Interpretability of AI weather models via intermediate decoding

Matthias Beylich, Kirsten Tempest and George Craig

Meteorological Institute Munich, Ludwig Maximilian University of Munich

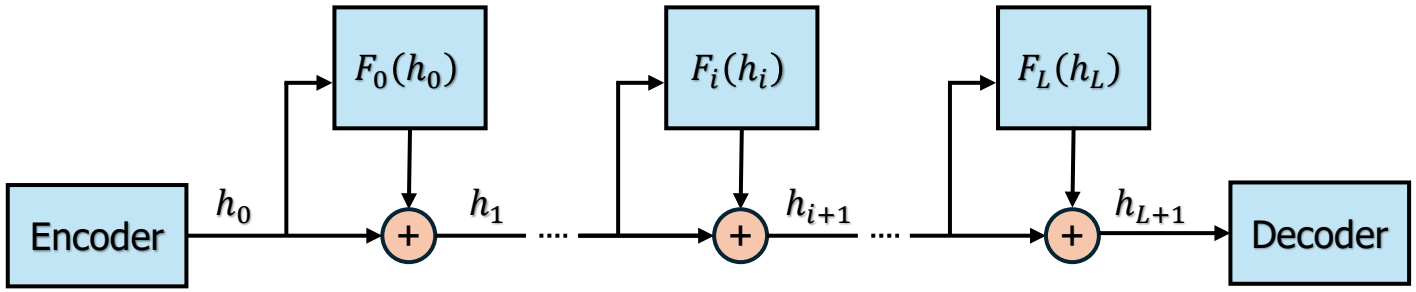
Introduction

- AI weather models are still black boxes
- Apply method of Intermediate Decoding to three different models:
  - Graphcast (Graph Neural Network)
  - Aurora (Transformer model)
  - ArchesWeather (Transformer model)
- How does the forecast evolve with increasing model depth?
  - Model loss
  - Spatial scale progression
  - Physical balances

Methods

- Models employ a residual network architecture:

$$\mathbf{h}_{\mathrm{i}+1} = \mathbf{h}_{\mathrm{i}} + F_{\mathrm{i}}(\mathbf{h}_{\mathrm{i}})$$



- Send intermediate model states to the decoder to create a prediction $\hat{X}^{t+1}$:

$$\hat{X}^{t+1}(\mathrm{i}) = Decoder(\mathbf{h}_{\mathrm{i}})$$

- Apply linear transformation (translator) before decoding:

$$\hat{X}^{t+1}(i) = Decoder(\mathbf{h}_{\mathrm{i}}\mathrm{W}_{\mathrm{i}} + \mathbf{b}_{\mathrm{i}})$$

Methods

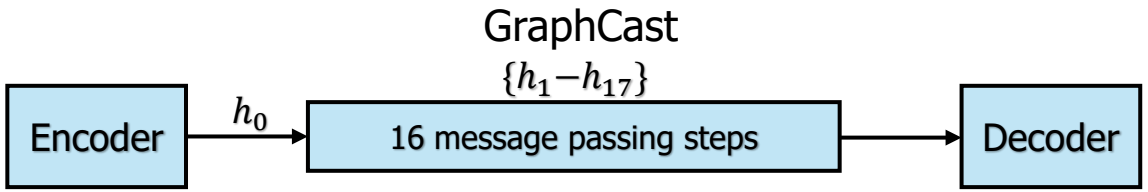
- Models generate a global forecast:

$$\hat{X}^{t+1} = Model(X^{t}, X^{t-1}), \qquad \Delta t_{GC,\,Au} = 6h, \Delta t_{Ar} = 24h$$

Methods

- Models generate a global forecast:

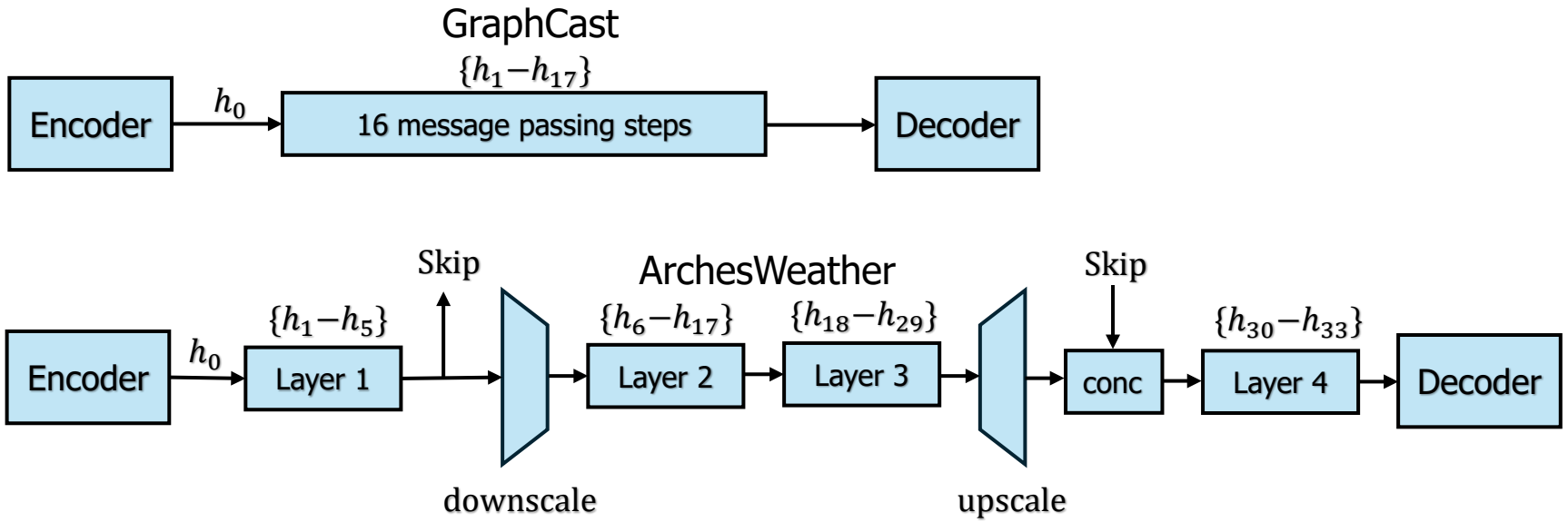
$$\hat{X}^{t+1} = Model(X^t, X^{t-1}), \qquad \Delta t_{GC,\,Au} = 6h, \Delta t_{Ar} = 24h$$



Methods

- Models generate a global forecast:

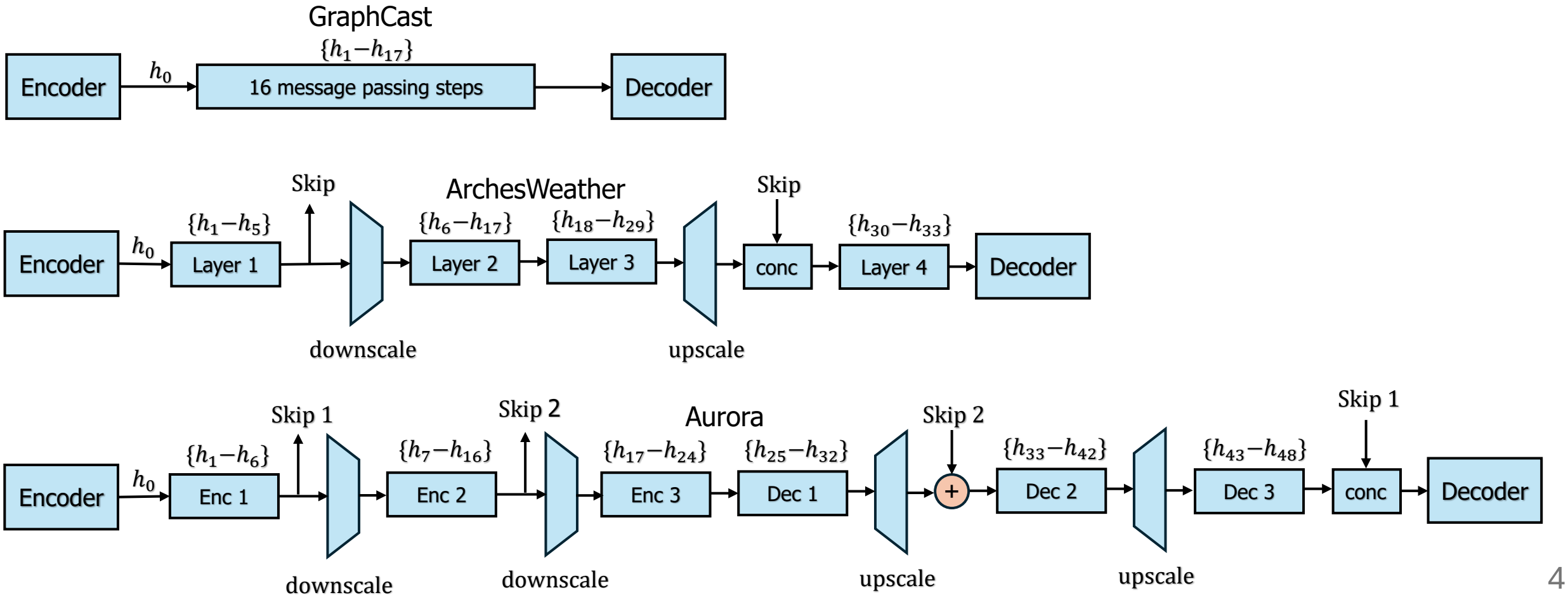
$$\hat{X}^{t+1} = Model(X^t, X^{t-1}), \qquad \Delta t_{GC,\,Au} = 6h, \Delta t_{Ar} = 24h$$



Methods

- Models generate a global forecast:

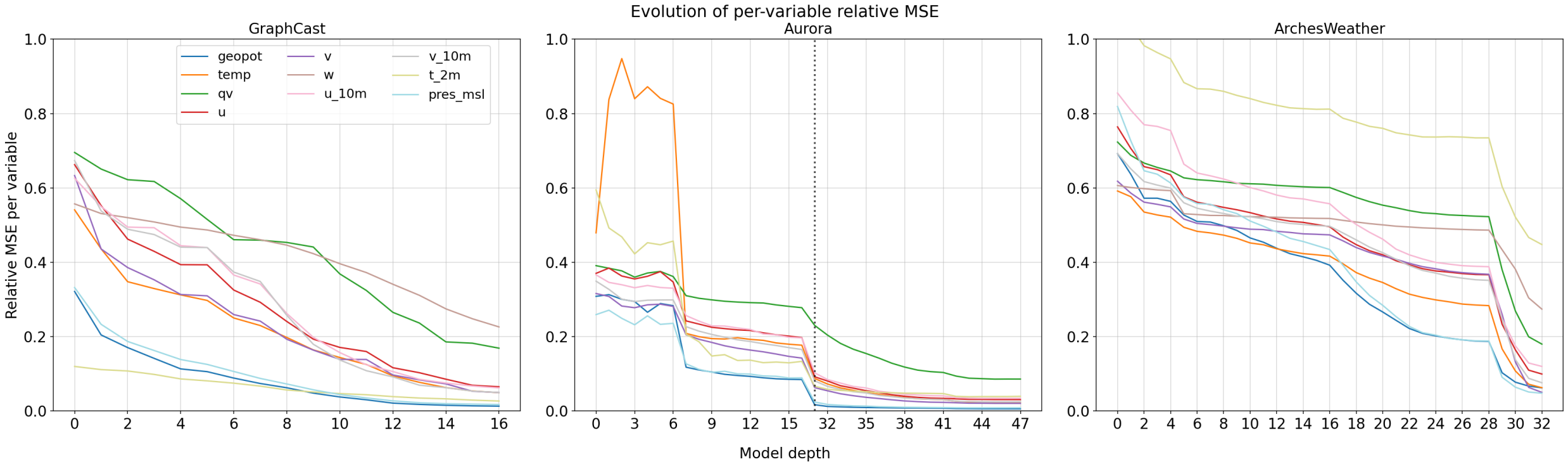
$$\hat{X}^{t+1} = Model(X^t, X^{t-1}), \qquad \Delta t_{GC,\,Au} = 6h, \Delta t_{Ar} = 24h$$



Forecast performance

- Calculate MSE for intermediate decodings:

$$MSE_{rel}(i) = \frac{MSE\left(\hat{X}_i^{t+1}, X^{t+1}\right)}{MSE(X^t, X^{t+1})}$$

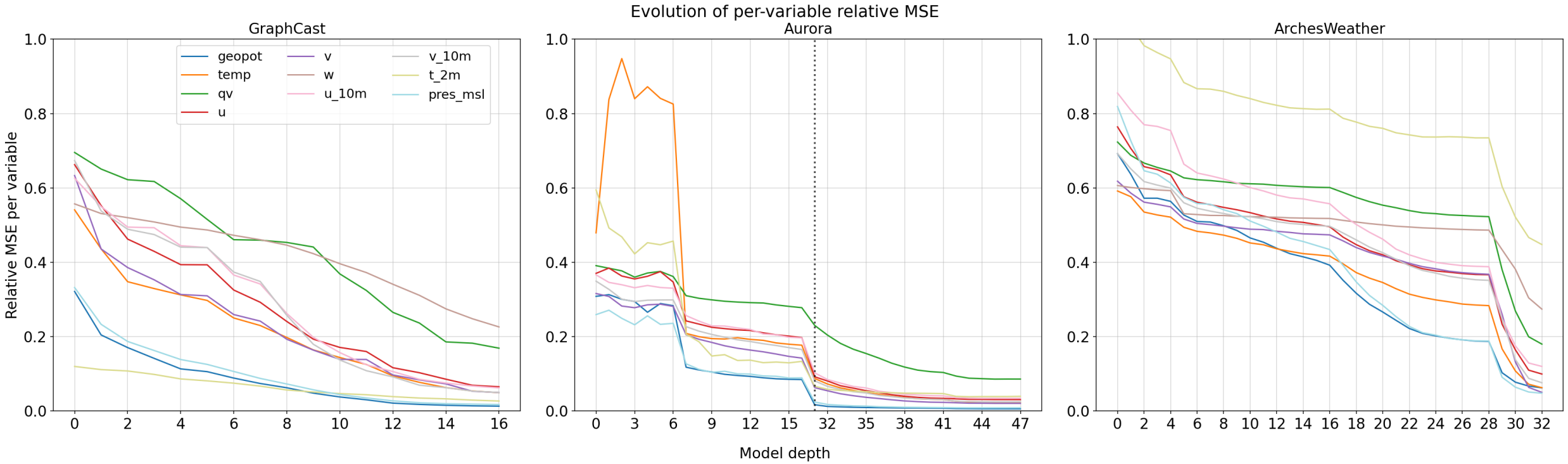


Forecast performance

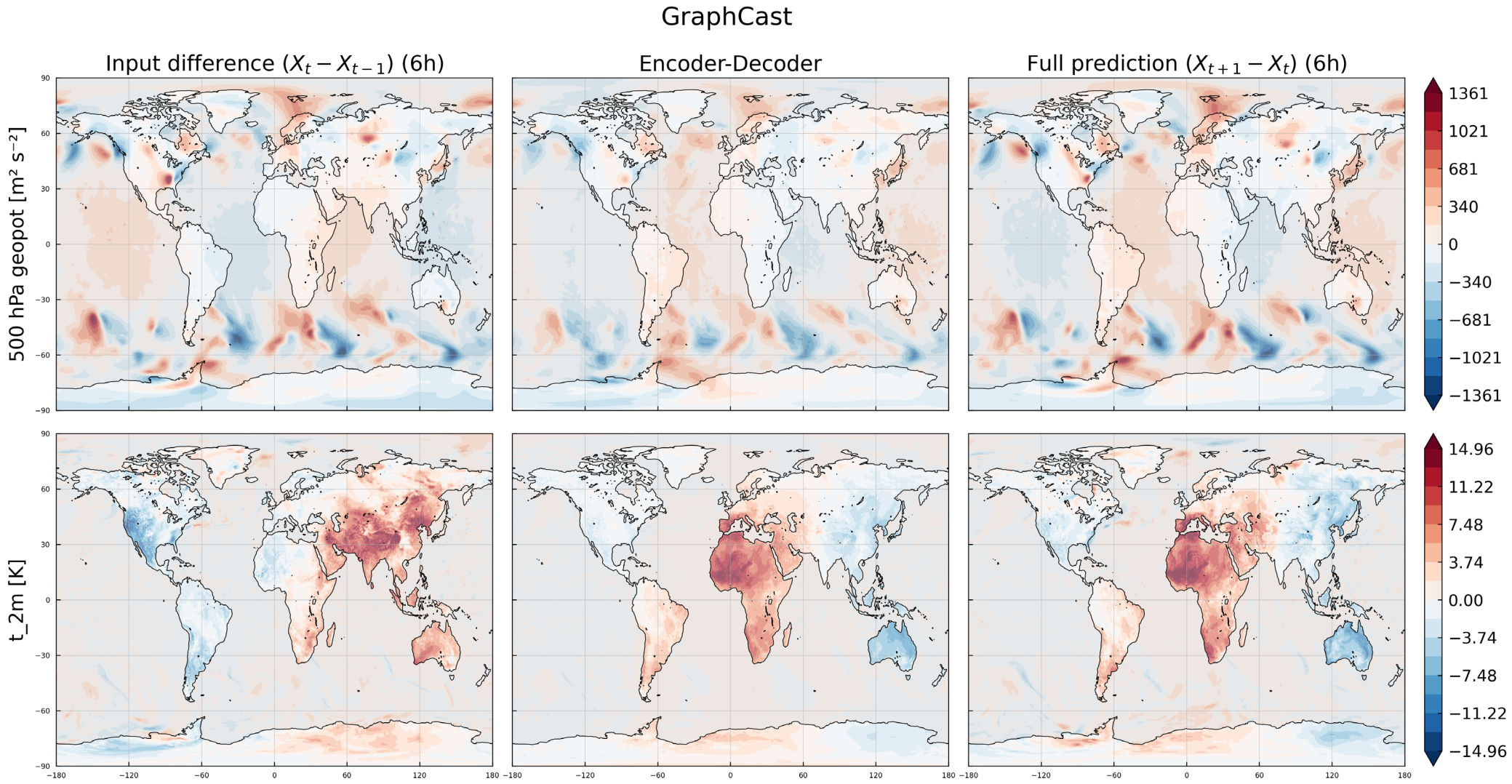
- Calculate MSE for intermediate decodings:

$$MSE_{rel}(i) = \frac{MSE\left(\hat{X}_i^{t+1}, X^{t+1}\right)}{MSE(X^t, X^{t+1})}$$

- ➢ Encoder immediately creates prediction
- ➢ Stage structure evident
- ➢ Gradual convergence towards final prediction within consistent stages



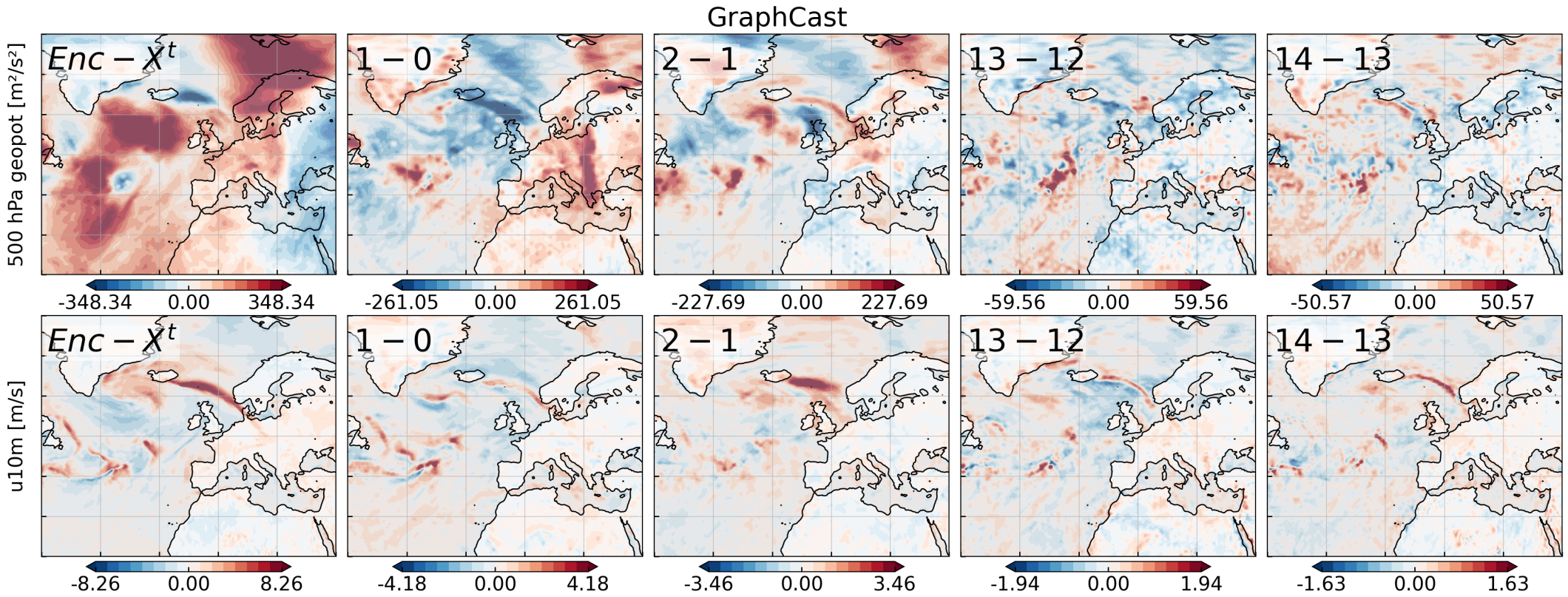
Encoder-Decoder only forecast



Forecast progression

- Calculate increment between intermediate decodings:

$$\Delta\hat{X}^{t+1}(i) = \hat{X}^{t+1}(i) - \hat{X}^{t+1}(i-1)$$

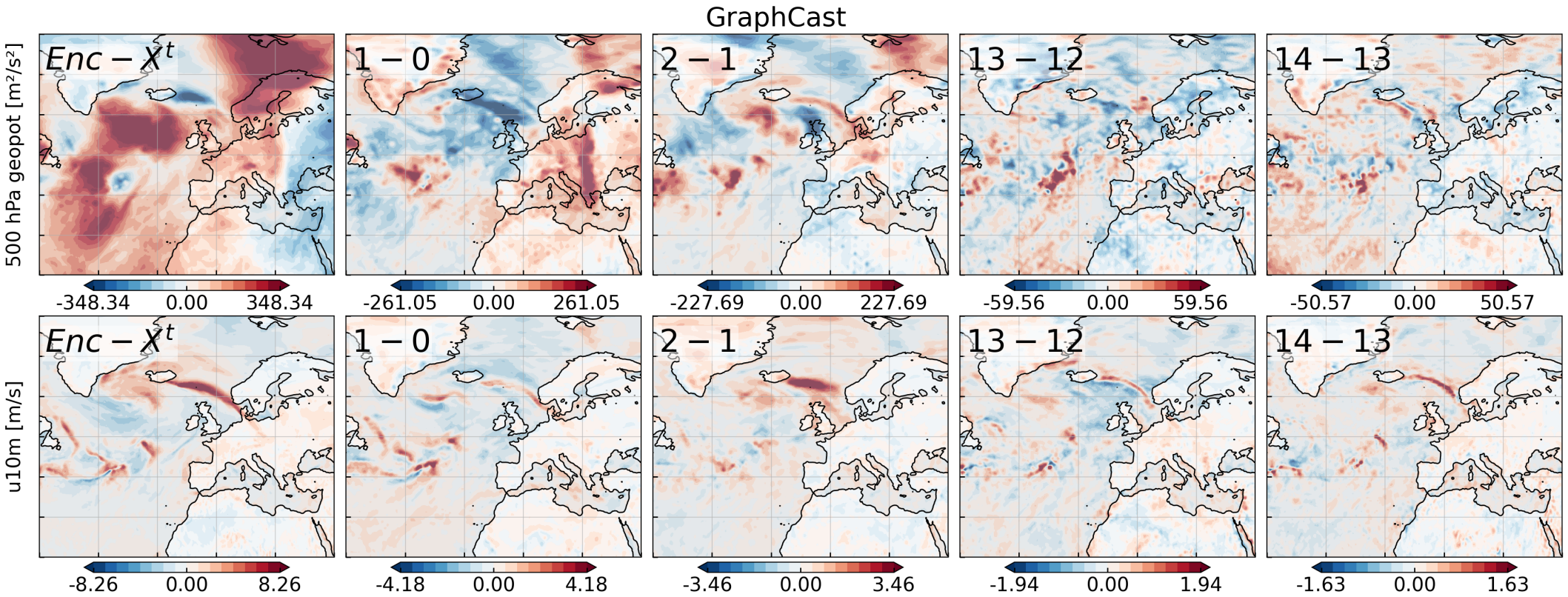


Forecast progression

- Calculate increment between intermediate decodings:

$$\Delta\hat{X}^{t+1}(i) = \hat{X}^{t+1}(i) - \hat{X}^{t+1}(i-1)$$

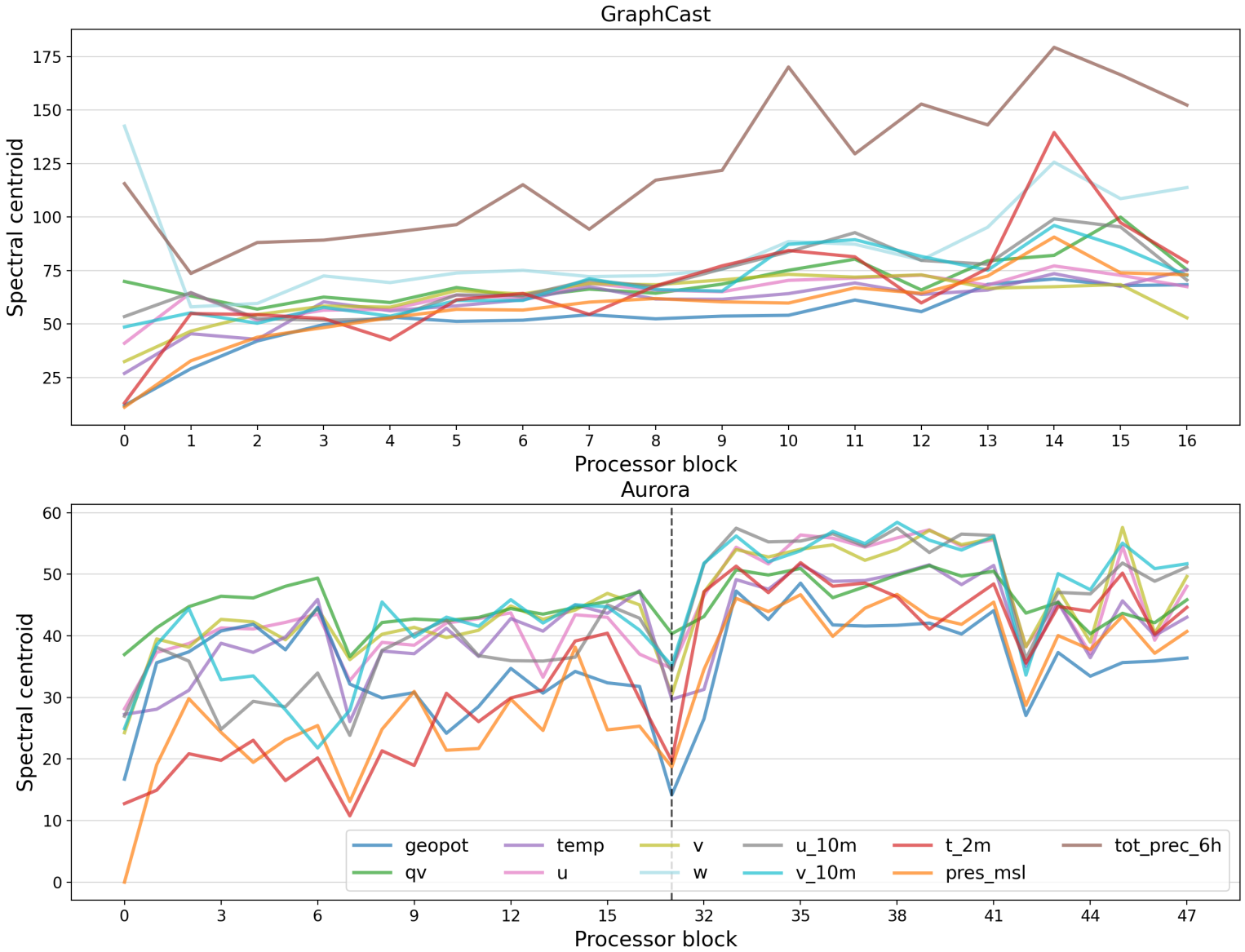
- ➢ Progression toward smaller spatial scales and reduced amplitudes with increasing model depth
- ➢ Features are iteratively refined



Forecast progression: Scales

- Calculate spectral centroids of increment fields:

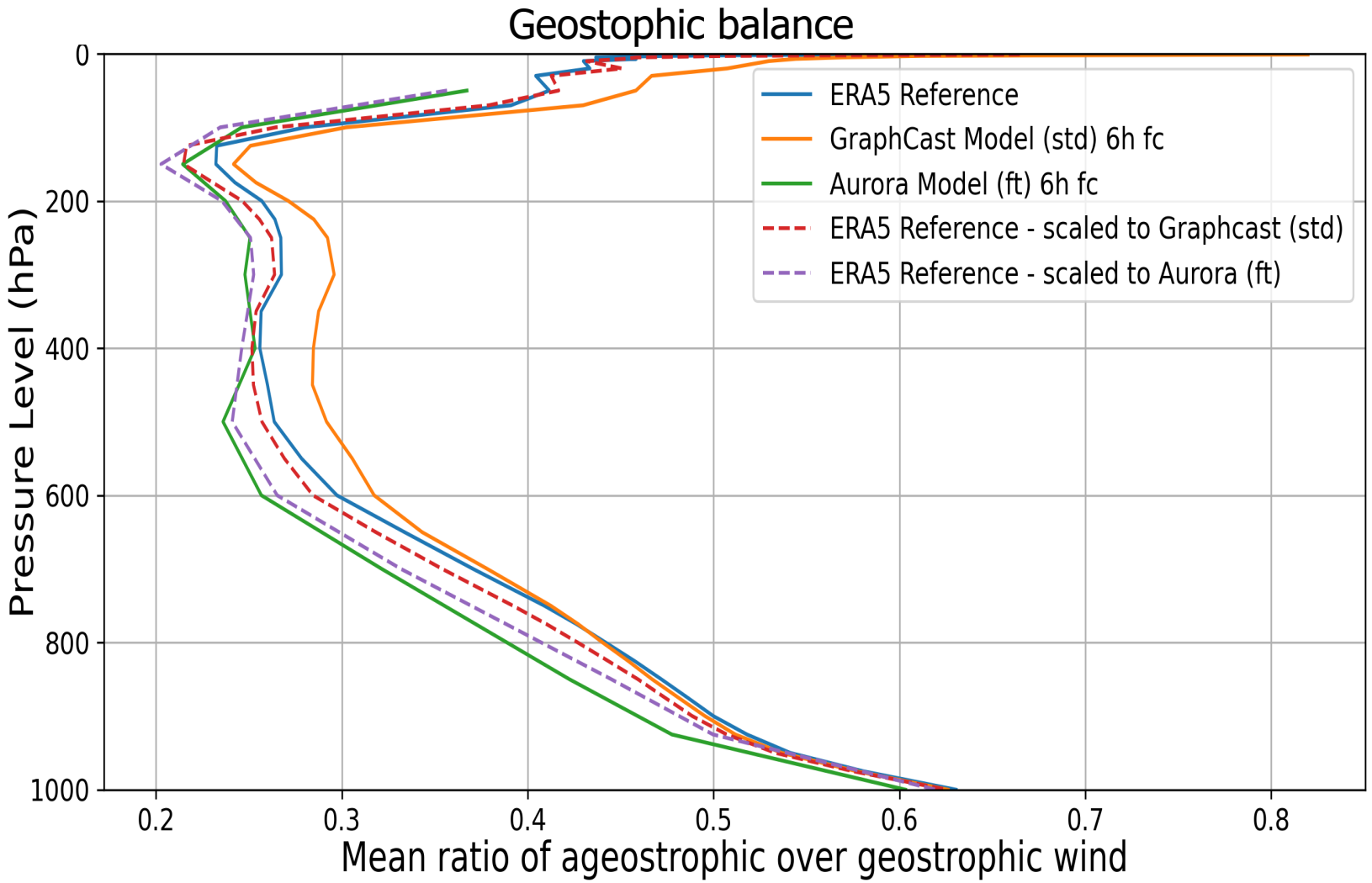
$$\mathrm{l}_{centroid} = \frac{\sum l\, P_l}{\sum l}$$



Geostophic balance

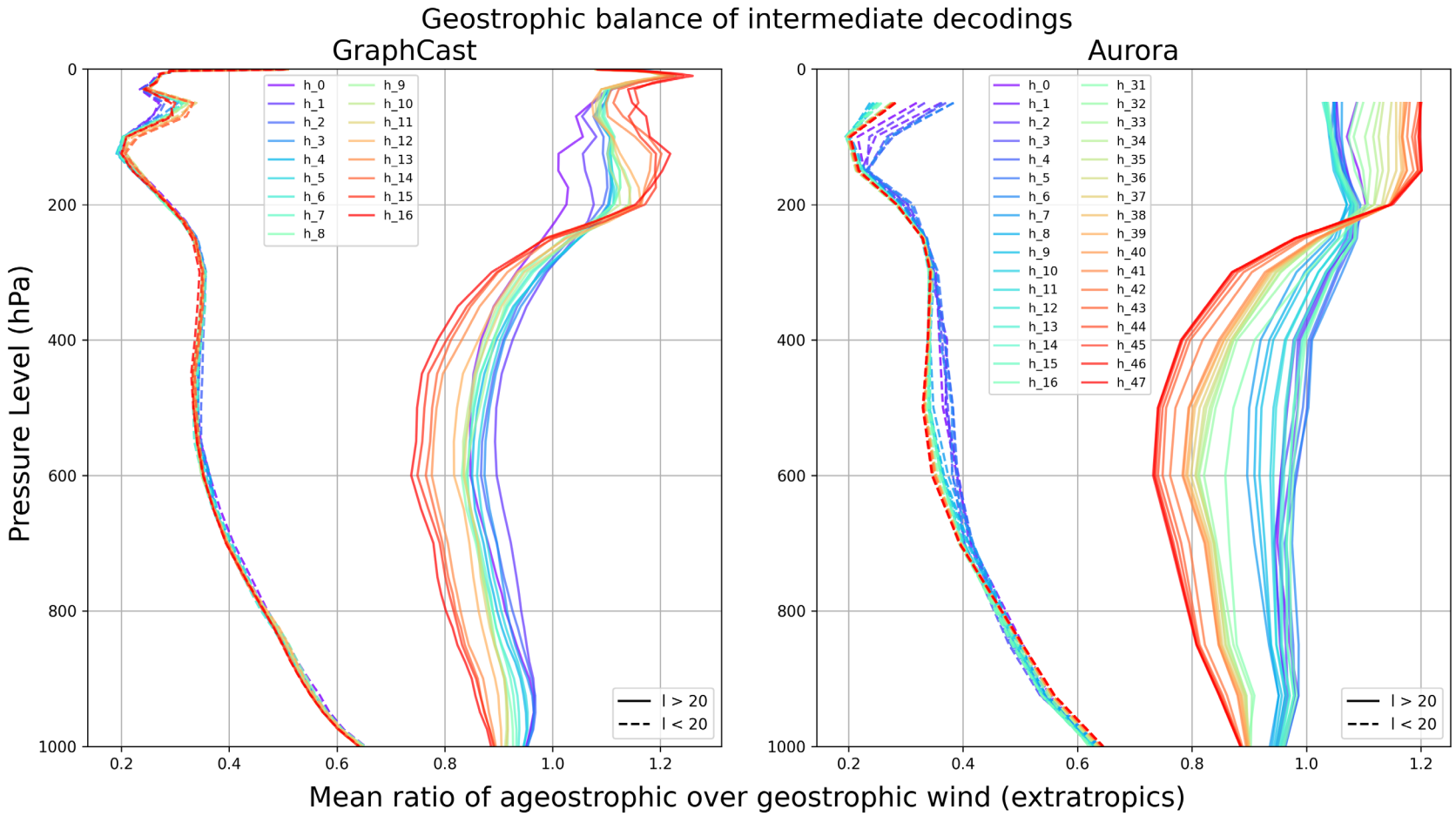
- Calculate geostrophic balance of model forecasts and compare to ERA5

$$r = \frac{\|\mathbf{u}_{\text{ageo}}\|}{\|\mathbf{u}_{\text{geo}}\|}, \quad lon > 20\,°$$



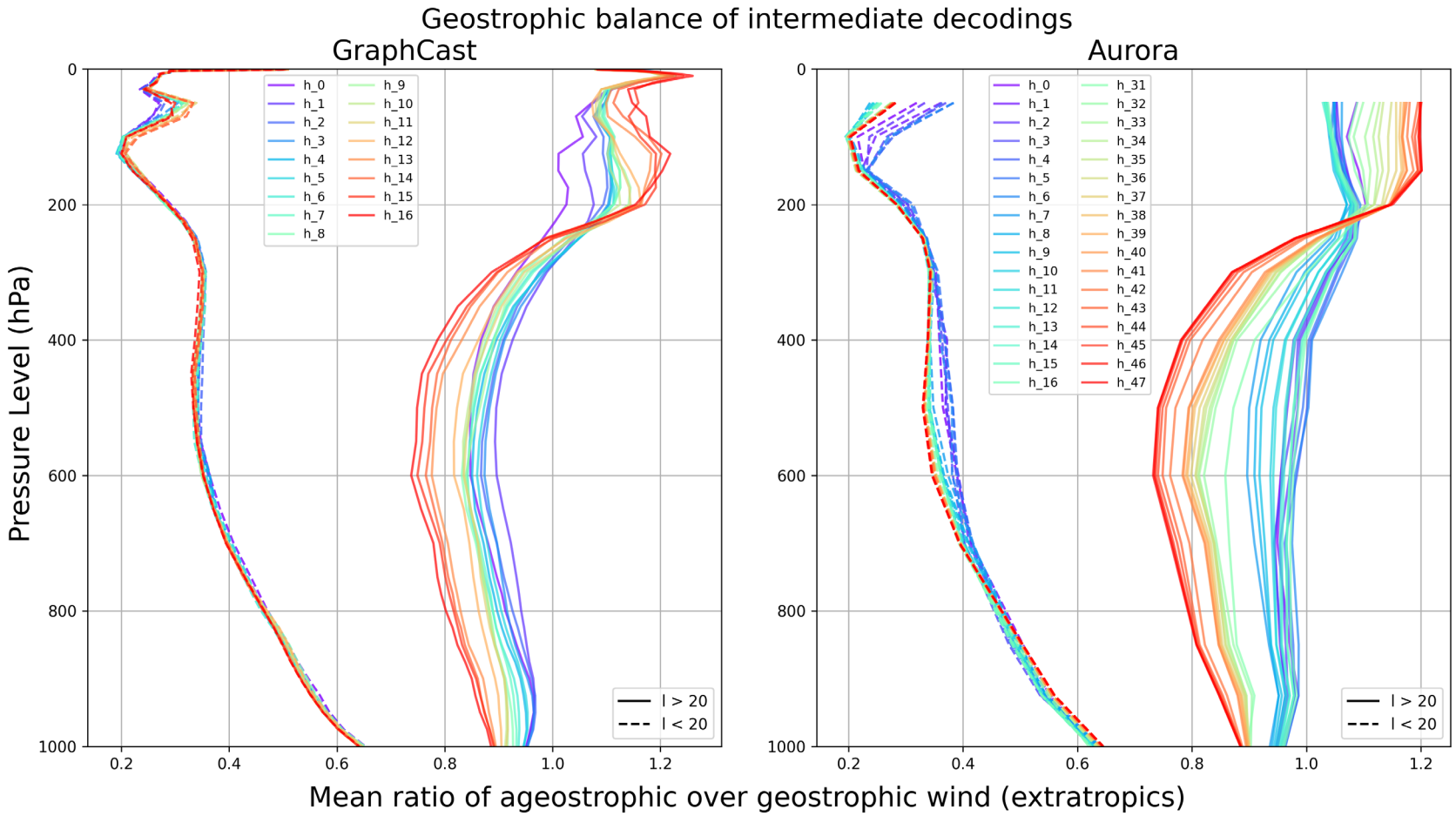
Geostophic balance: intermediate decodings

- Create intermediate decoding forecasts and filter scales:
  - Planetary/synoptic ($l \leq 20$)
  - Meso ($l > 20$)
- Calculate geostrophic balance on scale seperated intermediate decoding forecasts



Geostophic balance: intermediate decodings

- Create intermediate decoding forecasts and filter scales:
  - Planetary/synoptic ($l \leq 20$)
  - Meso ($l > 20$)
- Calculate geostrophic balance on scale seperated intermediate decoding forecasts



➢ Geostrophic balance increases with model depth, driven by small scales

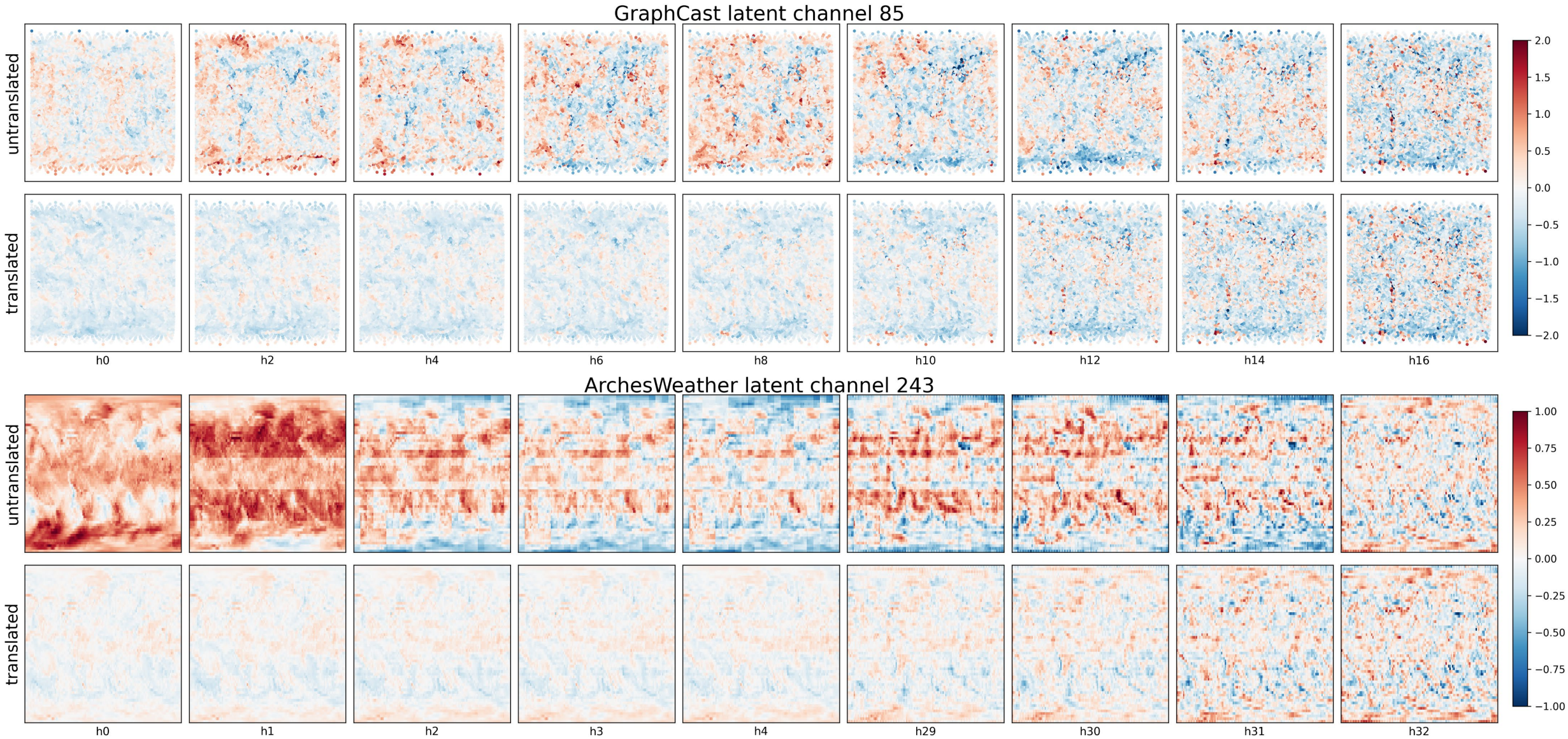
Conclusion

- We apply a method to track how forecasts evolve in physical space:
  - Stage structure relevant for forecast construction
  - Encoder-Decoder create sensible first guess
  - Progression to smaller scales
  - Geostrophic balance with later layers
- However: No direct access to interpretable latent features
- Next steps: Examine latent space directly, PCA, sparse autoencoders, Interactive tool (https://github.com/ktempestuous/latent_space_visualiser_weather_models)

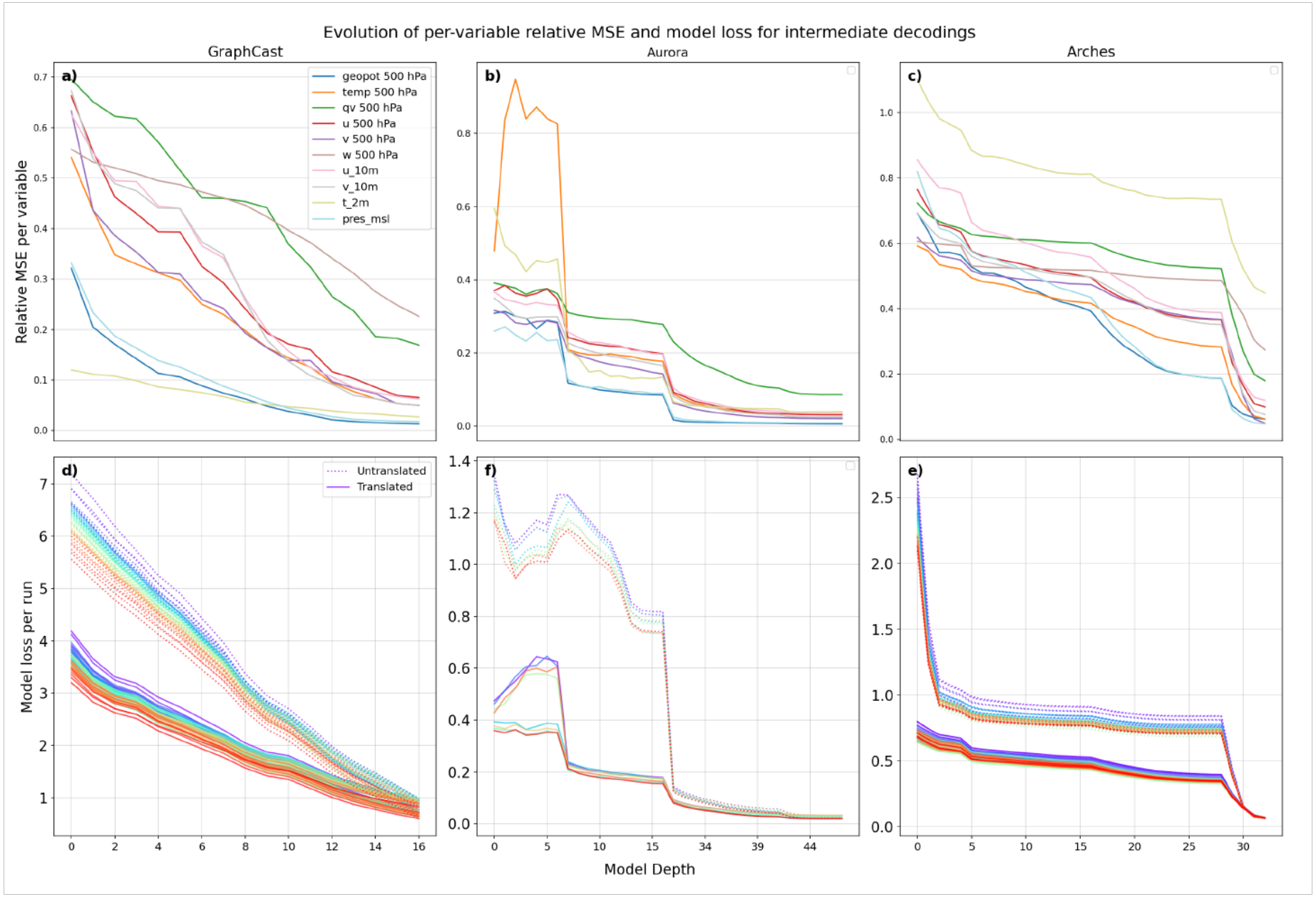
(k.tempest@physik.uni-muenchen.de)

Appendix

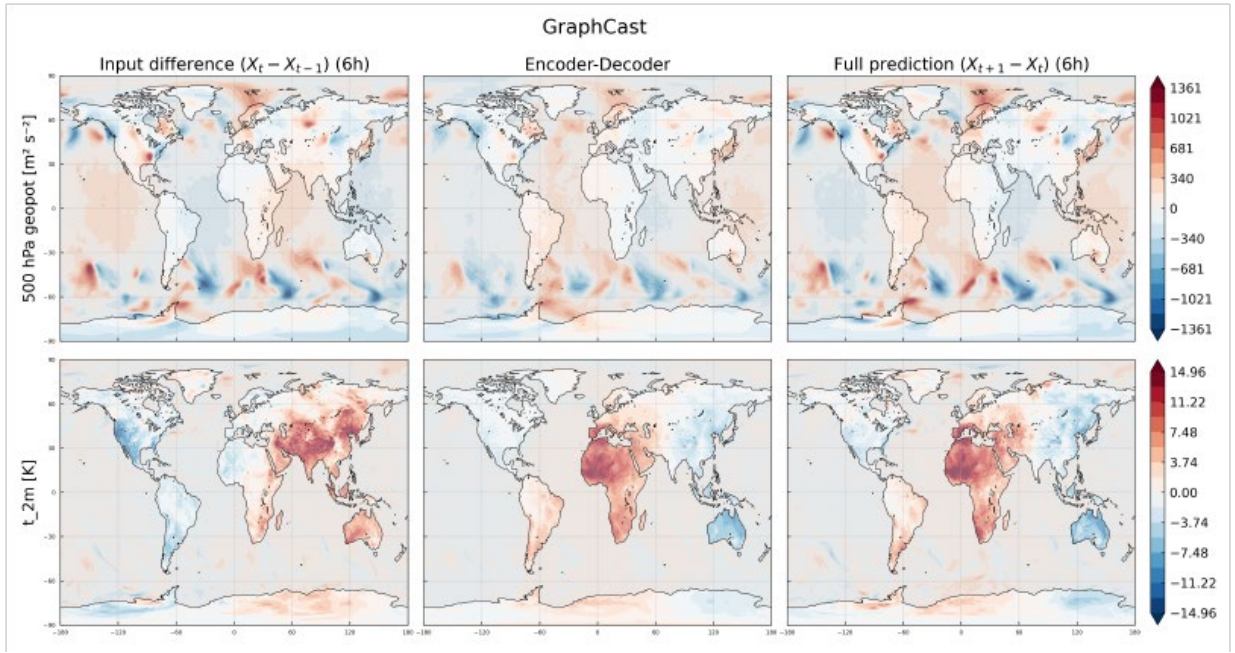
Latent channel developement

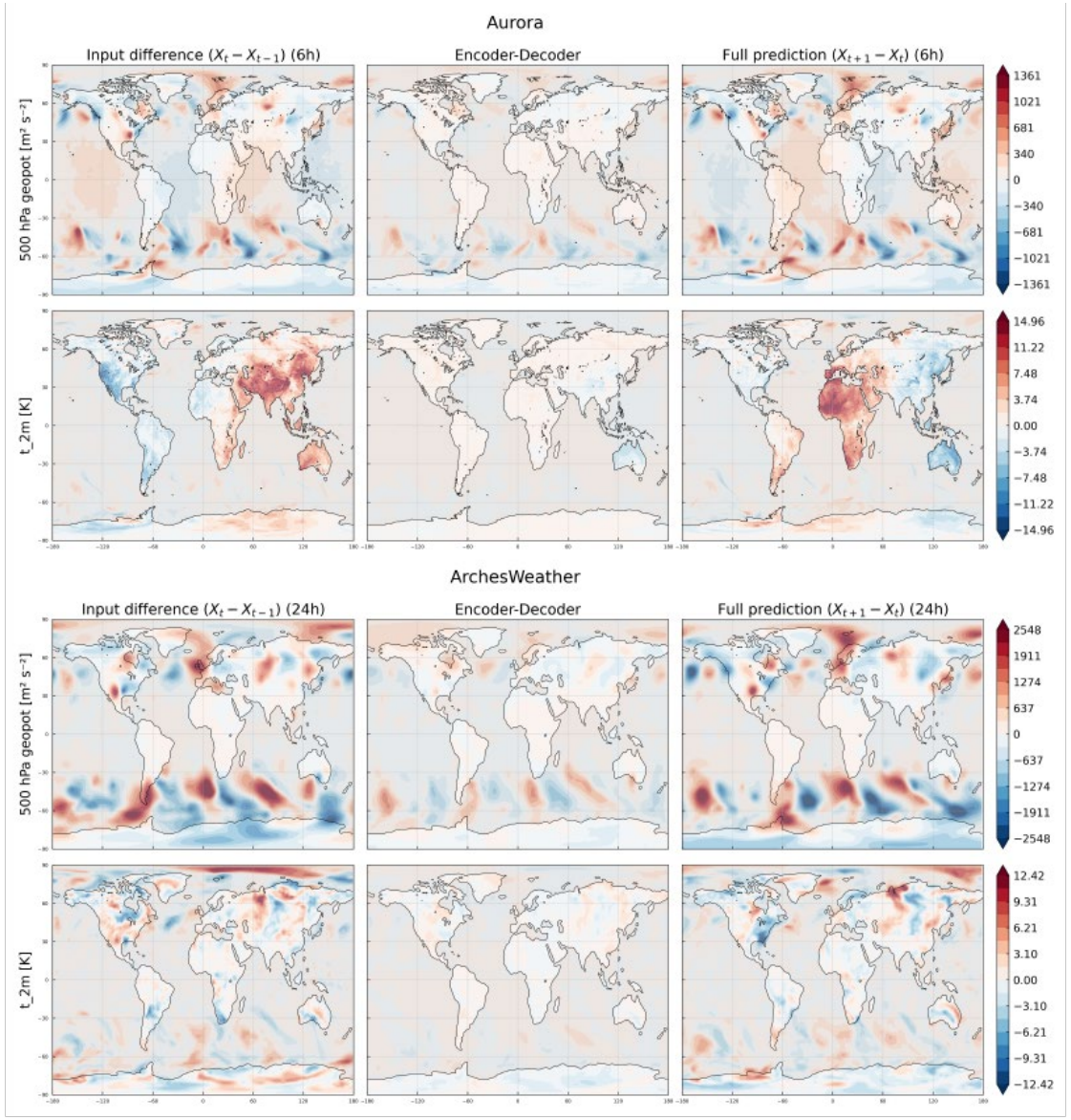


Forecast progression

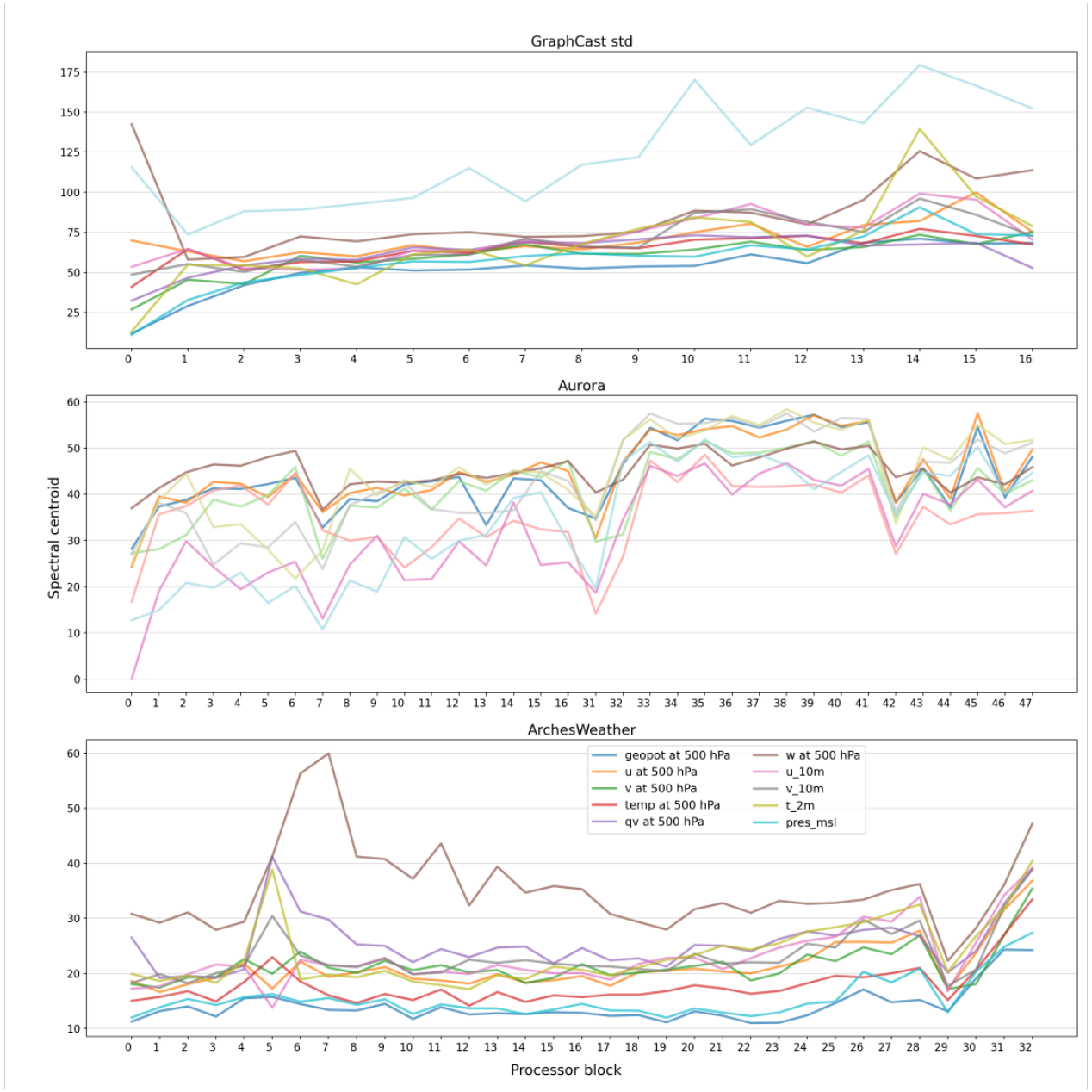


Encoder-Decoder

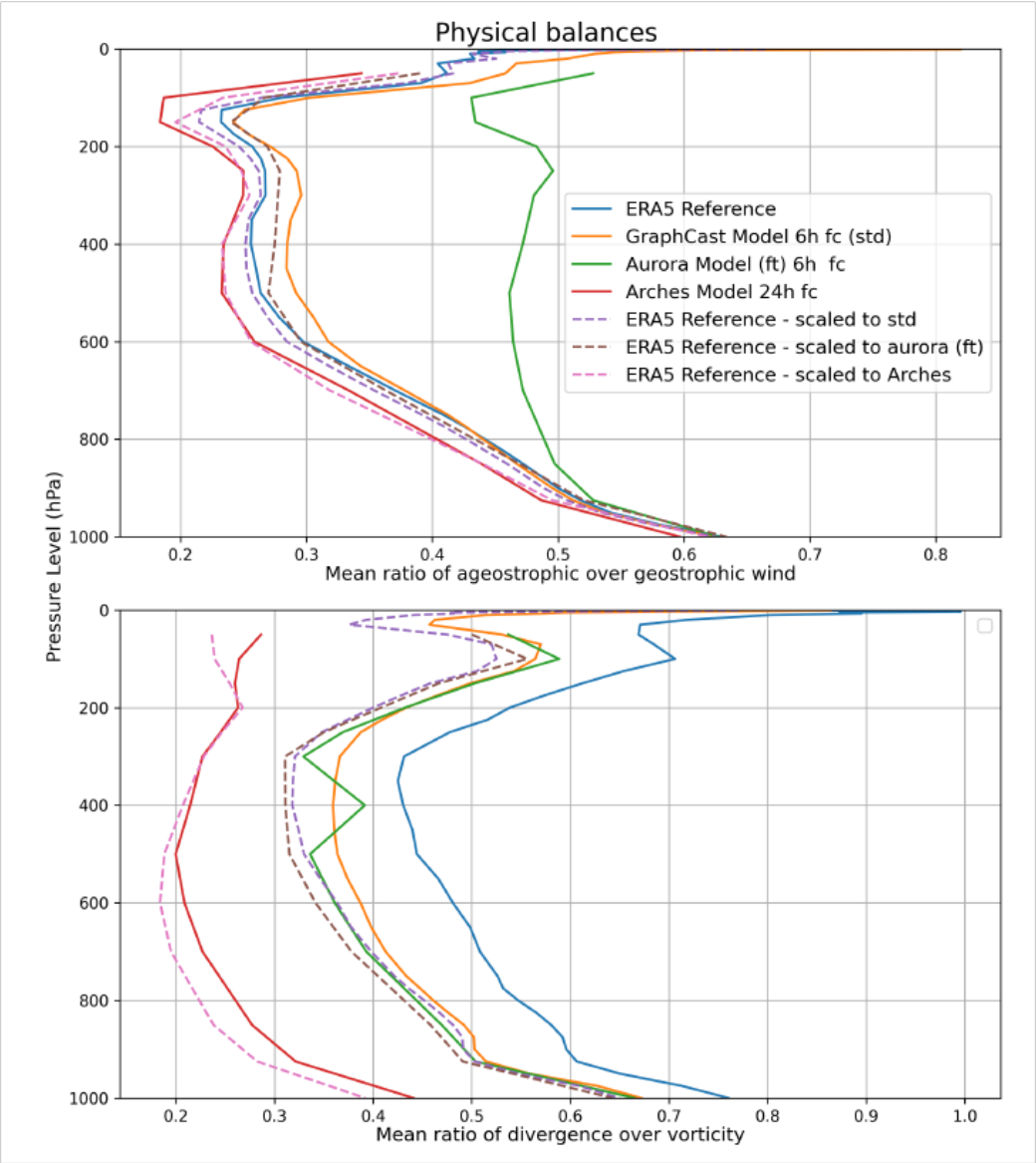




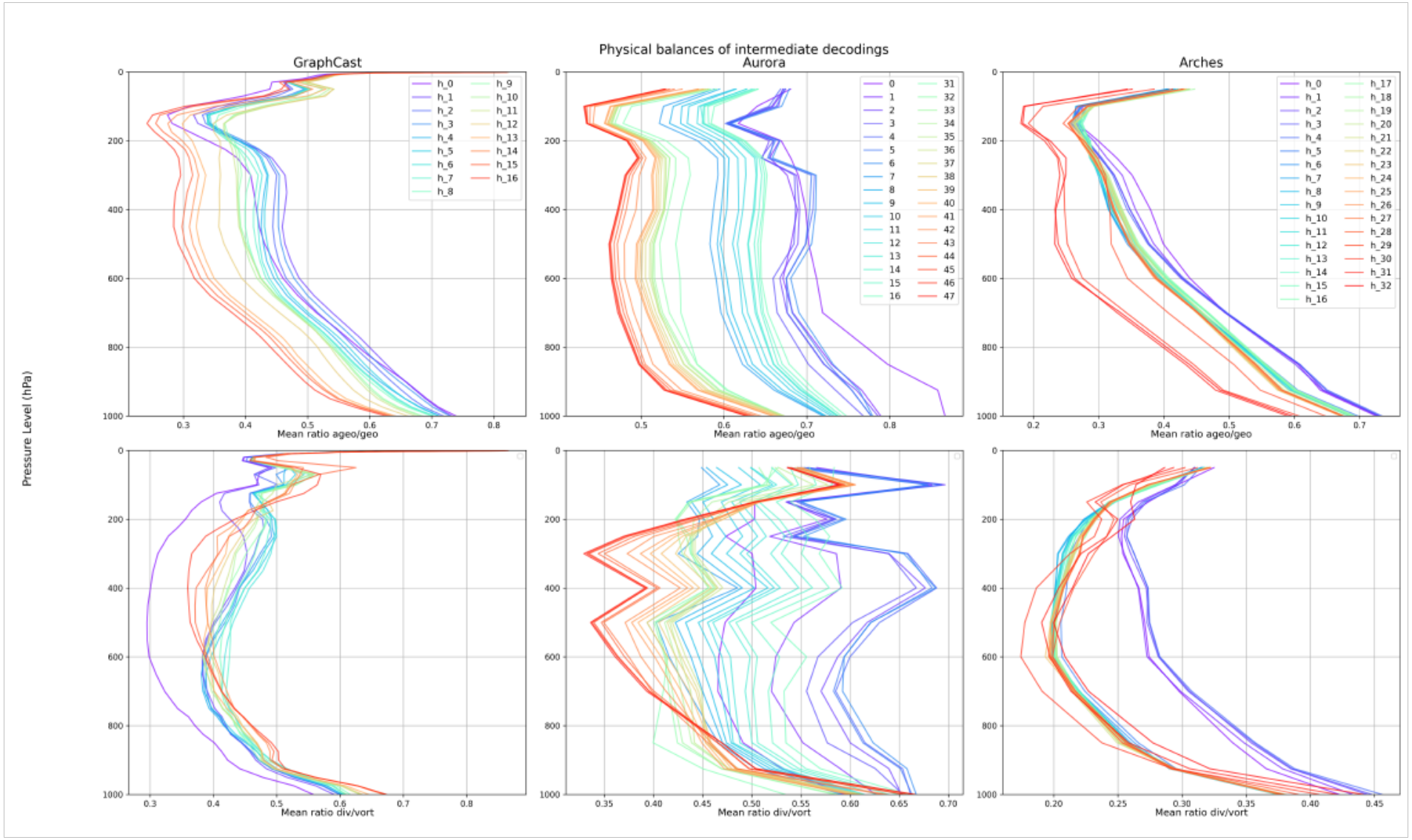
Scales


GraphCast std
Aurora
Spectral centroid
ArchesWeather
geopot at 500 hPa
u at 500 hPa
v at 500 hPa
temp at 500 hPa
qv at 500 hPa
w at 500 hPa
u_10m
v_10m
t_2m
pres_msl
Processor block

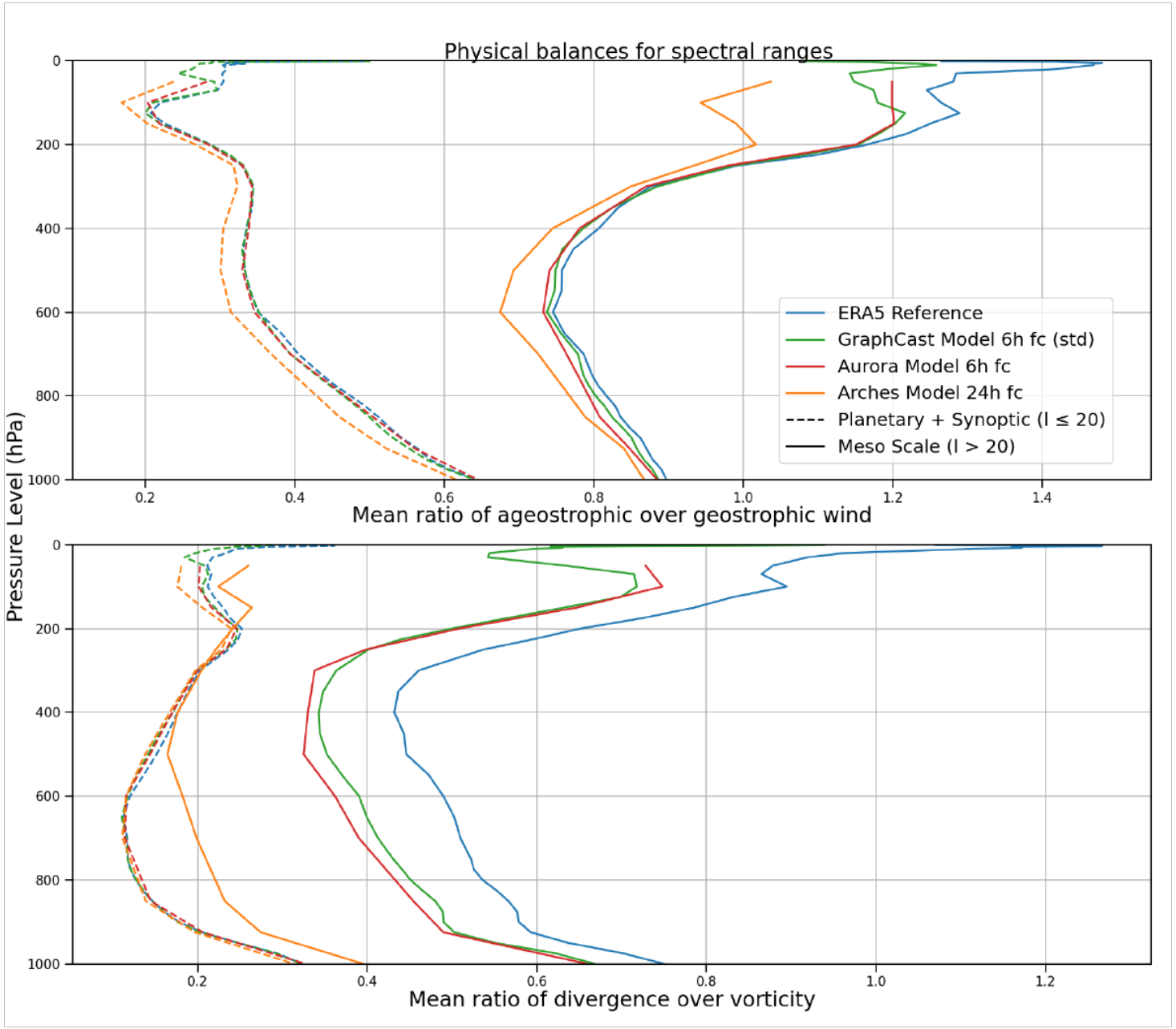
Physical balance



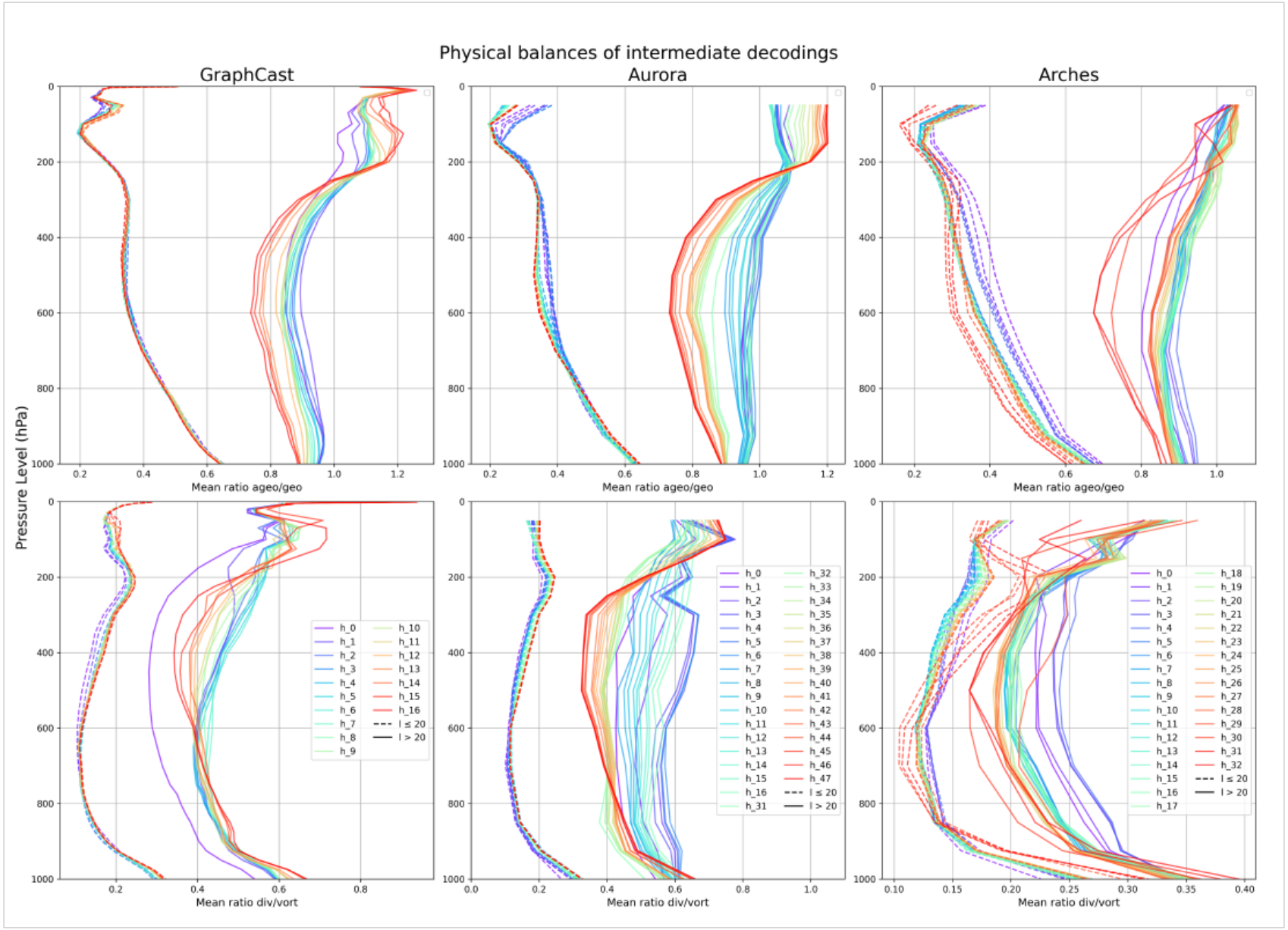
Physical balance



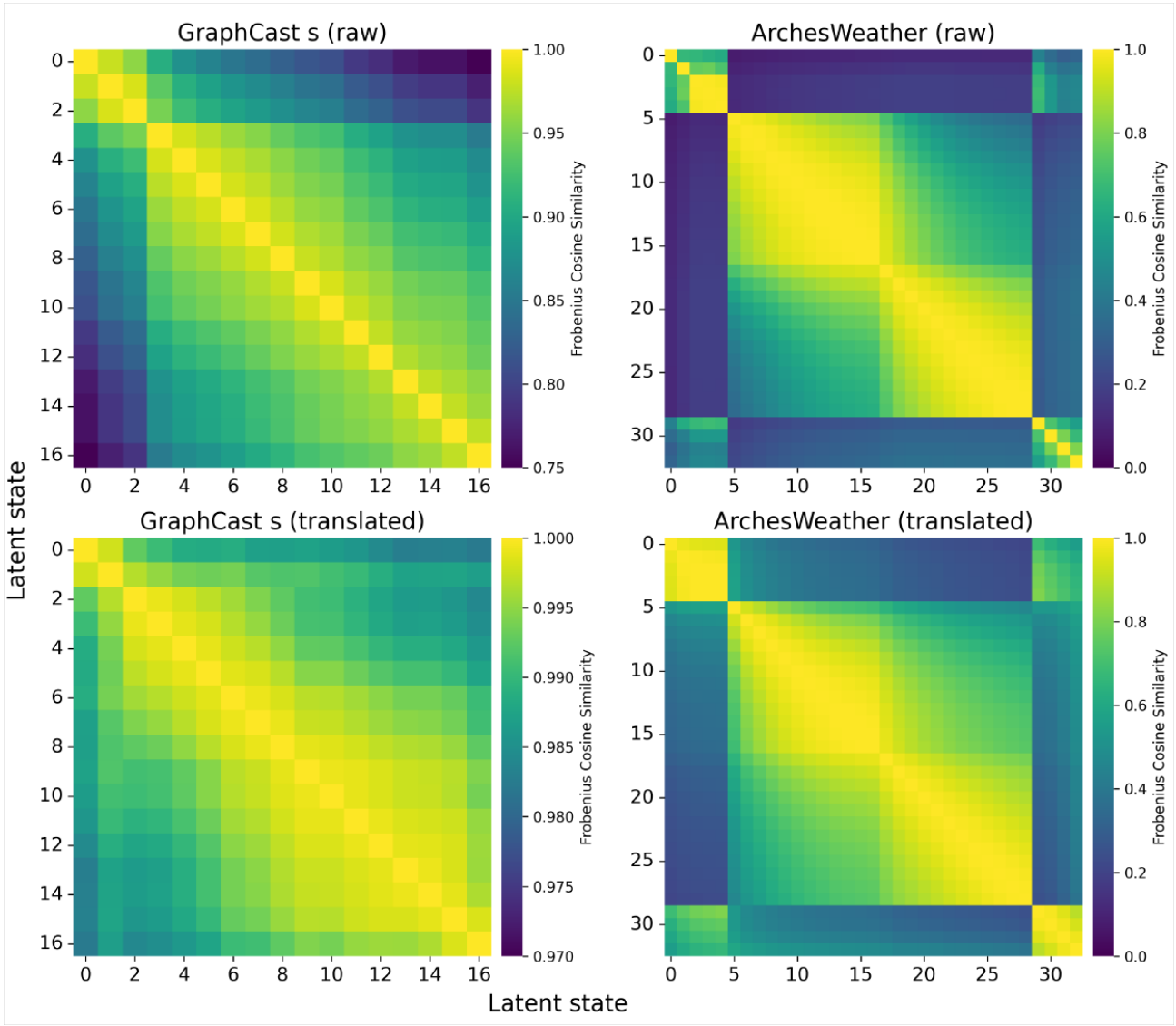
Physical balance



Physical balance

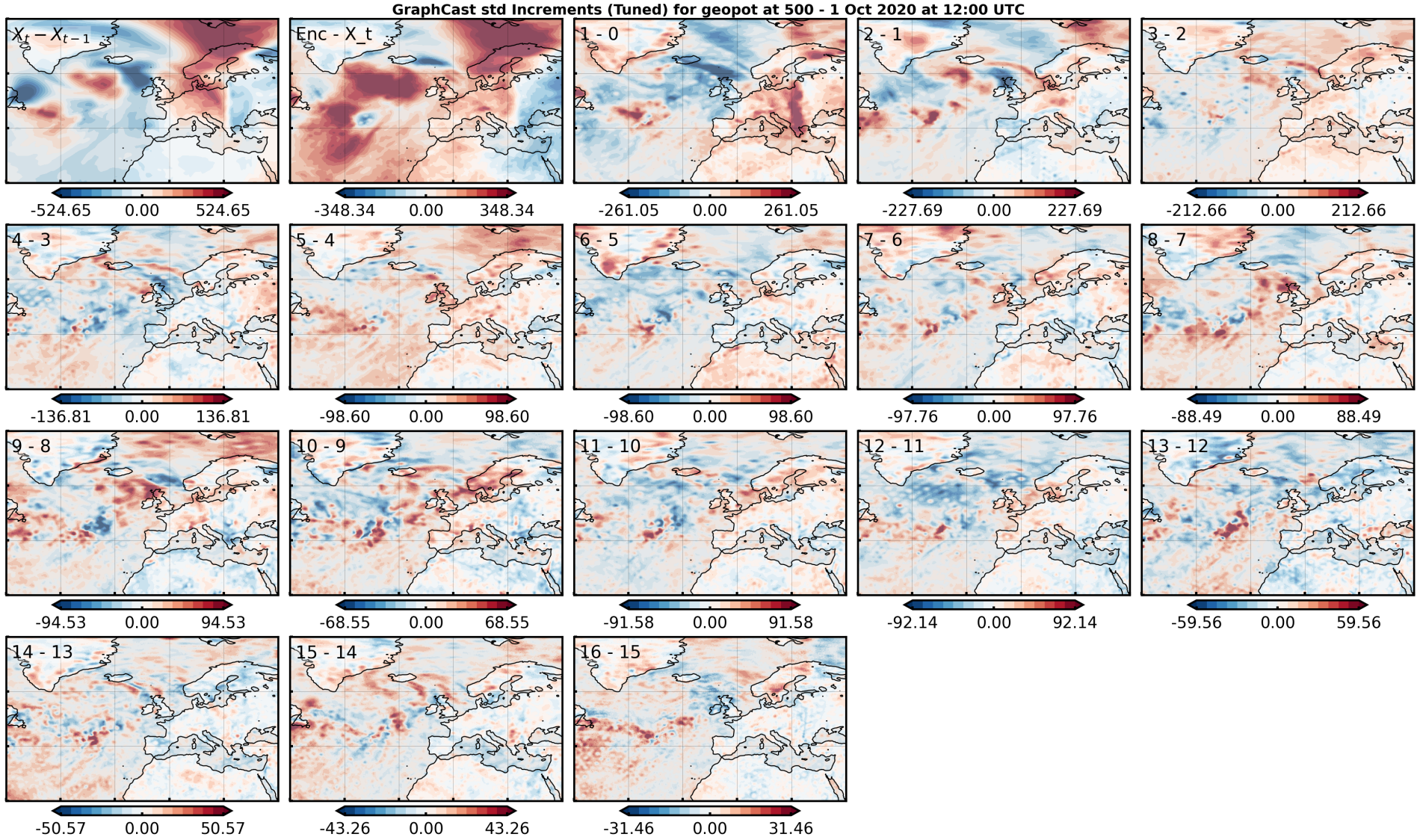


Representational stability

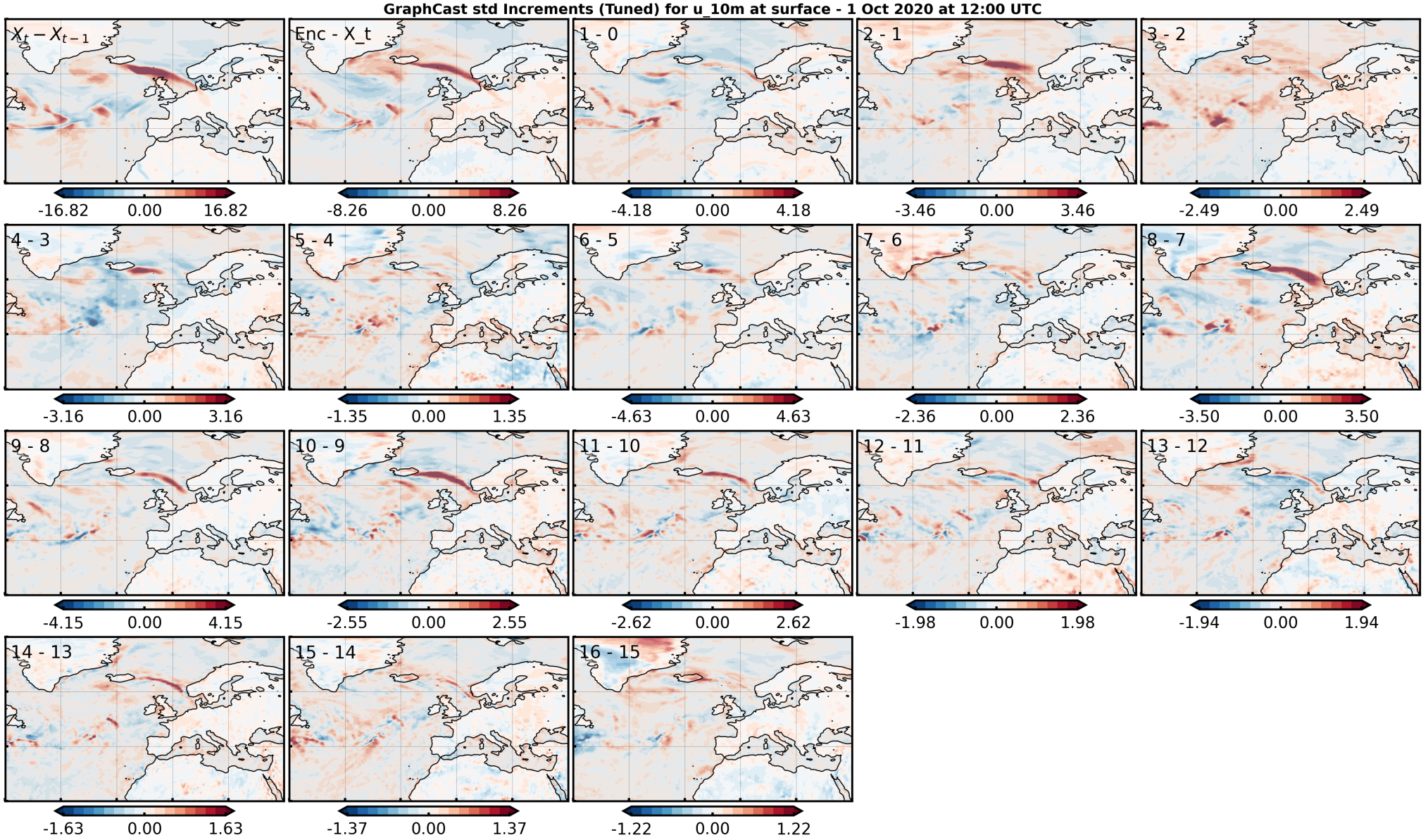


$$\cos_F(A,B) = \frac{\langle A,B\rangle_F}{\|A\|_F\,\|B\|_F},$$

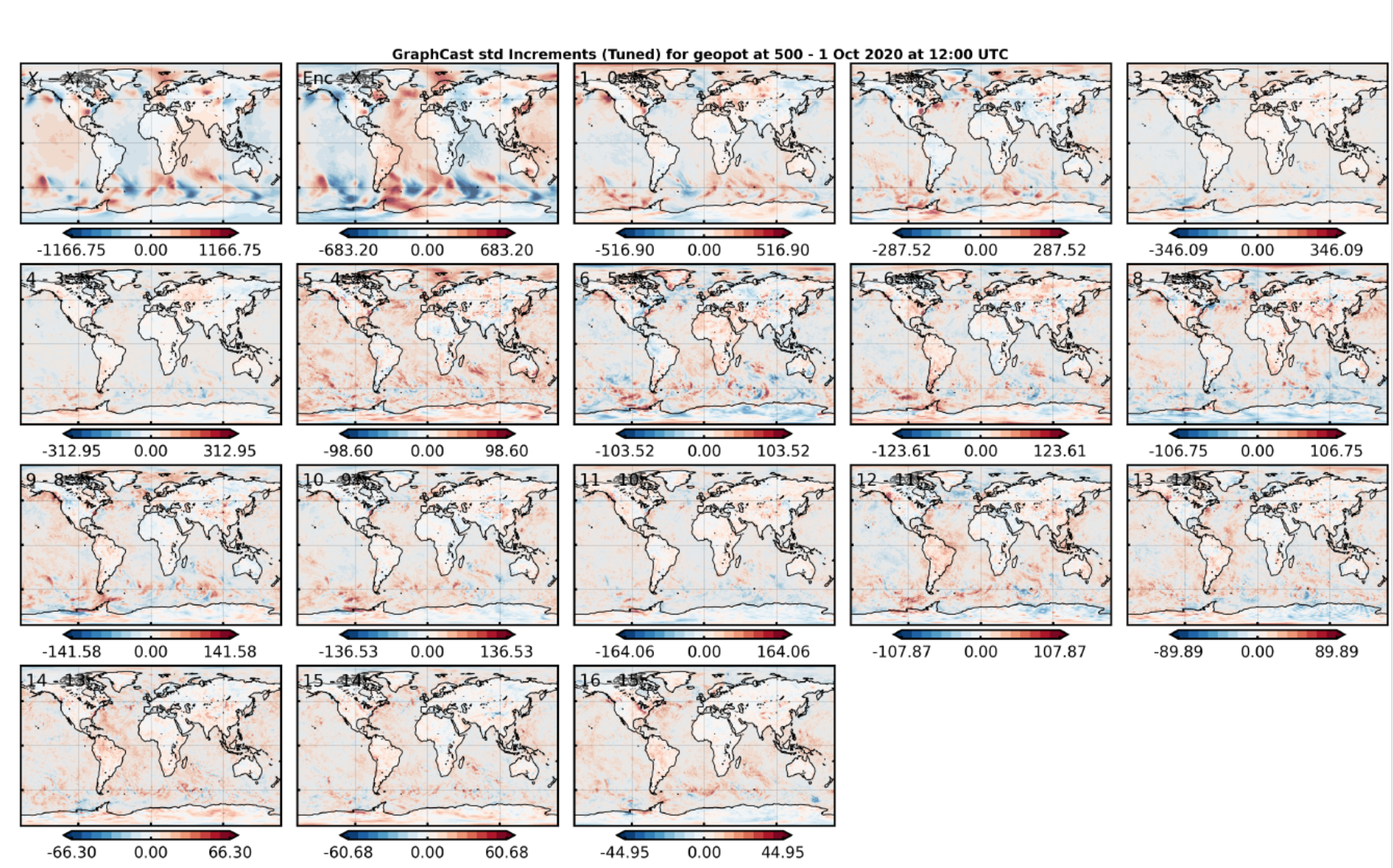
GraphCast all increments



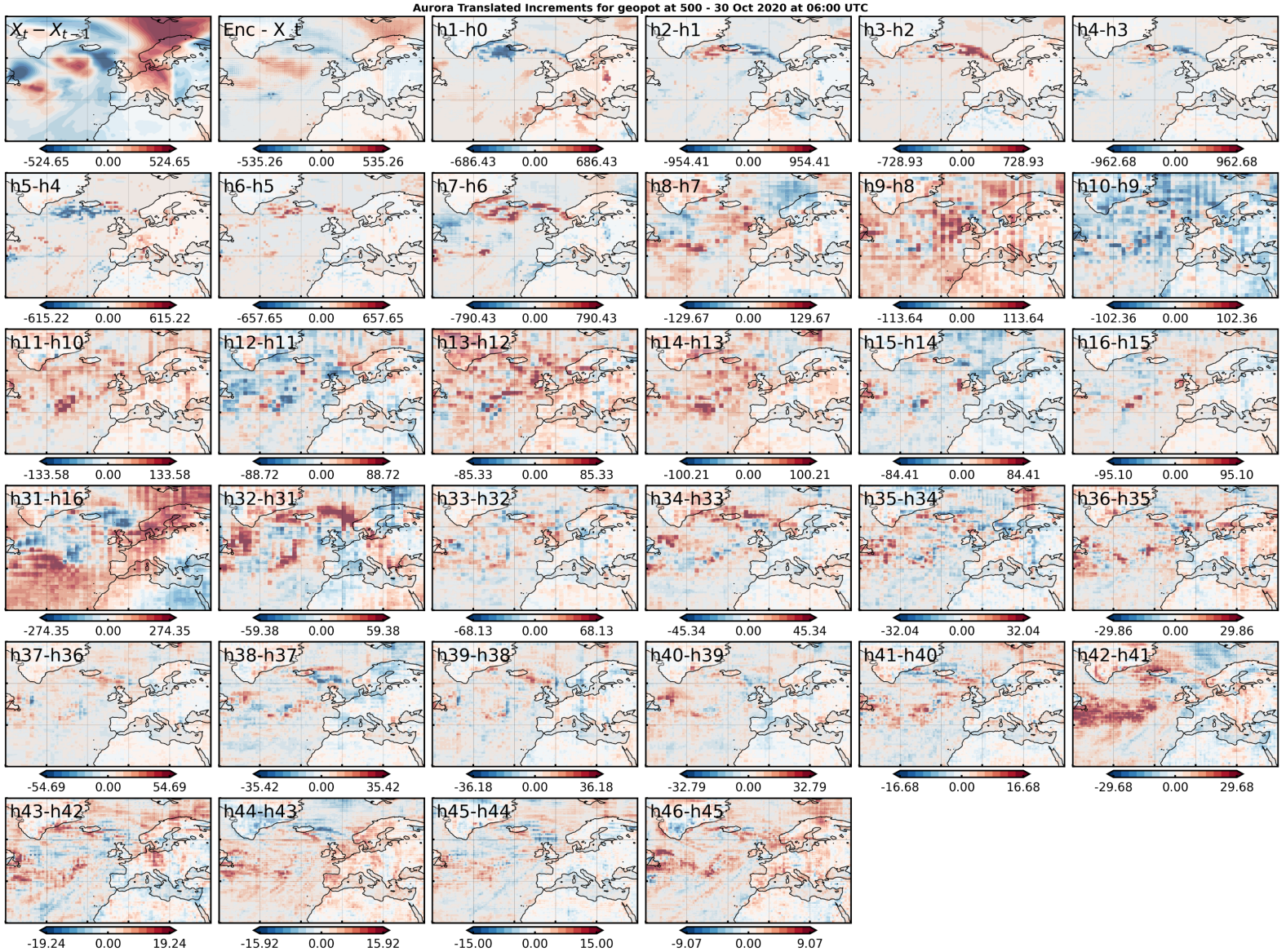
GraphCast all increments



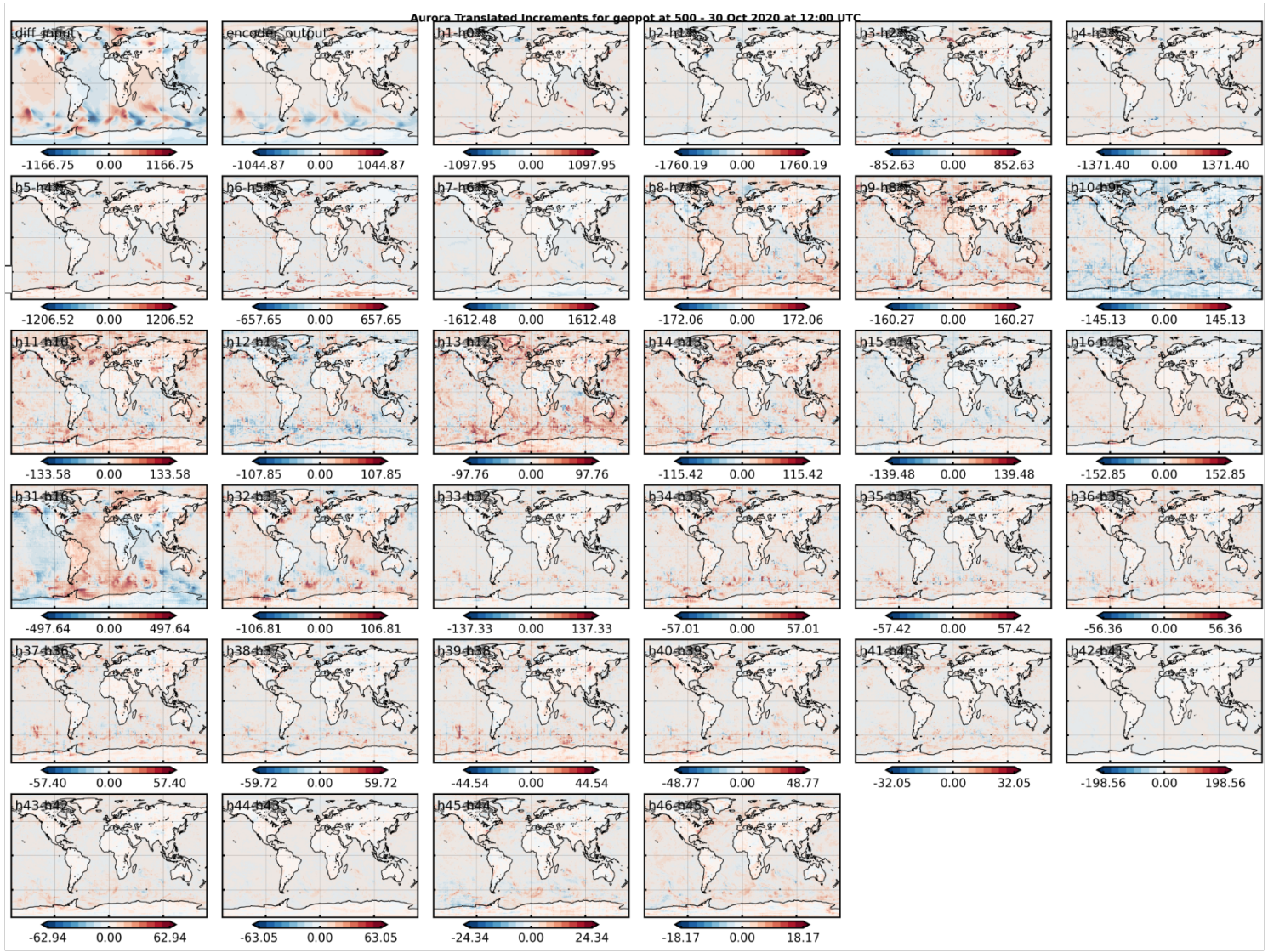
GraphCast all increments



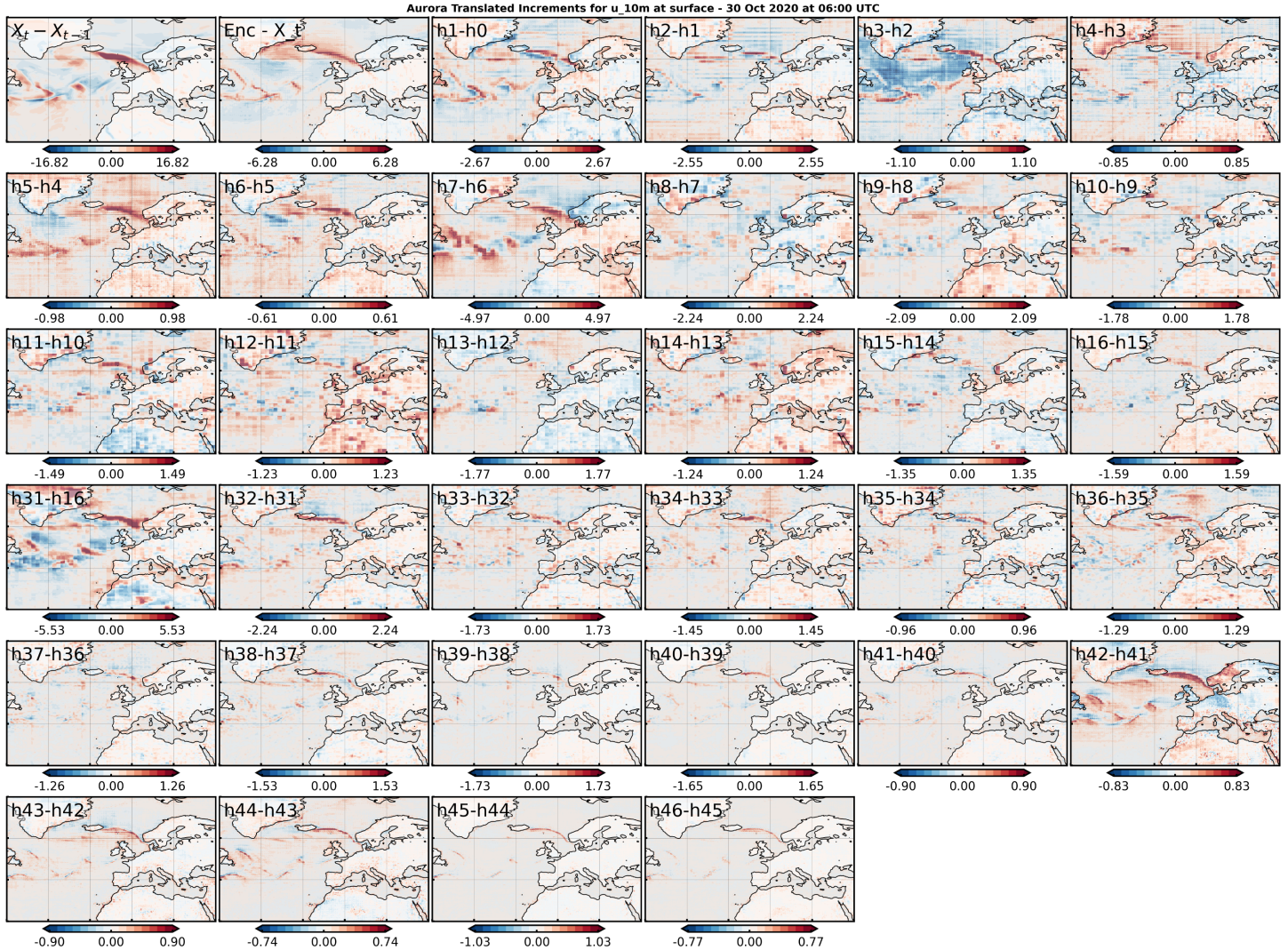
Aurora all increments



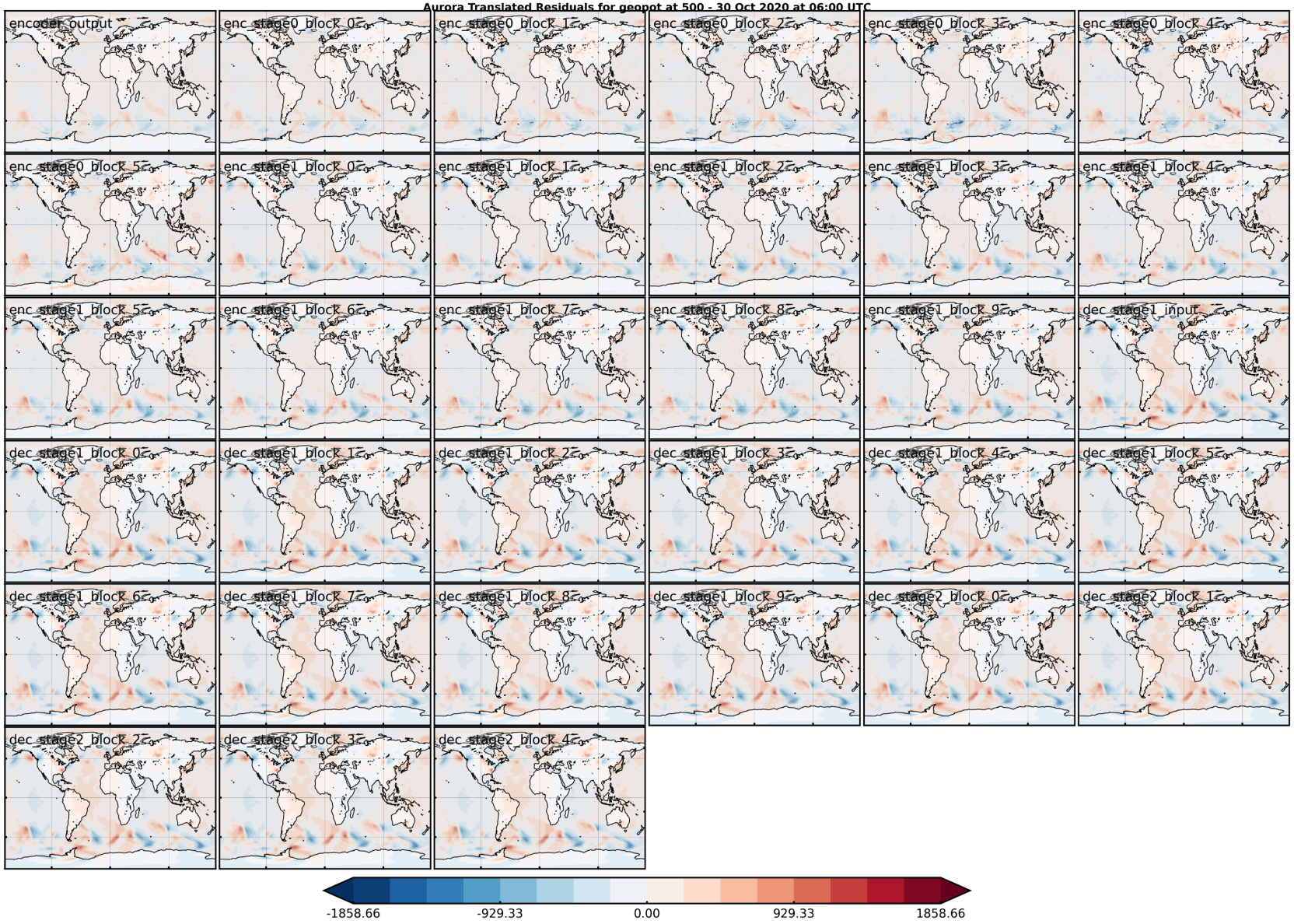
Aurora all increments



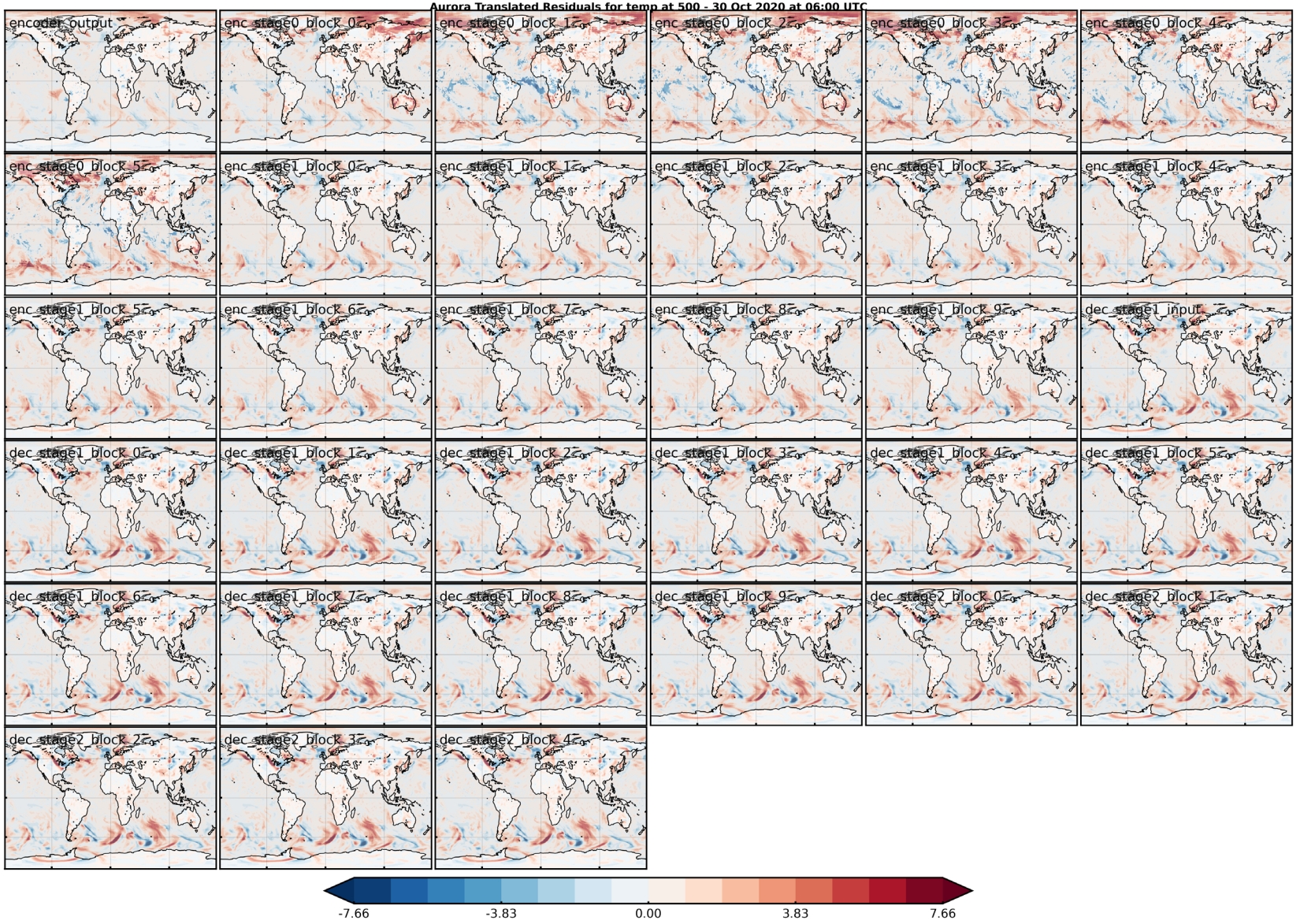
Aurora all increments



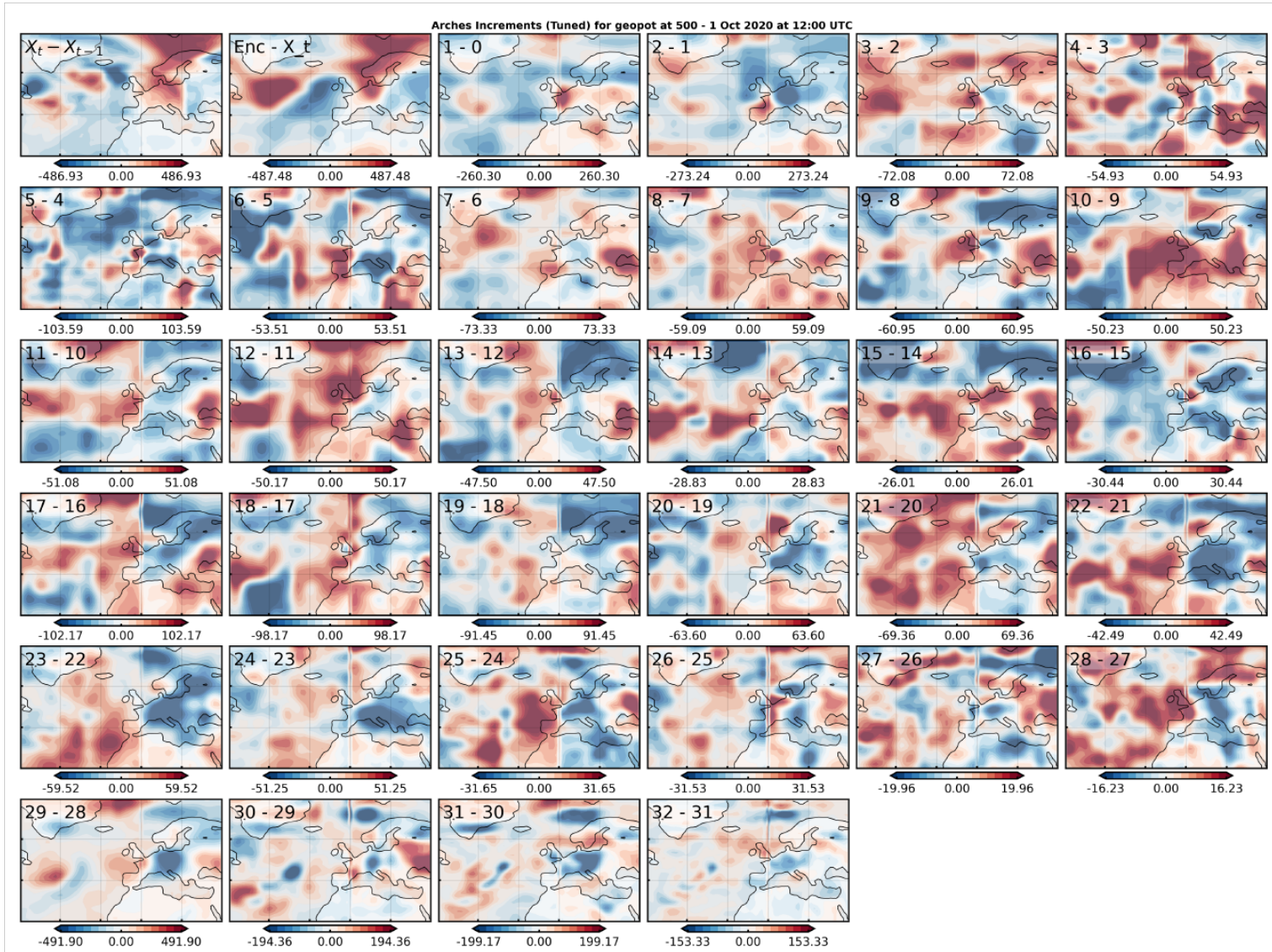
Aurora all increments



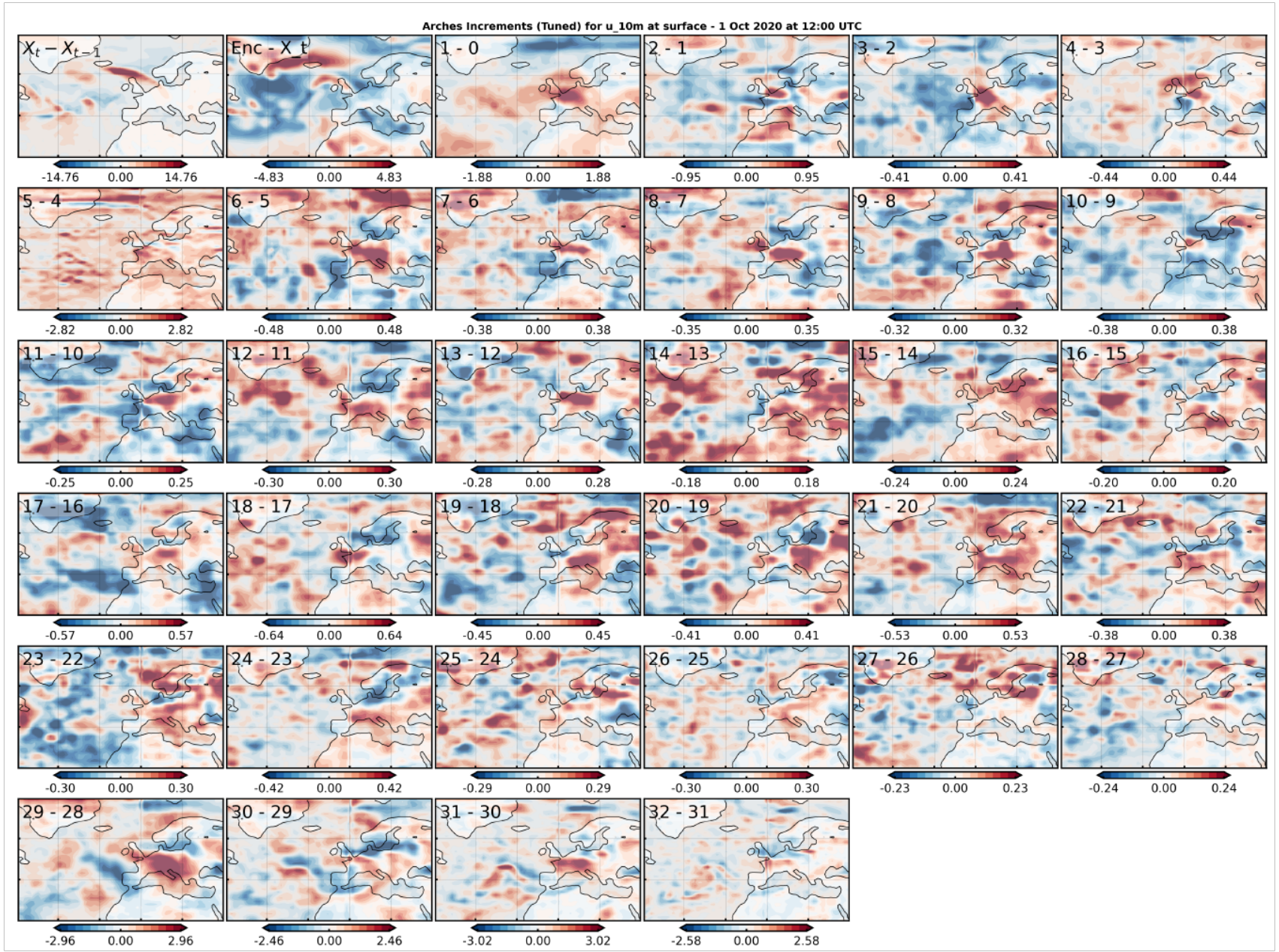
Aurora all increments



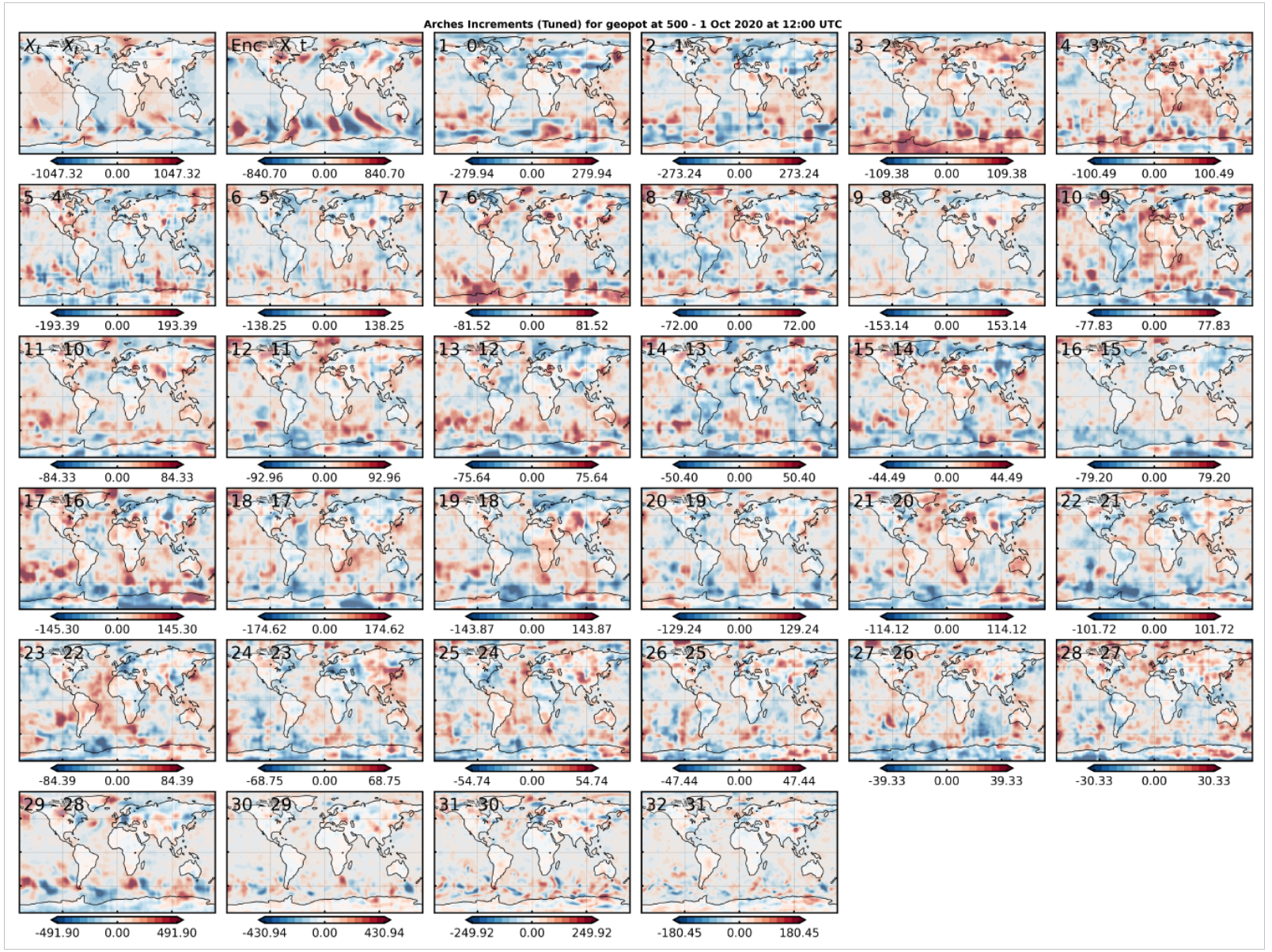
Arches all increments



Arches all increments



Arches all increments



Arches all increments