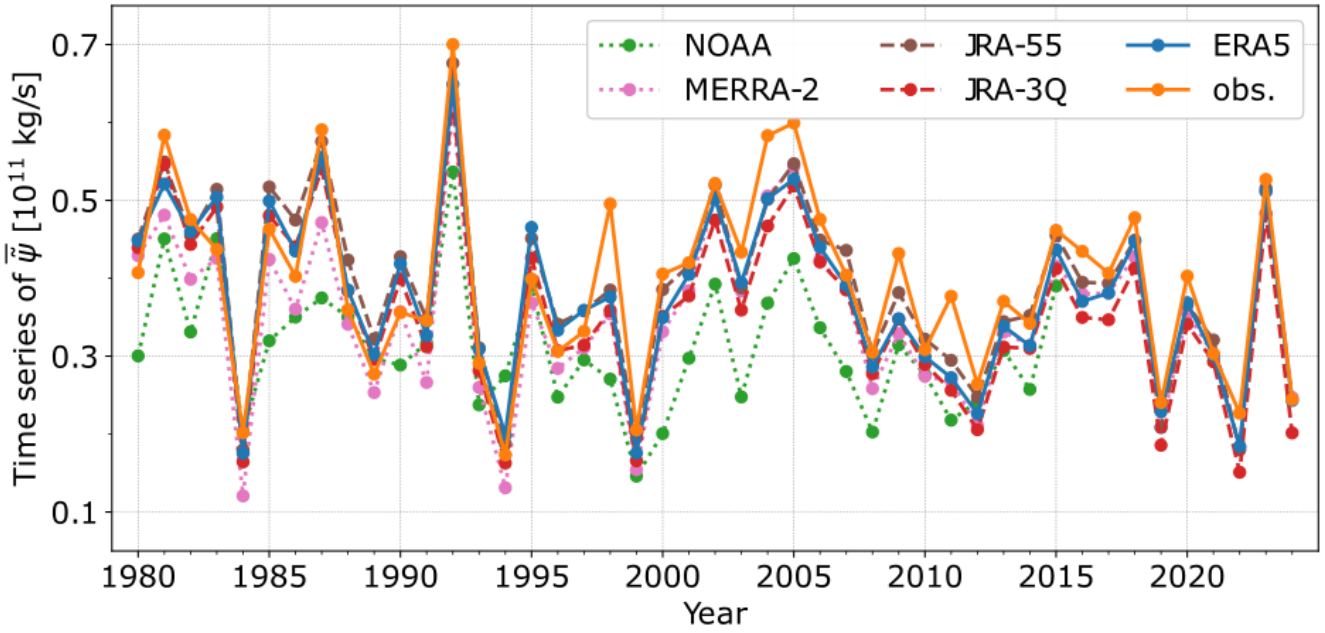
- Raw reanalyses mostly indicate strengthening of Northern Hadley cell (including ERA5)
- This is in contrast to climate models (Chemke and Polvani, 2019, Nature Geoscience)
- This is also in contrast to observation proxies (ICOADS SLP, Chemke and Yuval, 2023, Nature) $\Psi_{500} \propto \frac{1}{|f|}\overline{\frac{\partial \mathrm{PSL}}{\partial y}}$

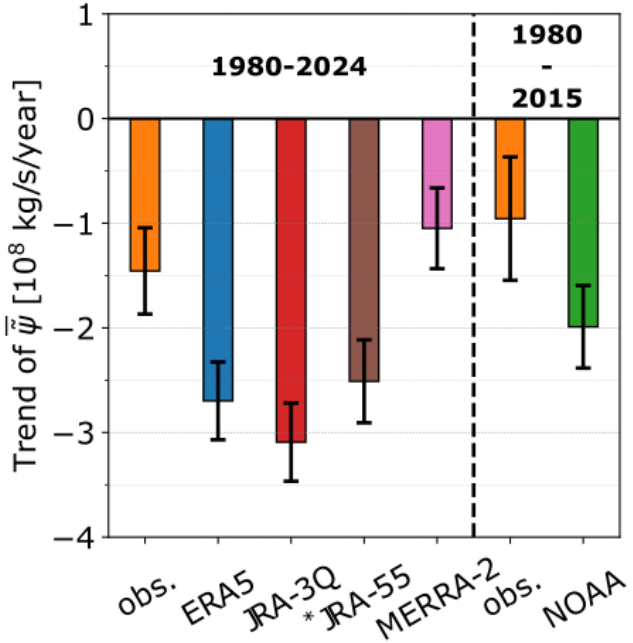


# Miscaptured sensitive climate trends

- Raw reanalyses mostly indicate strengthening of Northern Hadley cell (including ERA5)
- This is in contrast to climate models (Chemke and Polvani, 2019, Nature Geoscience)
- This is also in contrast to observation proxies (ICOADS SLP, Chemke and Yuval, 2023, Nature) $\Psi_{500} \propto \frac{1}{|f|}\overline{\frac{\partial \mathrm{PSL}}{\partial y}}$
- Circulation trends in the reanalyses are susceptible to rapidly changing global observing system → employing a temporally consistent set of radiosonde observations relaxes the problem
- Radiosonde-only reanalysis might still be needed!



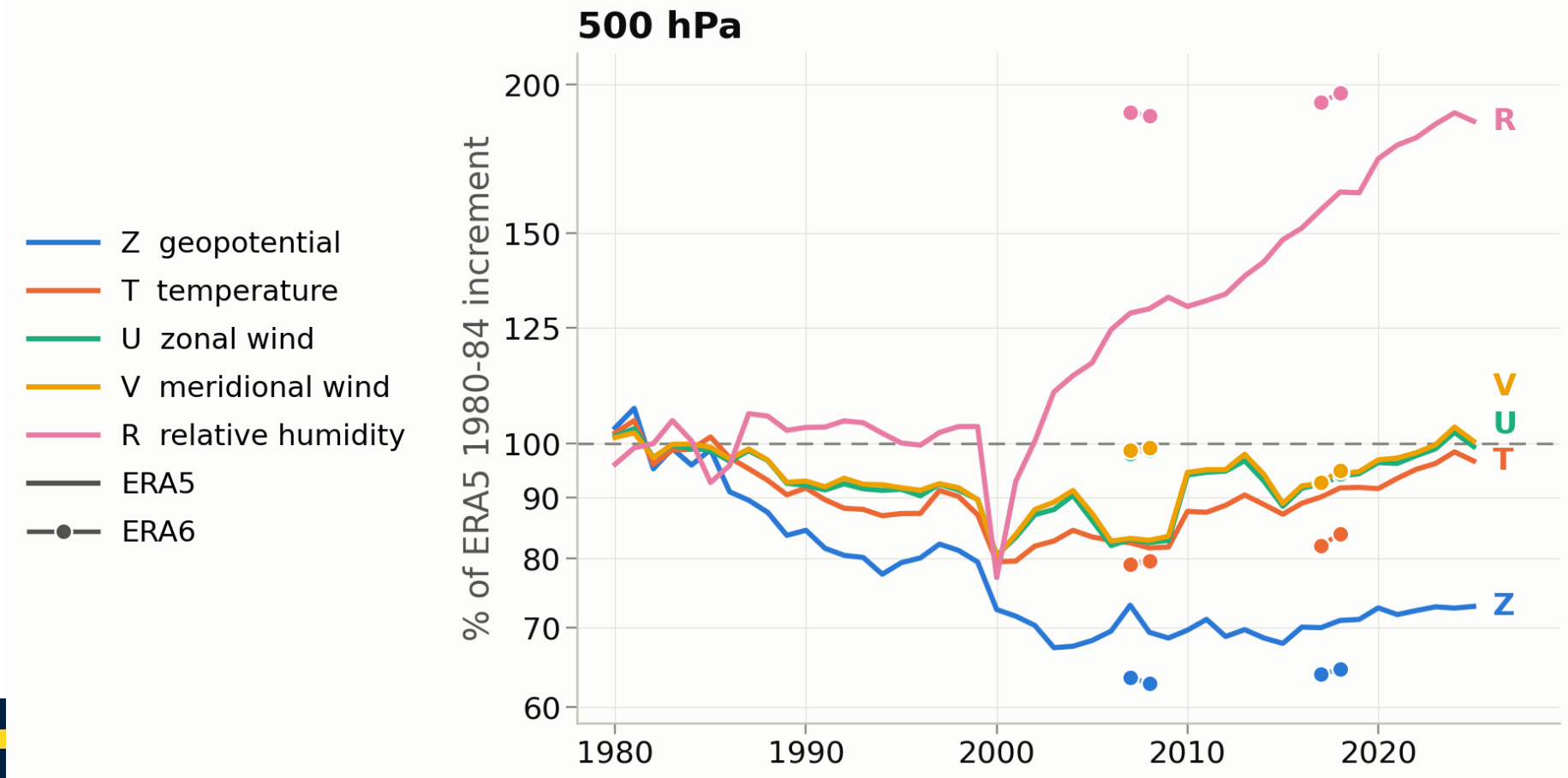
# Weaknesses of current reanalysis systems

## Growing magnitude of analysis increments

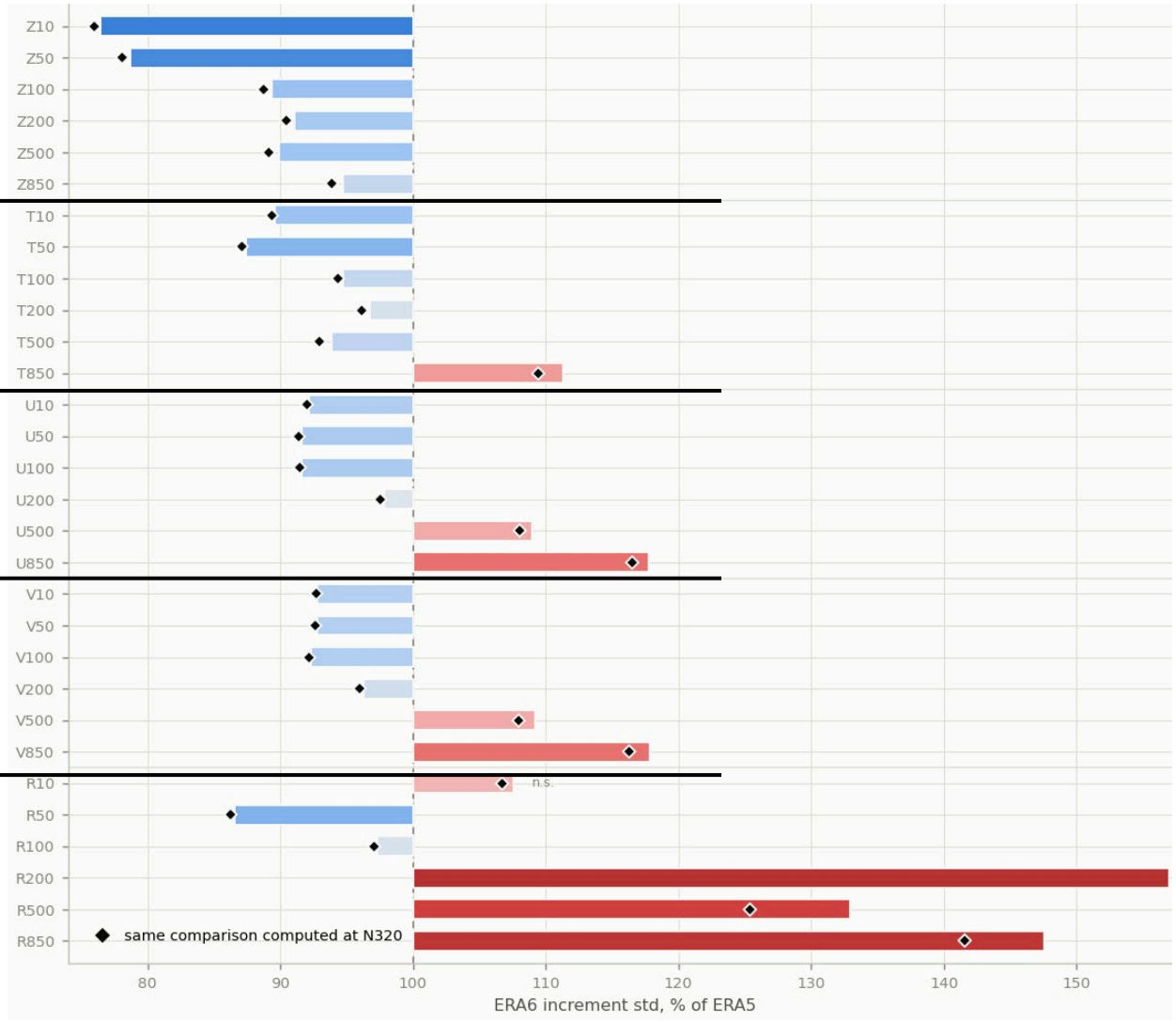
# The increments should be getting smaller with time?

The magnitude of ERA5 analysis increments diminished until ~2010 for u,v,Z, since then on, all are increasing
→ increasing number of obs?
But it is increasing also until 2010. Besides, this should better constrain the background (closer to truth).

→ The real atmospheric dynamics is increasingly different to the (constant) model dynamics.
Shifting attractor due to climate change ?



# ERA6 vs ERA5 analysis increments magnitude




ERA6 has smaller increments than ERA5

ERA6 has larger increments than ERA5

Magnitude of increments is measured as a standard deviation of the field and averaged over common periods (2007, 2008, 2017, 2018)

# Weaknesses of current reanalysis systems

- Representation of extreme events (deep tropical cyclones, intense precipitation)
- Discontinuities between 4D-Var windows
- Mean-state biases in primary atmospheric circulation
- Misrepresented sensitive climate trends

- NWP reanalysis separated from atmospheric-composition (CAMS) reanalysis
- Weakly-coupled data assimilation in different Earth-system components

# DA methodologies to further improve the reanalyses

Longer data assimilation windows

# Longer windows




## CORE OBJECTIVES

Reduce diurnal-cycle inconsistencies and analysis discontinuities for Climate and AI/ML training.

Strengthen coupling Atmosphere-Land, Atmosphere-Ocean

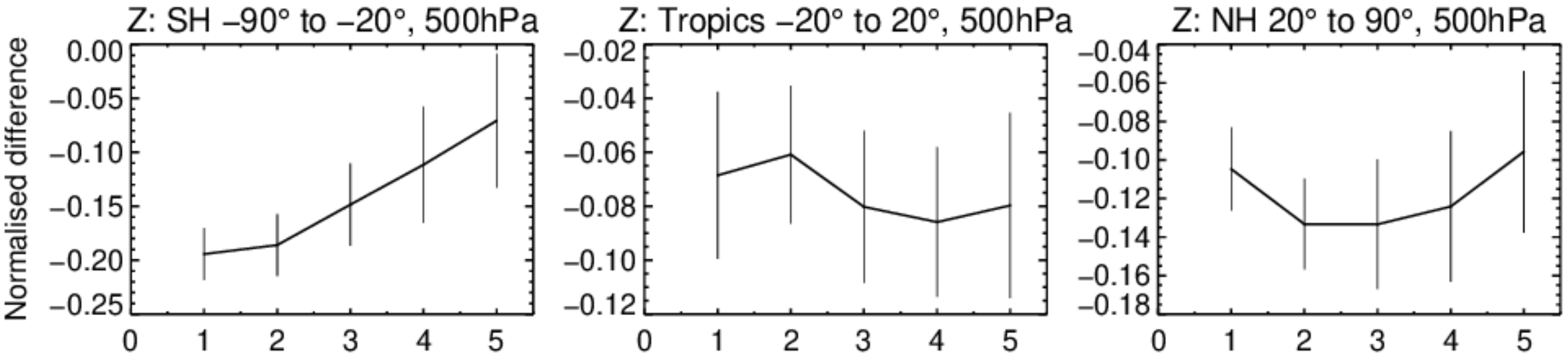
Extract more information from observations through 4D-Var internal adjustment and tracing.

Large forecast-skill benefit from common initial times.

# Longer windows

- Comparison of forecast skill from analysis using 24-hour 4D-Var window and

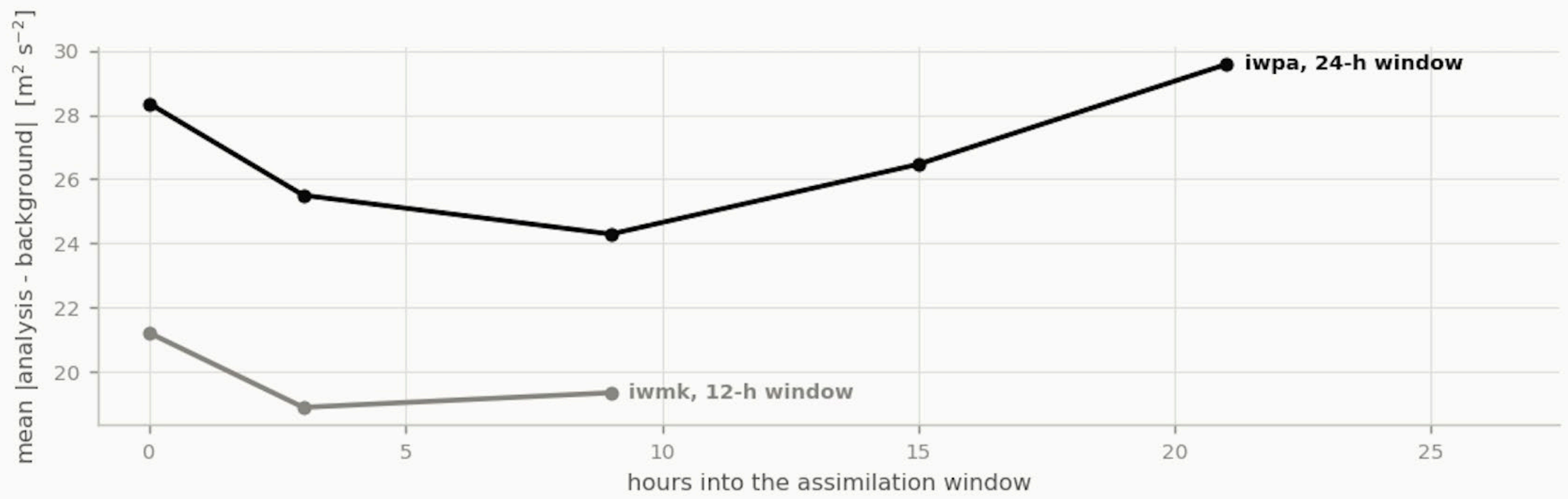
- Reduction in RMSE: **negative** means 24h window is better




- This is an early test, lots of room for improvement:
  - Background-error variances should be retuned
  - anoffset=21, however the forecasts should start closer to the middle of the 4D-Var window (→ next slide)

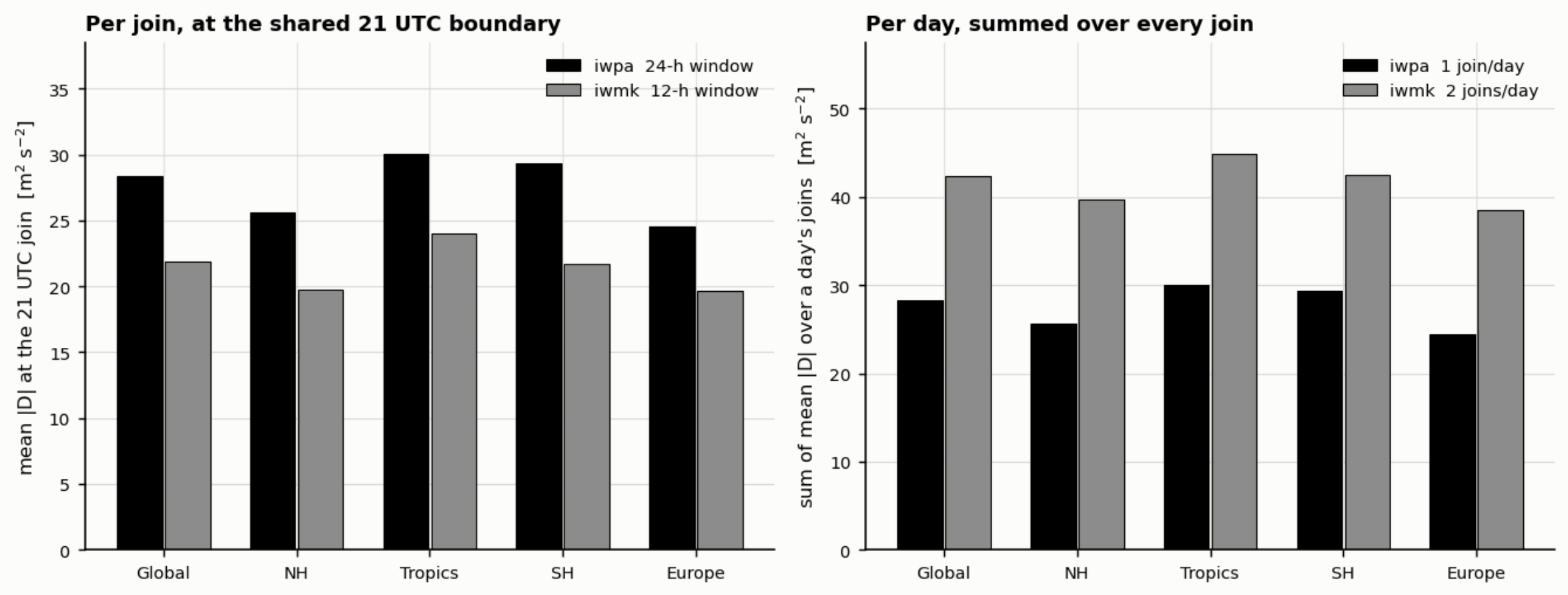
# Longer windows

- The discrepancy of the 24-hour analysis against the background is larger than the discrepancy of the 12-hour window throughout the common time of the 4D-Var window
- The initial time for issuing a forecast likely also moves towards 9 hours into the assimilation window



# Longer windows

- Discontinuities are slightly larger in 24-hour window
- However, the sum of all jumps per day is decreased



# Longer overlapping windows

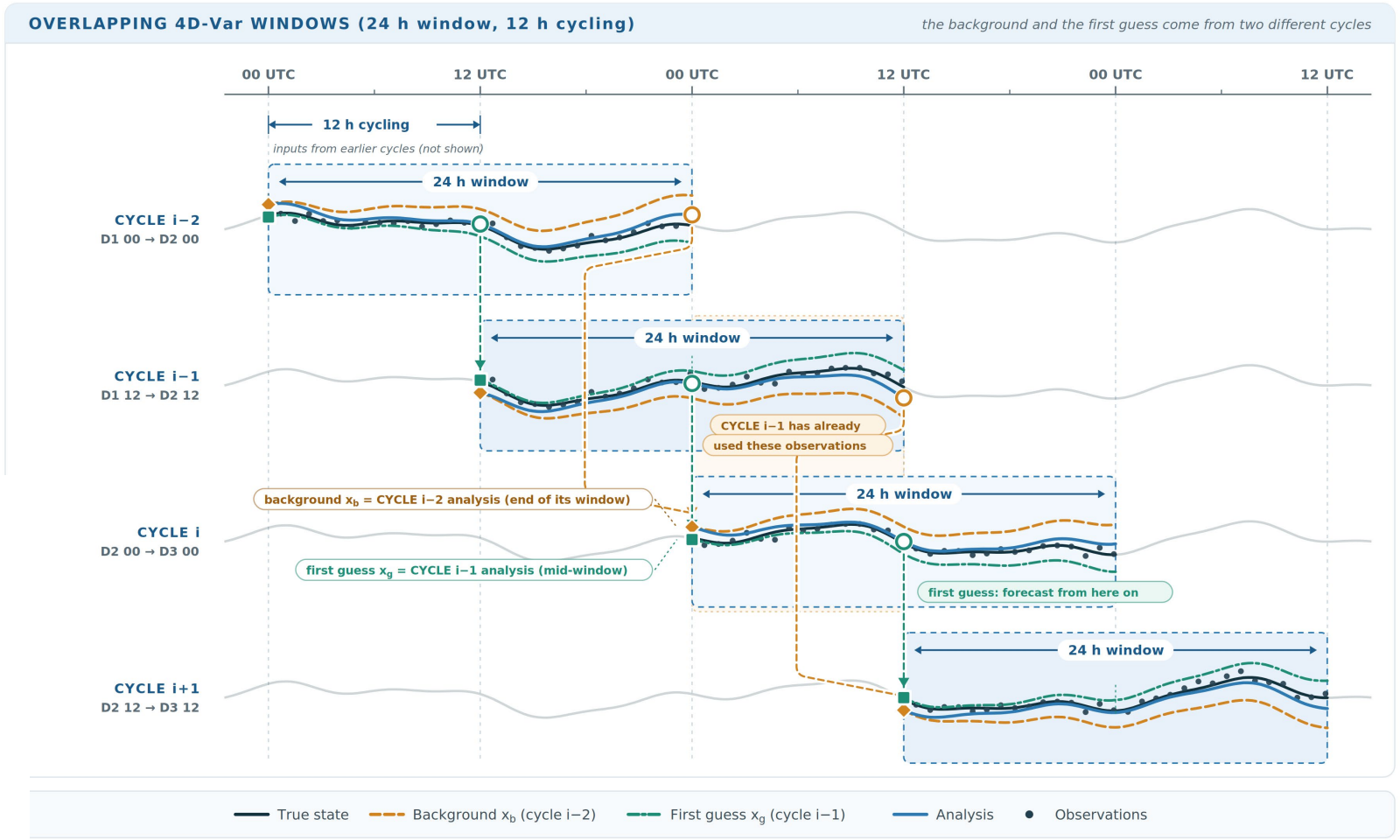
The background for cycle $i$ comes from cycle $i-2$, so that the same window does not see the observations twice.

The first guess for cycle $i$ comes from cycle $i-1$, which ensures less iterations.

Pros:

- Reduced discontinuities
- Less iterations as we start from an improved guess
- Incremental 4D-Var allows $x_g \neq x_b$ with initial $J_b$ term equal to:

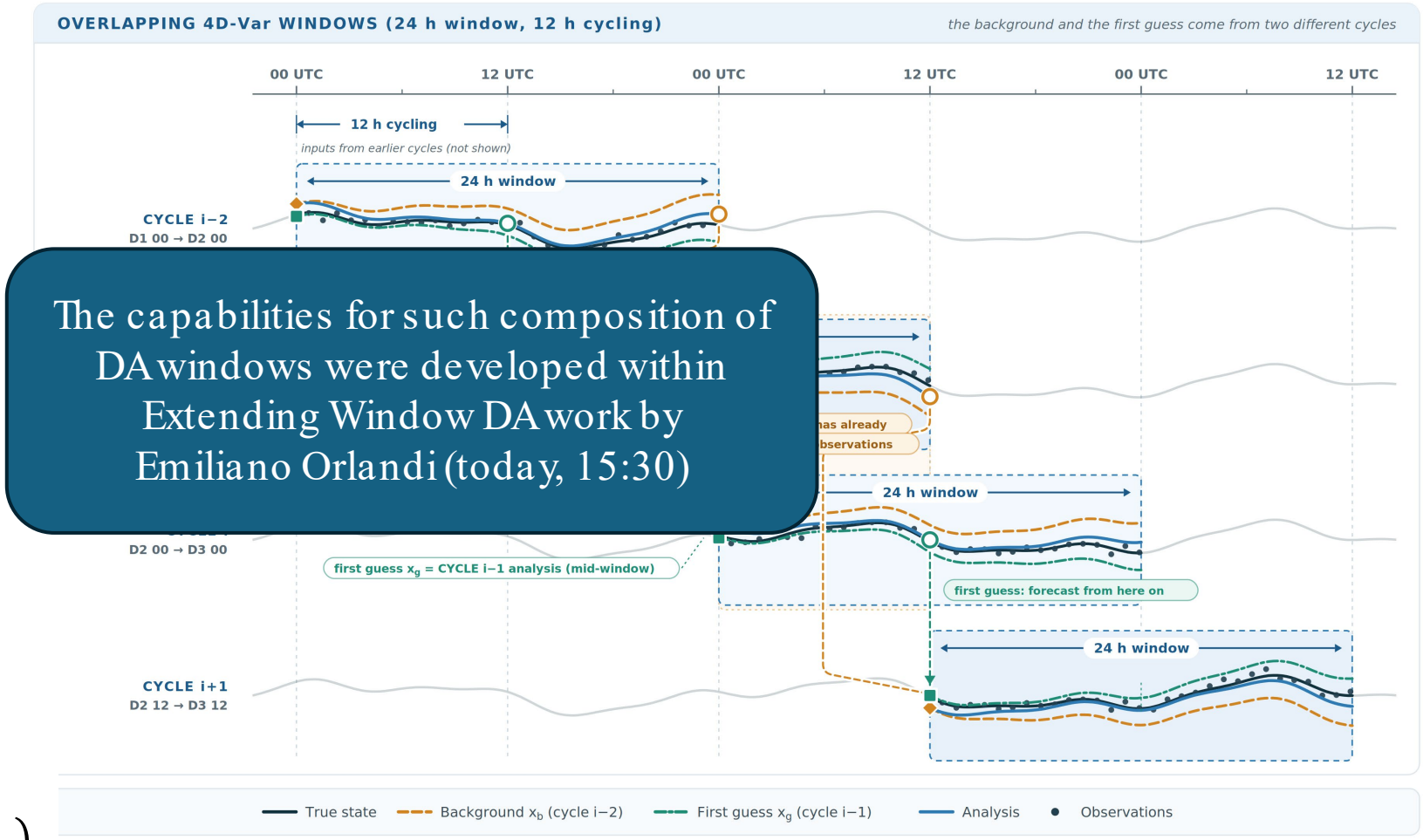
$$J_b = \frac{1}{2}\left(x_g + \delta x - x_b\right)^T B^{-1}\left(x_g + \delta x - x_b\right)$$



# Longer overlapping windows

The background for cycle $i$ comes from cycle $i-2$, so that the same window does not see the observations twice.

The first guess for cycle $i$ comes from cycle $i-1$, which ensures less iterations.




The capabilities for such composition of DA windows were developed within Extending Window DA work by Emiliano Orlandi (today, 15:30)

Pros:

- Reduced discontinuities
- Less iterations as we start from an improved guess
- Incremental 4D-Var allows $x_g \neq x_b$ with initial $J_b$ term equal to:

$$J_b = \frac{1}{2}\left(x_g + \delta x - x_b\right)^T B^{-1}\left(x_g + \delta x - x_b\right)$$

# Latent space data assimilation

We have established a theory of latent space data assimilation for NWP DA in several papers:

- Melinc and Zaplotnik (2024), 3D-Var data assimilation using a variational autoencoder, QJRMS
- Melinc, Perkan and Zaplotnik (2026), A unified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation, JAMES [within the ECMWF visiting scientist scheme]
- Zaplotnik, Melinc, Kugler, Bouillon, Bonavita (2026, in prep.), Revealing Hidden Atmospheric Balances through Graph-Based Latent Space Data Assimilation

# DA methodologies to further improve the reanalyses

## Latent-space data assimilation

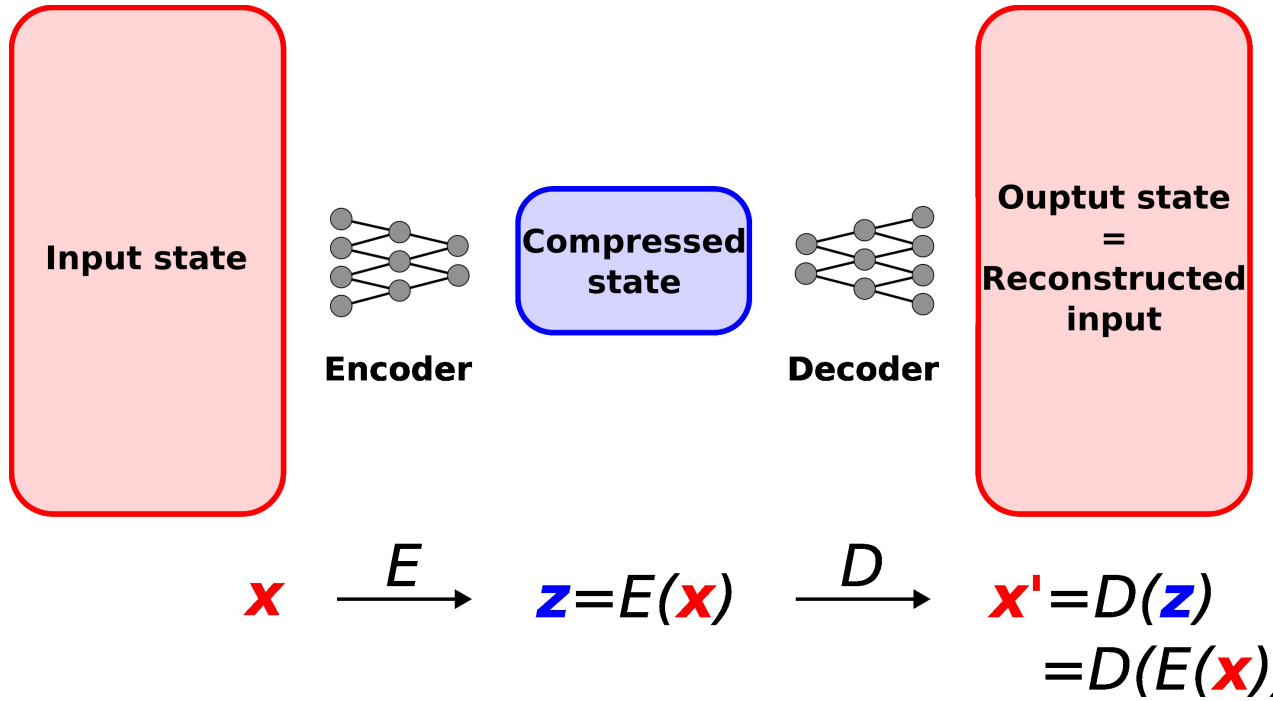
# Latent space data assimilation

We have established a theory of latent space data assimilation for NWP DA in several papers:

- Melinc and Zaplotnik (2024), 3D-Var data assimilation using a variational autoencoder, QJRMS
- Melinc, Perkan and Zaplotnik (2026), A unified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation, JAMES [within the ECMWF visiting scientist scheme]
- Zaplotnik, Melinc, Kugler, Bouillon, Bonavita (2026, in prep.), Revealing Hidden Atmospheric Balances through Graph-Based Latent Space Data Assimilation

**Core idea:**

- The autoencoder encodes atmospheric states into a reduced latent space (data compression; analogue of PCA, ICA, but nonlinear)
- Perform 3D-Var data assimilation in the latent-space of a (variational) autoencoder



# Latent space data assimilation

We have established a theory of latent space data assimilation for NWP DA in several papers:

- Melinc and Zaplotnik (2024), 3D-Var data assimilation using a variational autoencoder, QJRMS
- Melinc, Perkan and Zaplotnik (2026), A unified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation, JAMES [within the ECMWF visiting scientist scheme]
- Zaplotnik, Melinc, Kugler, Bouillon, Bonavita (2026, in prep.), Revealing Hidden Atmospheric Balances through Graph-Based Latent Space Data Assimilation

**Core idea:**

- The autoencoder encodes atmospheric states into a reduced latent space (data compression; analogue of PCA, ICA, but nonlinear)

- Perform 3D-Var data assimilation in the latent-space of a (variational) autoencoder

$$\mathcal{J}_z(\mathbf{z}) = \mathcal{J}_{bz} + \mathcal{J}_{oz} = \\ = \frac{1}{2}(\mathbf{z}-\mathbf{z}_b)^{\mathrm{T}}\mathbf{B}_z^{-1}(\mathbf{z}-\mathbf{z}_b) + \frac{1}{2}[\mathbf{y}-H\{D(\mathbf{z})\}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y}-H\{D(\mathbf{z})\}]$$

$$\mathbf{x}_a = S^{-1}D(\mathbf{z}_a)$$

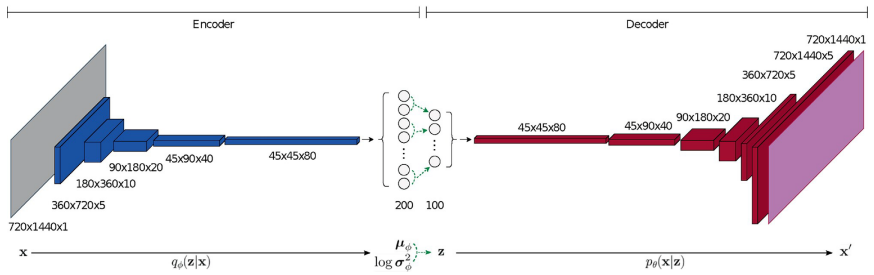
# Latent space data assimilation

Requires a “good” autoencoder:

- Properly reconstructs the initial fields
- Performs a significant data compression (O(10x) to O(100x))
- Learns to decorrelate latent space components (spatially, spectrally, feature-wise)
- … and many more (follow the talk on Thursday at 13.30 on latent space DA)

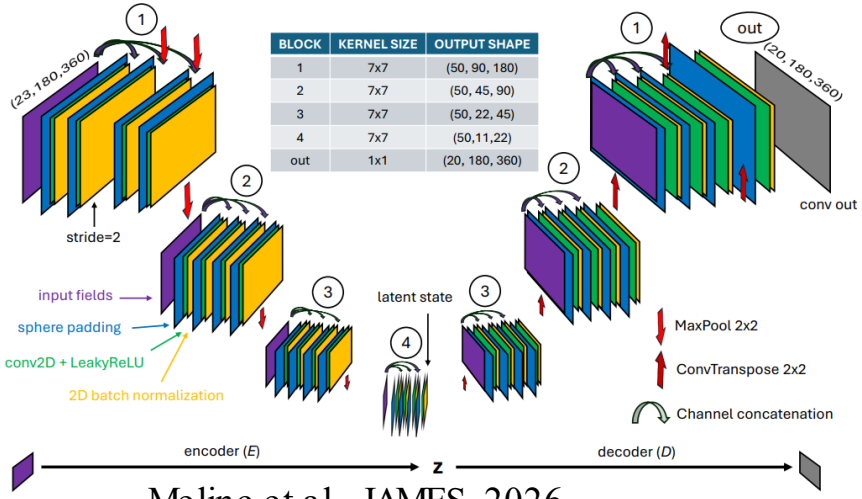
TIMELINE

Proof of concept: simple univariate CNN




Melinc and Zaplotnik, QJRMS, 2024

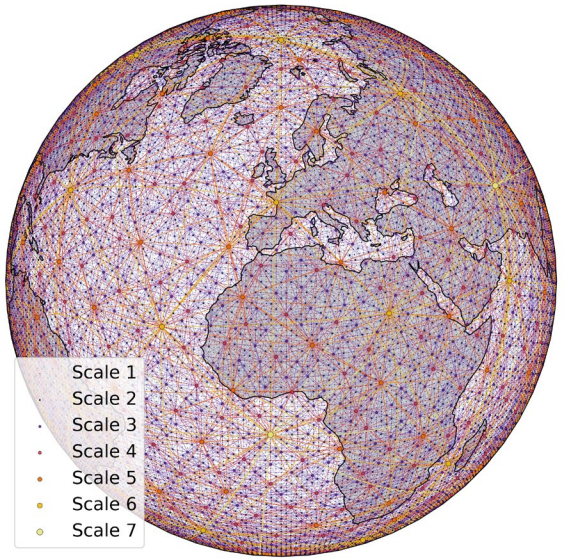
Multivariate multilevel CNN




| BLOCK | KERNEL SIZE | OUTPUT SHAPE |
|---|---|---|
| 1 | 7x7 | (50, 90, 180) |
| 2 | 7x7 | (50, 45, 90) |
| 3 | 7x7 | (50, 22, 45) |
| 4 | 7x7 | (50,11,22) |
| out | 1x1 | (20, 180, 360) |

Melinc et al., JAMES, 2026

Fully-coupled GNN

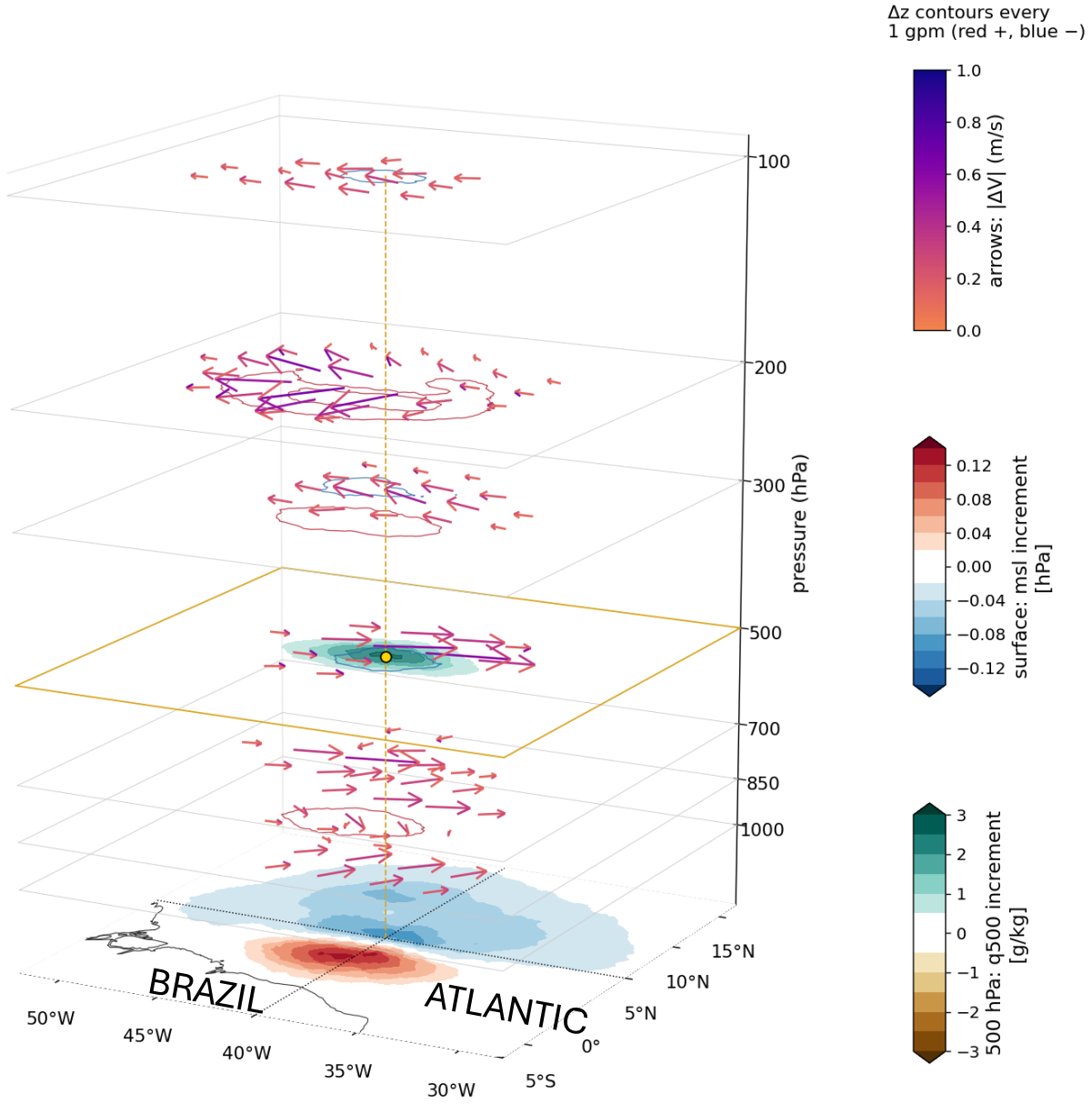



Zaplotnik et al., 2026 (in prep.)

# Latent space data assimilation – single obs experiment

A single observation experiment assimilating synthetic observation of 500 hPa specific humidity with +4g/kg departure on 2024-04-14 at 5N 40W (north of the coast of Brazil):

- Localised horizontally
- Physically-consistent dynamic and thermodynamic response (e.g. pressure-difference driven friction flow at the bottom
- Gill-like response to heat-induced perturbations (upper-divergence and lower-level convergence)



# Fully-coupled 4D-Var latent-space data assimilation

Wavelet-based B →ML latent representation of coupled covariances

Cross-Earth system increments: Assimilation of T2m observation with −5K departure above the Bering Sea in two distinct sea ice concentration (SIC) background scenarios: (top) near-maximum yearly SIC on Northern Hemisphere on March 1, 2023, and (bottom) near-minimum yearly SIC on Northern Hemisphere on September 15, 2023

Moving beyond Outer Loop coupling heuristics!

No localization needed,
no specified boundaries of land-sea-ice,
fully-coupled B by definition

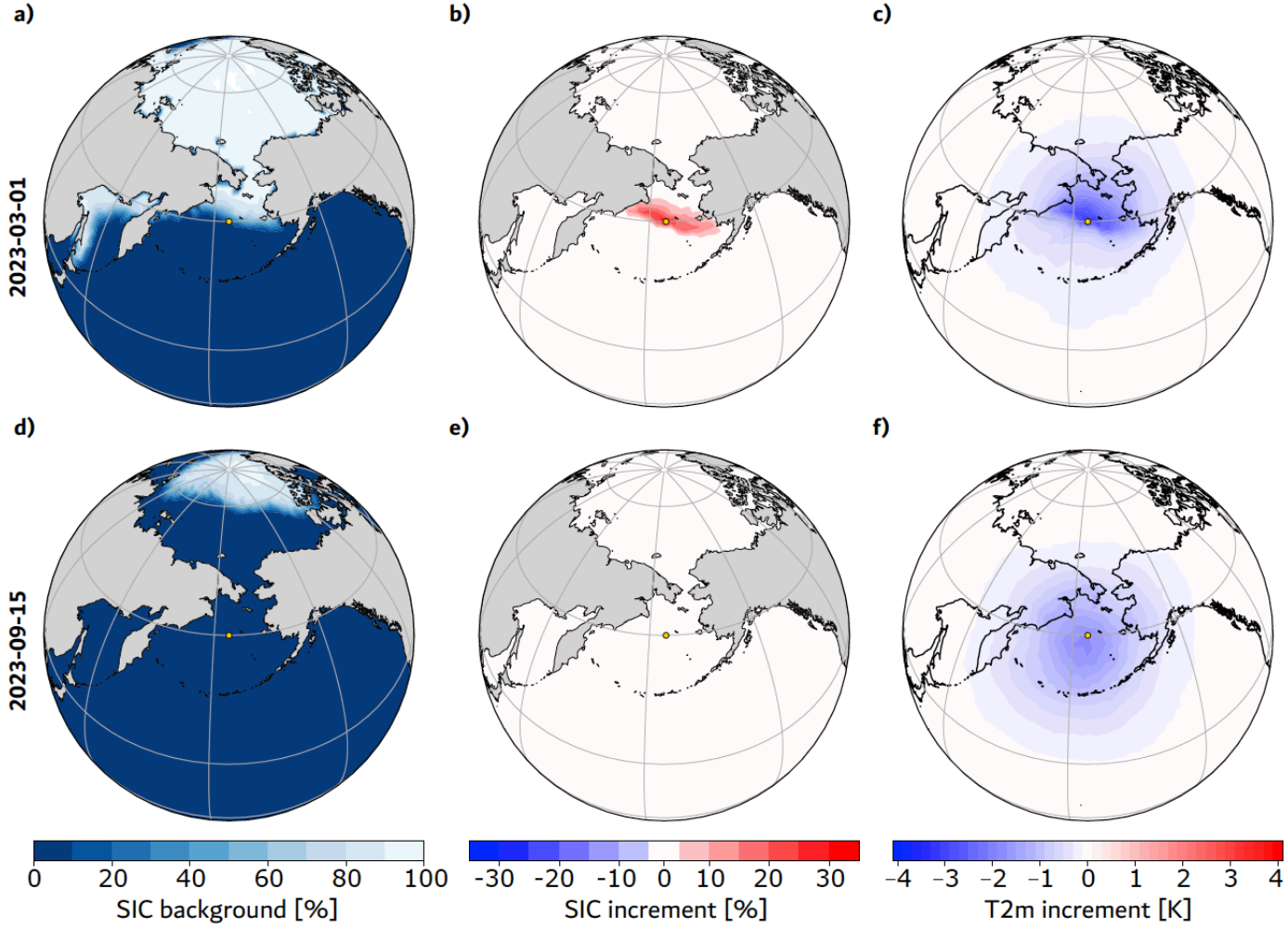



Figure courtesy of Lukas Kugler (University of Ljubljana)

# Latent-space DA

- Allows unified background-error covariance model for the midlatitudes and tropics
- Crucial capability for correcting the representation of tropical climate variables, e.g. strength of Hadley circulation
- Allows fully-coupled atmosphere-ocean-land DA with a common **B**
- Allows unified NWP-air quality DA

# How could a future hybrid-physical Earth-system reanalysis look like?

# How could a future reanalysis look like?

### Overlapping 24h 4DVar
Temporal continuity, stronger coupling, large forecast-skill benefit
Emiliano: Tuesday, 15:30

### Latent-space DA
ML-enabled coupled B modelling
Ziga: Thursday, 13:30

### Coupled Earth System DA
Physically consistent interface increments

### Composition DA
Unified state and forcing estimation

### ML EDA emulation
Low-cost flow-dependent uncertainty
Wei: Thursday, 9:00

### ML model error
Characterise and correct systematic errors
Patrick: Tuesday, 13:30
Alban and Marcin: Thursday, 14:15

### Towards an Earth-system reanalysis
Atmosphere, ocean, land, air quality (aerosols, greenhouse gases, chemistry) + carbon and methane cycle

# Conclusions

- The reanalyses are being increasingly used for assesments of climate trends (from trends in thermodynamic variables to atmospheric dynamics and circulation, and recently also extreme events)
- The requirements of reanalyses are different to requirements of weather prediction
  → the reanalyses should be the best possible description of the past atmospheric/ocean/land state
  → the reanalyses should obey physical constraints (mass, energy, concepts of dynamic meteorology…)
  → physical models coupled with objective DA algorithms (e.g. 4D-Var) were so far able to deliver state-of-the-art reanalyses and we should keep developing those
- ERA6 is an important step forward in addressing the limitations of reanalyses, but weakneses remain

- The developments of the next-generated Earth-system reanalysis will be supported by the CLIMATE-NEXT project: HORIZON-CL5 Next Generation Climate Monitoring (led by Massimo Bonavita)

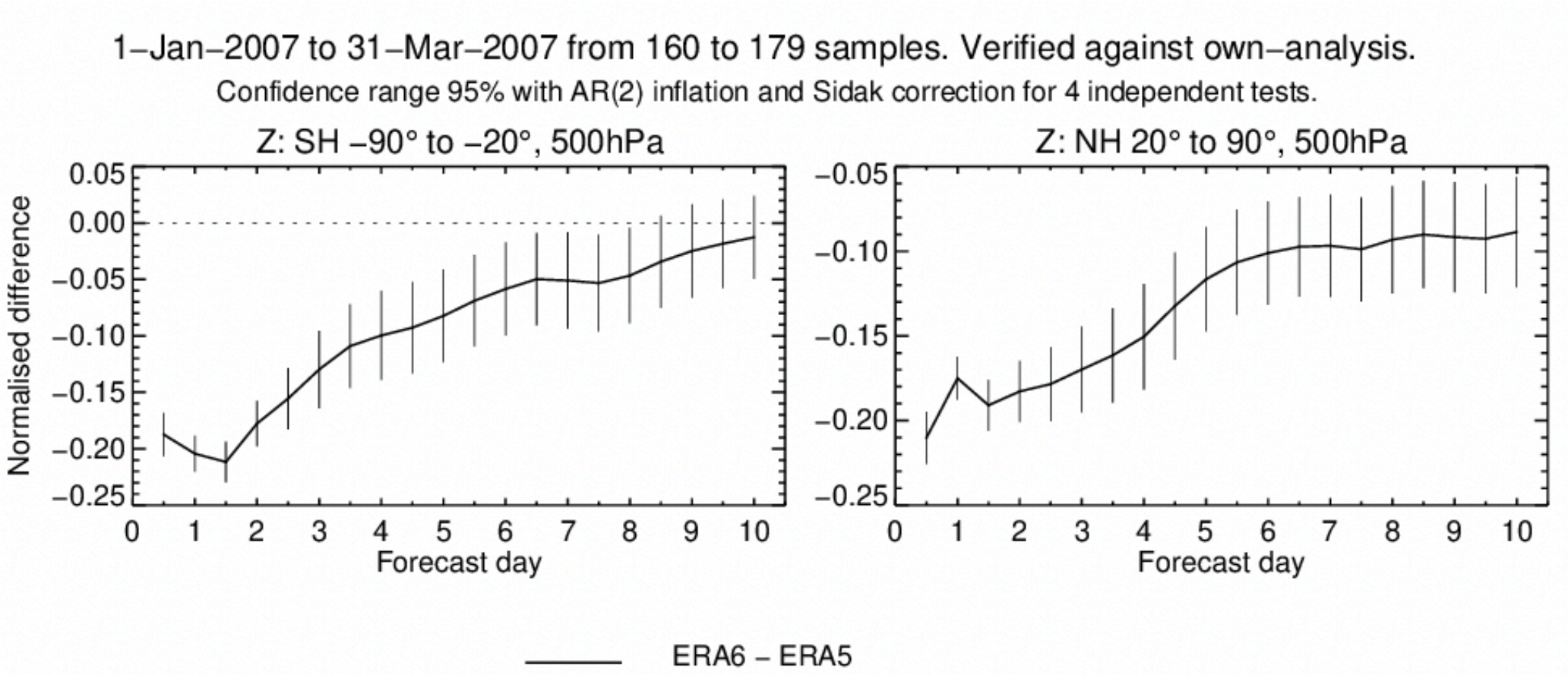


Ziga Zaplotnik, Massimo Bonavita

Thanks to Hans Hersbach, Paul Poli, Marcin Chrust, Matic Pikovnik, Bostjan Melinc, Lukas Kugler

www.ecmwf.int

# ERA6 supersedes ERA5



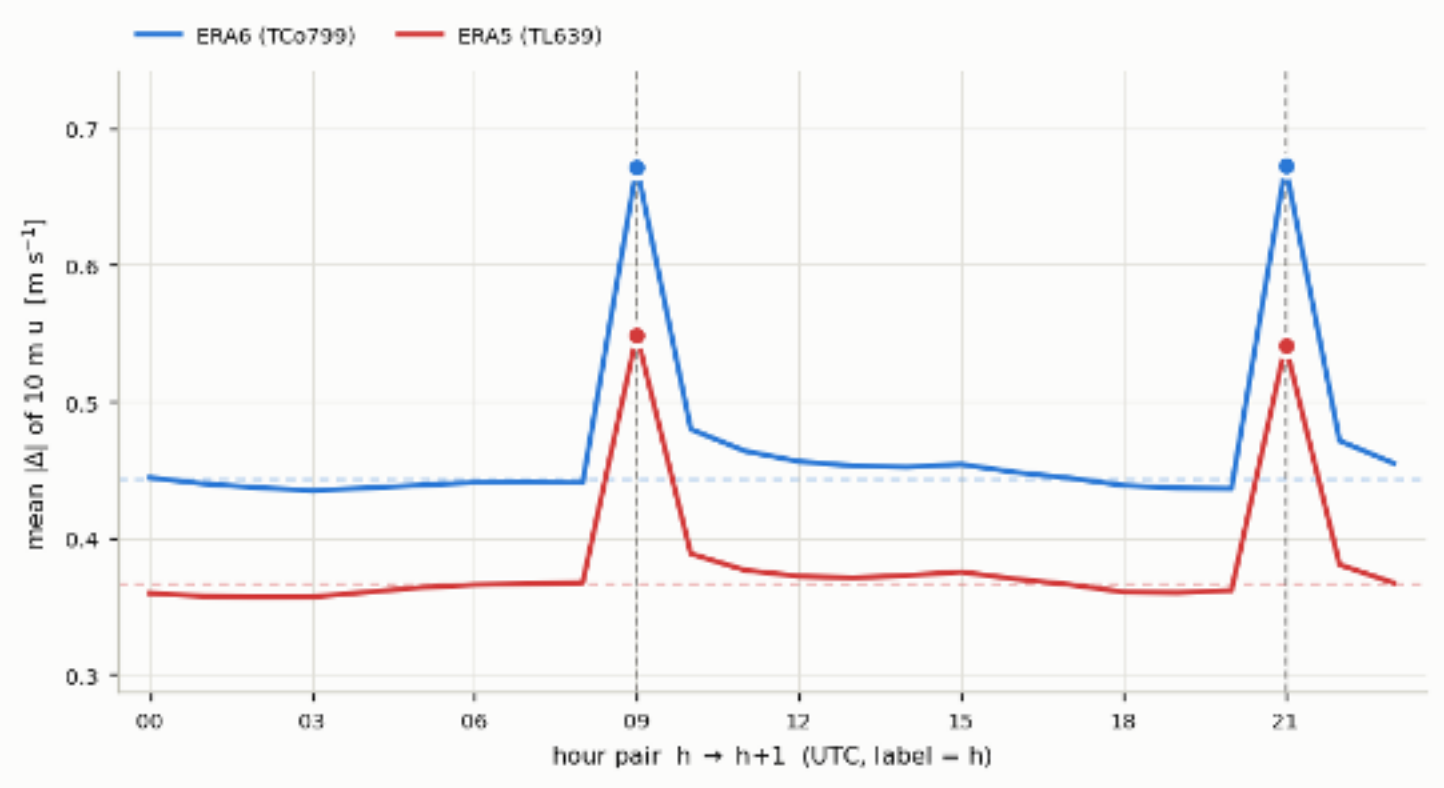

ERA6 gains approximately 12 hours
in accuracy of initial conditions.

Paul Poli and Reanalysis Team

# Discontinuities between 4D-Var windows

How large are the discontinuities?
We can compare $|\mathbf{x}(t+1) - \mathbf{x}(t)|$ of the two transition steps with the 22 non-transition steps

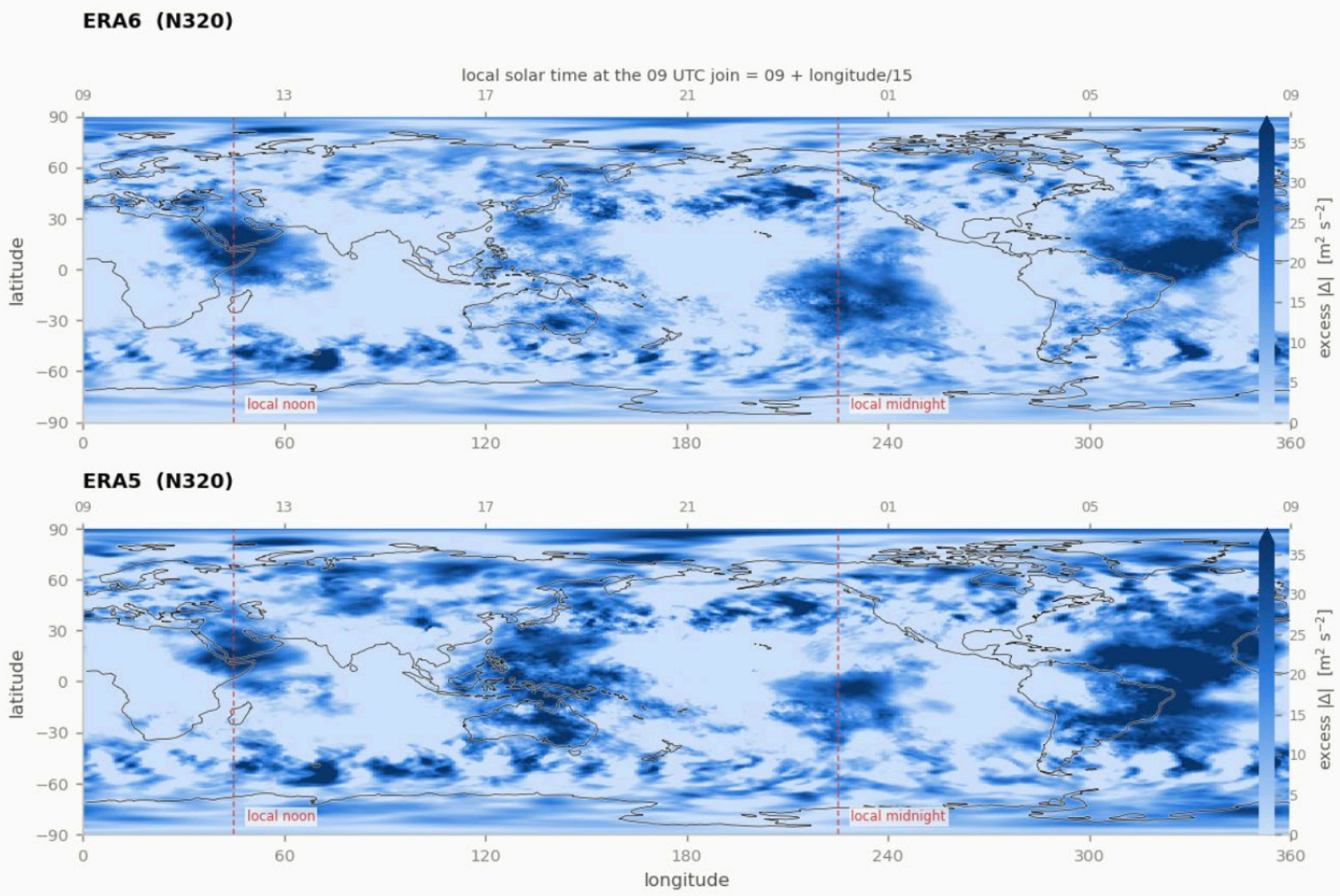
Mean absolute jump in u10m: ERA6 vs ERA5




× 1.50

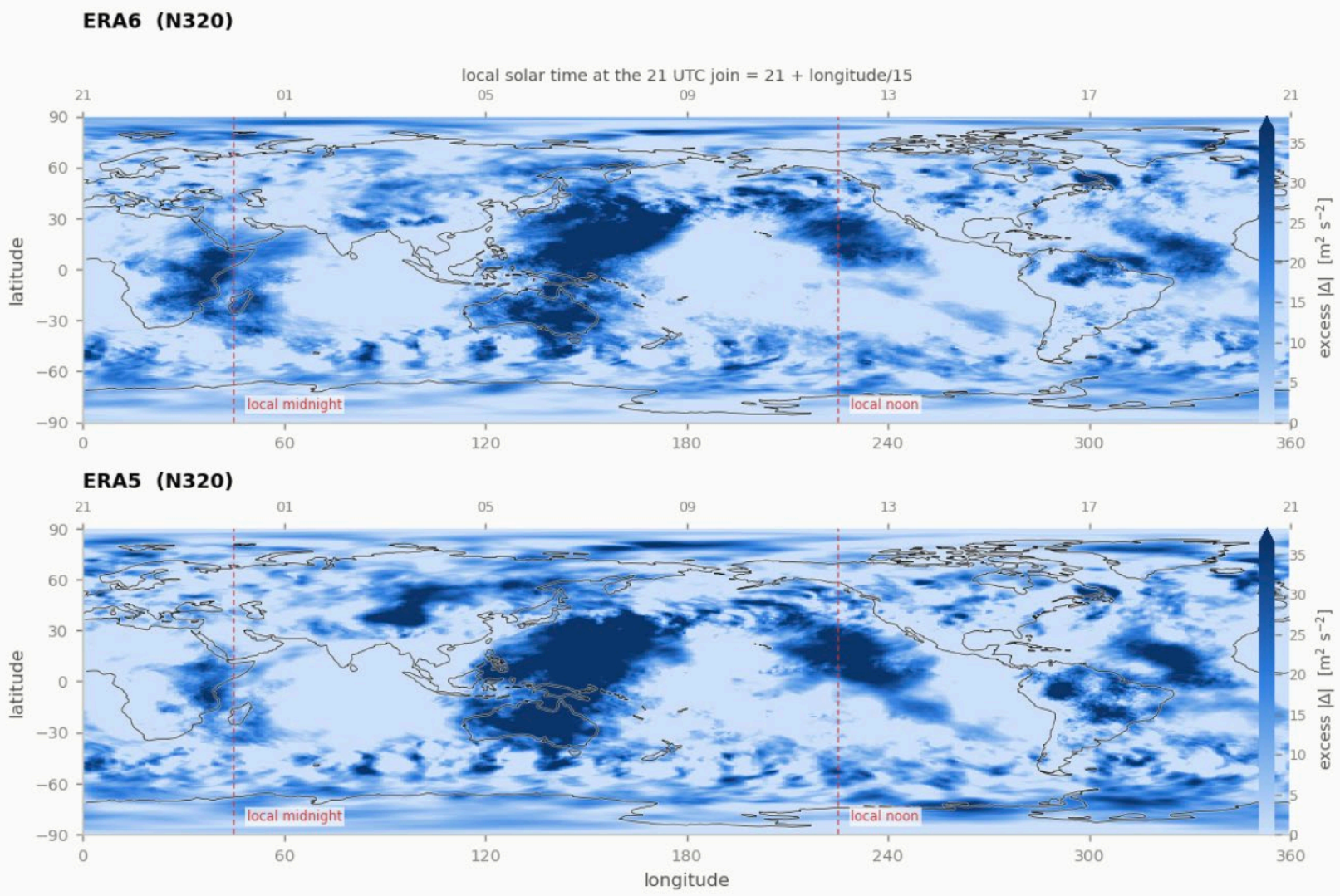
# Extra slides

Z500, winter,
09 UTC



# Extra slides

Z500, winter,
21 UTC



# Lorenz-63

| scheme | nRMSE | jump RMS | joins/day |
|---|---|---|---|
| non-overlapping 6 h | 0.1692 | 0.1099 | 4 |
| non-overlapping 12 h | 0.1473 | 0.1195 | 2 |
| non-overlapping 24 h | 0.3084 | 0.1999 | 1 |
| overlapping 6/3 h | 0.1645 | 0.0876 | 8 |
| overlapping 12/6 h | 0.1393 | 0.1008 | 4 |
| **overlapping 24/12 h** | **0.1161** | 0.0904 | 2 |

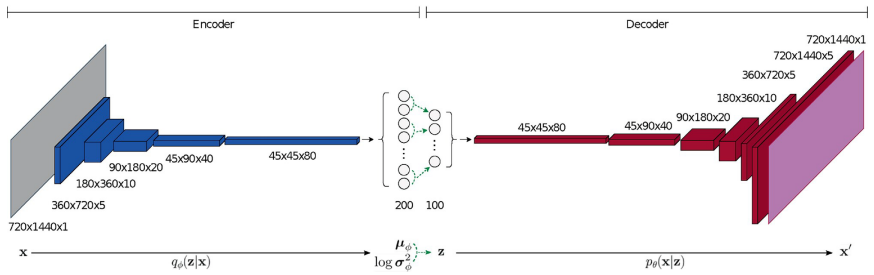
# Latent space data assimilation – a good autoencoder is all you need ☺

Requirements of a “good” autoencoder:

- Properly reconstructs the initial fields
- Performs a significant data compression (O(10x) to O(100x))
- Learns to decorrelate latent space components (spatially, spectrally, feature-wise)
- … and many more (follow the talk on Thursday at 13.30 on latent space DA)

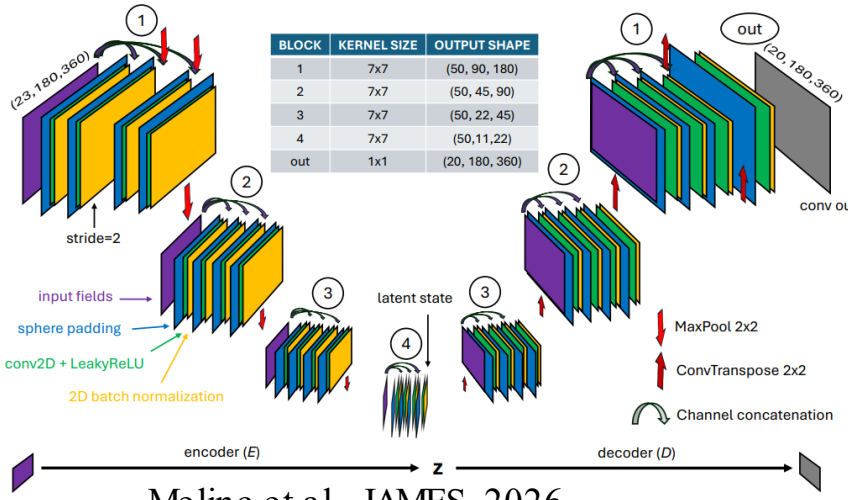
TIMELINE

Proof of concept: simple univariate CNN




Melinc and Zaplotnik, QJRMS, 2024

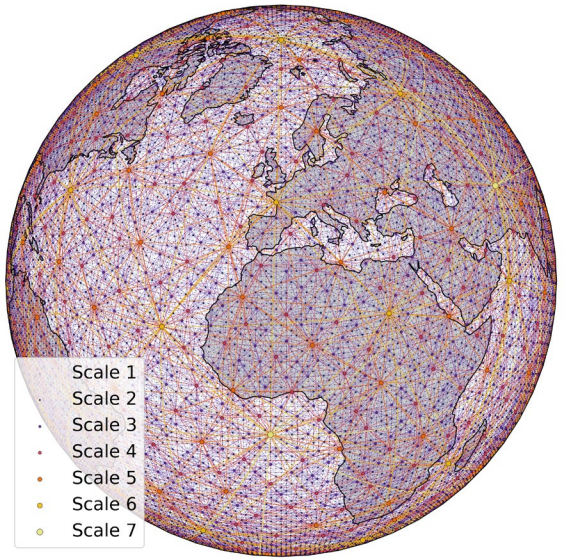
Multivariate multilevel CNN




| BLOCK | KERNEL SIZE | OUTPUT SHAPE |
|---|---|---|
| 1 | 7x7 | (50, 90, 180) |
| 2 | 7x7 | (50, 45, 90) |
| 3 | 7x7 | (50, 22, 45) |
| 4 | 7x7 | (50,11,22) |
| out | 1x1 | (20, 180, 360) |

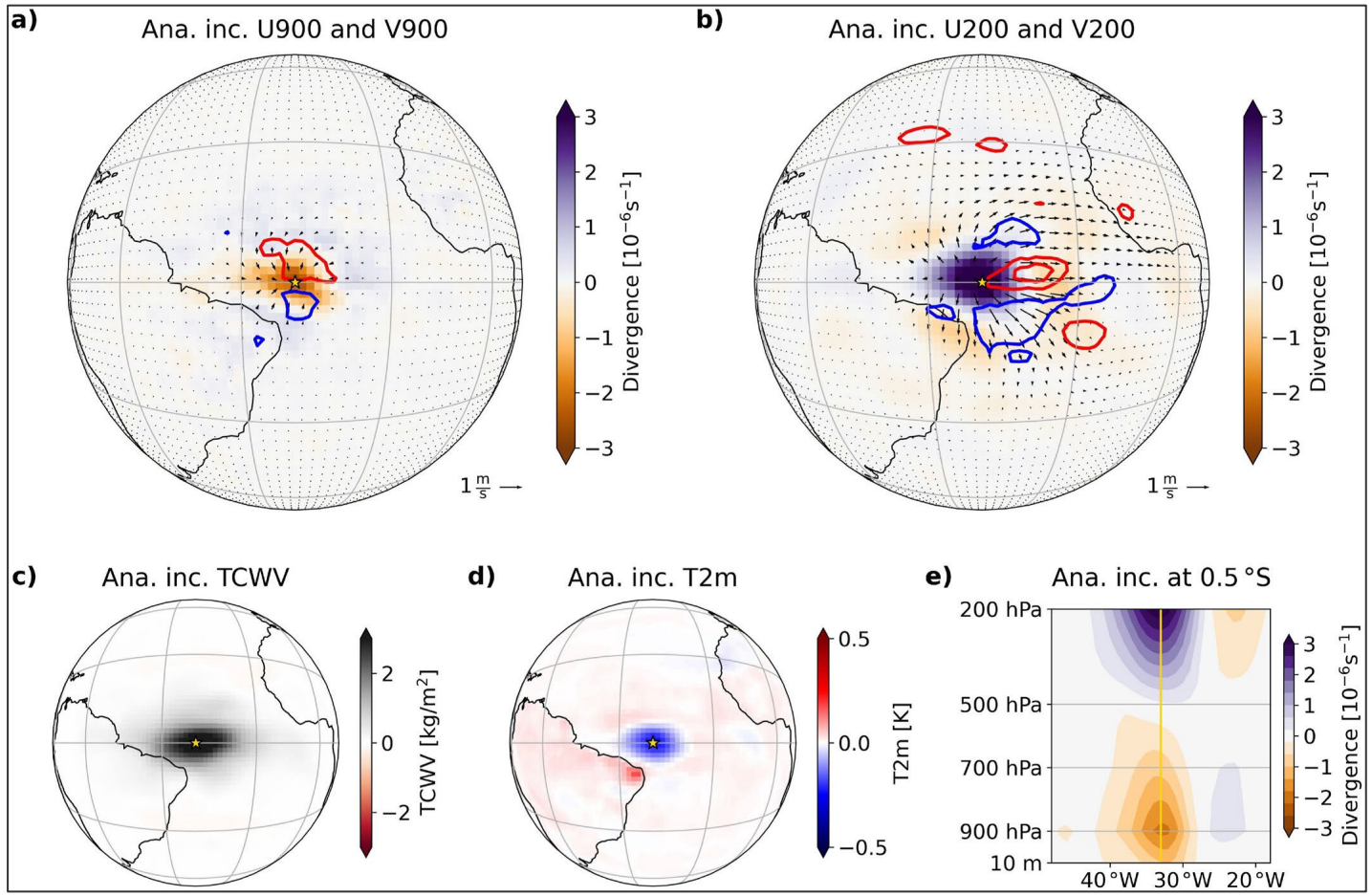
Melinc et al., JAMES, 2026

Fully-coupled GNN




Zaplotnik et al., 2026 (in prep.)

# Latent space data assimilation




## Wavelet-based B →ML latent representation of coupled covariances

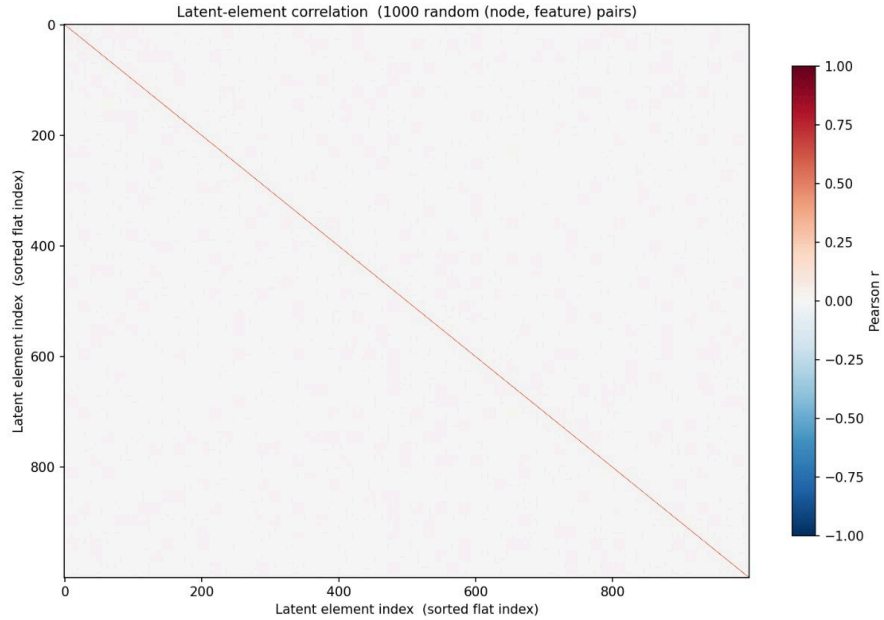
Cross-variable increments: Assimilation of Total Column Water Observation in the tropical Atlantic induces fully physically consistent dynamic and thermodynamic response in IC.

Crucial capability for correcting the representation of tropical climate variables, e.g. strength of Hadley circulation


Melinc, B., Perkan, U., &Zaplotnik, Ž. (2026). Aunified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation. Journal of Advances in Modeling Earth Systems, 18, e2025MS005360. https://doi.org/10.1029/2025MS005360

# Latent space data assimilation – measuring the $B_z$-matrix

Encodes EDA perturbed members i = 1,…,9 for every 3 valid dates across 2023–2025. For each date t, nine consecutive circular pairs of perturbed forecasts at 6-hour lead time (initialised at 18 UTC, valid for 00 UTC, on which data the autoencoder is trained) $(i,j) \in \{(1,2), (2,3), \ldots, (8,9), (9,1)\}$ yield latent-space differences:




$$\varepsilon_{ij}(t) = z_{b,i}(t) - z_{b,j}(t)$$

Because both members are drawn from the same background-error distribution N(0,B), the expected outer product is

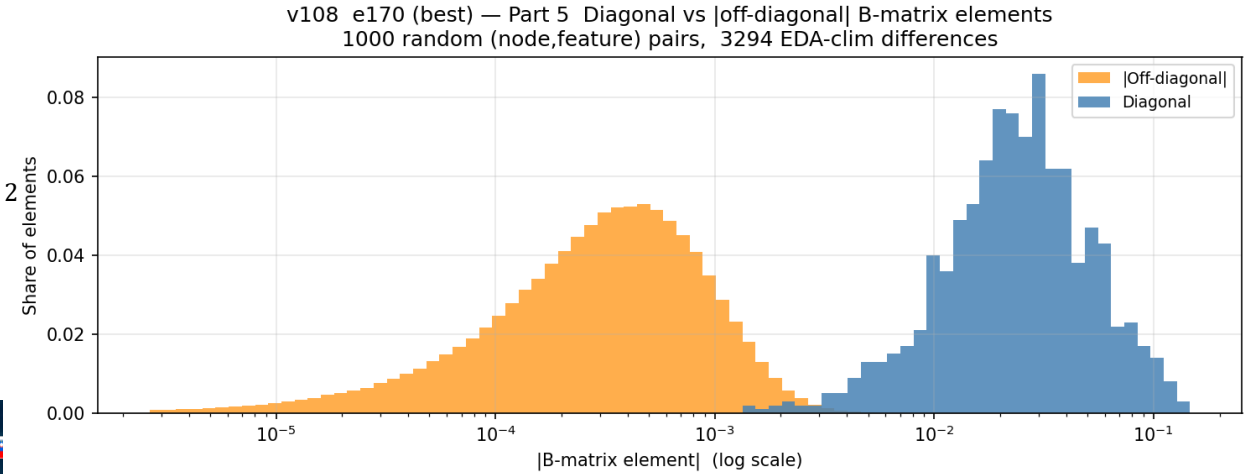
$$E[\varepsilon_{ij}\, \varepsilon_{ij}^{T}] = 2B$$

All accumulators are divided by 2. Note that the control (i = 0) was not used in the computations.

Diagonal B-matrix:

$$B_z^{\text{EDA-clim}}[n,f] = 1/(2T) \sum_{t=1,\ldots,T} \sum_{(i,j)} \varepsilon_{ij}[n,f](t)^2$$

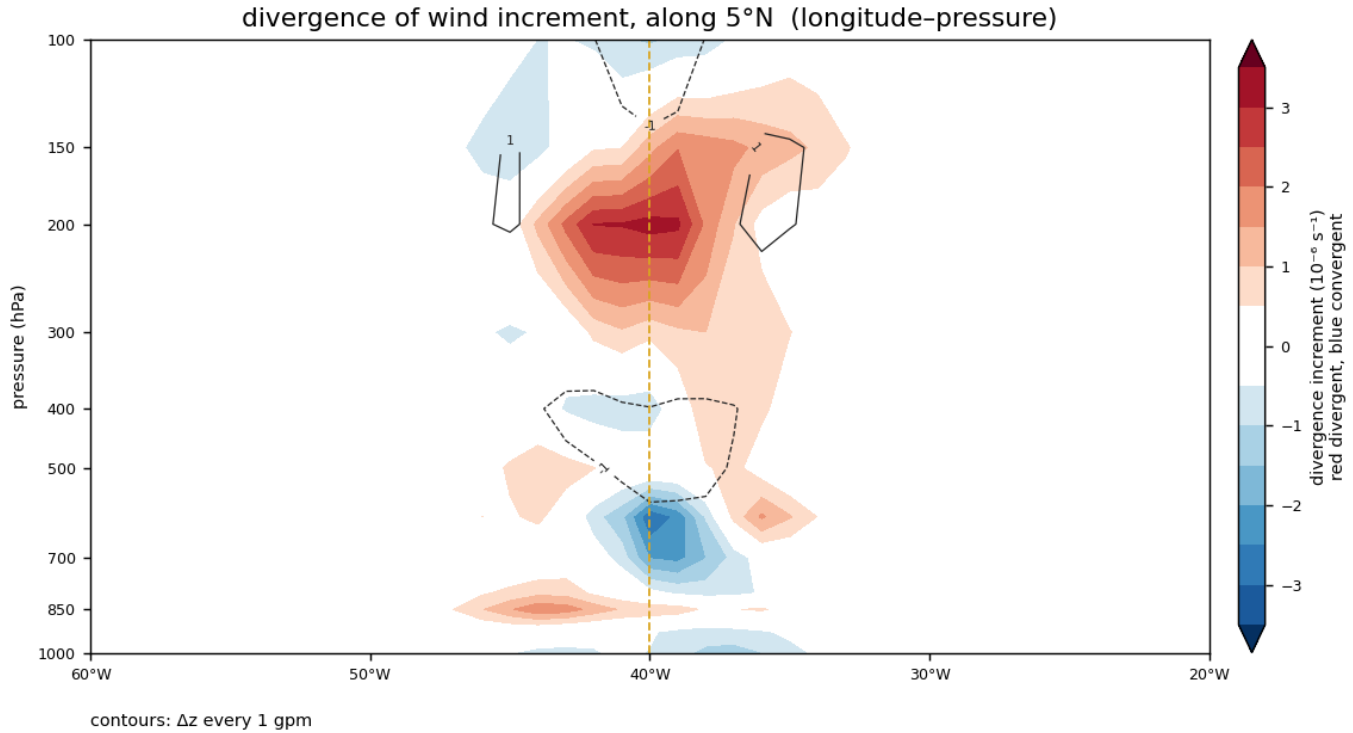
n…node
f…feature



# Latent space data assimilation – single obs experiment

A single observation experiment assimilating synthetic observation of 500 hPa specific humidity with +4g/kg departure on 2024-04-14 at 5N 40W (north of the coast of Brazil):

- Localised horizontally
- Physically-consistent dynamic and thermodynamic response (e.g. pressure-difference driven friction flow at the bottom
- Gill-like response to heat-induced perturbations (upper-divergence and lower-level convergence)



# Latent space data assimilation

- For the application in IFS/future reanalyses, we will replace the control variable transform with the linearised decoder and its adjoint
- Incremental 4D-Var cost function (only Jo and Jb terms) and its gradient are currently:

$$J(\boldsymbol{\chi}) = \frac{1}{2}(\boldsymbol{\chi}+\boldsymbol{\chi}_g)^T(\boldsymbol{\chi}+\boldsymbol{\chi}_g) + \frac{1}{2}\sum_{k=0}^{N}(\mathbf{d}_k - \mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{L}\boldsymbol{\chi})^T\mathbf{R}_k^{-1}(\mathbf{d}_k - \mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{L}\boldsymbol{\chi}) \qquad \delta\mathbf{x} = \mathbf{L}\boldsymbol{\chi}$$

$$\nabla_{\boldsymbol{\chi}}J(\boldsymbol{\chi}) = \boldsymbol{\chi} + \boldsymbol{\chi}_g + \sum_{k=0}^{N}\mathbf{L}^T\mathbf{M}_{t_0\to t_k}^T\mathbf{H}_k^T\mathbf{R}_k^{-1}\left(\mathbf{d}_k - \mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{L}\boldsymbol{\chi}\right)$$

- Square-root preconditioned incremental latent-space 4D-Var becomes:

$$J_{\chi_z}(\boldsymbol{\chi}_z) = \frac{1}{2}(\boldsymbol{\chi}_z+\boldsymbol{\chi}_{zg})^T(\boldsymbol{\chi}_z+\boldsymbol{\chi}_{zg})+\frac{1}{2}\sum_{k=0}^{N}(\mathbf{d}_k-\mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{S}^{-1}\mathbf{D}\mathbf{L}_z\boldsymbol{\chi}_z)^T\mathbf{R}_k^{-1}(\mathbf{d}_k-\mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{S}^{-1}\mathbf{D}\mathbf{L}_z\boldsymbol{\chi}_z)$$

$$\nabla_{\boldsymbol{\chi}_z}J(\boldsymbol{\chi}_z) = \boldsymbol{\chi}_z + \boldsymbol{\chi}_{zg} + \sum_{k=0}^{N}\mathbf{L}_z^T\mathbf{D}^T\mathbf{S}^{-1}\mathbf{M}_{t_0\to t_k}^T\mathbf{H}_k^T\mathbf{R}_k^{-1}\left(\mathbf{H}_k\mathbf{M}_{t_0\to t_k}\mathbf{S}^{-1}\mathbf{D}\mathbf{L}_z\boldsymbol{\chi}_z - \mathbf{d}_k\right)$$

  - Latent increment $\delta\mathbf{z} = \mathbf{z} - \mathbf{z}_g$
  - Guess increment $\delta\mathbf{z}_g = \mathbf{z}_g - \mathbf{z}_b$
  - Observation departure $\mathbf{d}_k = \mathbf{y}_k - \mathcal{H}_k\mathcal{M}_{t_0\to t_k}S^{-1}D(\mathbf{z}_g)$
  - Control variable transform $\mathbf{L}_z = \mathbf{B}_z^{1/2}$, such that $\delta\mathbf{z} = \mathbf{L}_z\boldsymbol{\chi}_z$

TL approximation: $D(\mathbf{z}) - D(\mathbf{z}_b) = \mathbf{D}\delta\mathbf{z}$

$$\mathbf{D} = (\partial/\partial\mathbf{z})\Big|_{\mathbf{z}=\mathbf{z}_g} D$$

# Latent space data assimilation – pathway to IFS implementation

1) N128 ML137 pytorch-based autoencoder successfully trained on a single GPU ✓
2) EDA-based climatological B constructed ✓
3) Include torch c++ copy of the autoencoder in IFS and test in research configuration (ongoing work with Marcin and Alban)

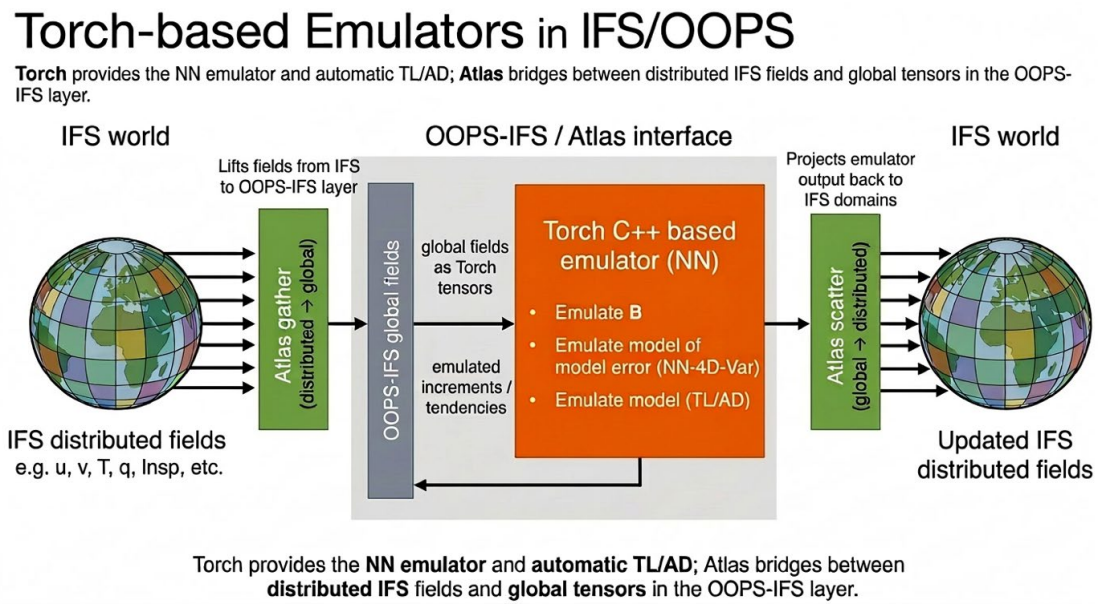



Figure courtesy of Marcin Chrust

See his post on "Supporting NN in OOPS-IFS" https://confluence.ecmwf.int/spaces/ESAS/pages/673320300/DA+Methodology+Team

4) N320 ML137 autoencoder being prepared (ERA5 data supplemented with ERA6 data)
5) Dual-resolution autoencoder (trained on ERA5 grid fine-tuned on native ERA6 grid at O800)