Sparse autoencoders applied to GraphCast

(and other techniques from the interp world)



Theo MacMillan

Causal Labs



WEATHER FORECASTING

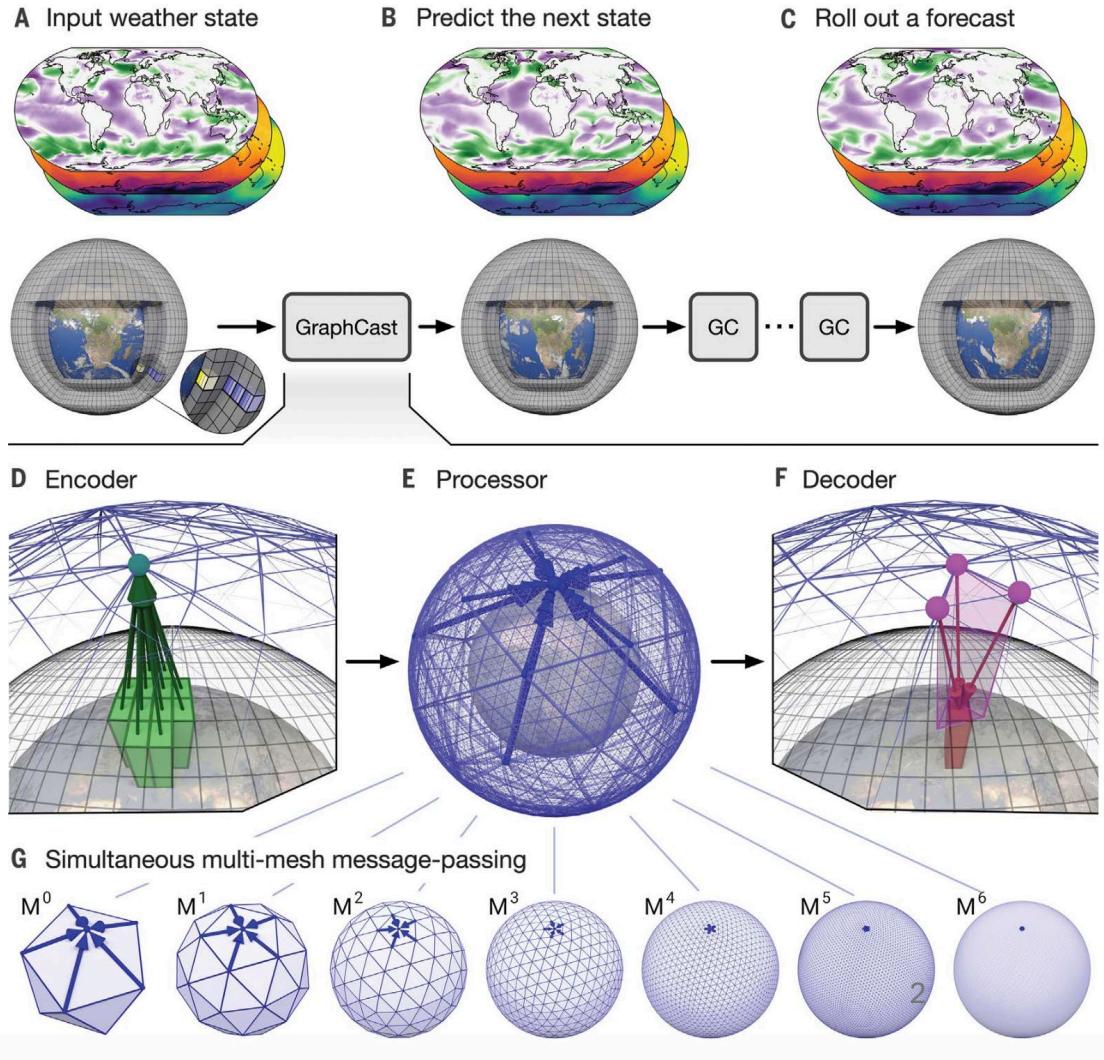
Learning skillful medium-range global weather forecasting

Remi Lam[1]*†, Alvaro Sanchez-Gonzalez[1]*†, Matthew Willson[1]*†, Peter Wirnsberger[1]†, Meire Fortunato[1]†, Ferran Alet[1]†, Suman Ravuri[1]†, Timo Ewalds[1], Zach Eaton-Rosen[1], Weihua Hu[1], Alexander Merose[2], Stephan Hoyer[2], George Holland[1], Oriol Vinyals[1], Jacklynn Stott[1], Alexander Pritzel[1], Shakir Mohamed[1]*, Peter Battaglia[1]*

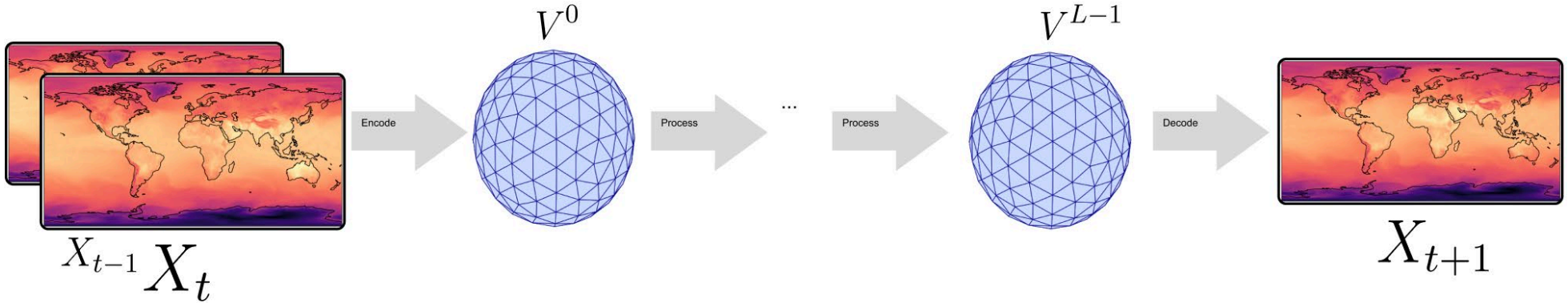
Global medium-range weather forecasting is critical to decision-making across many social and economic domains. Traditional numerical weather prediction uses increased compute resources to improve forecast accuracy but does not directly use historical weather data to improve the underlying model. Here, we introduce GraphCast, a machine learning–based method trained directly from reanalysis data. It predicts hundreds of weather variables for the next 10 days at 0.25° resolution globally in under 1 minute. GraphCast significantly outperforms the most accurate operational deterministic systems on 90% of 1380 verification targets, and its forecasts support better severe event prediction, including tropical cyclone tracking, atmospheric rivers, and extreme temperatures. GraphCast is a key advance in accurate and efficient weather forecasting and helps realize the promise of machine learning for modeling complex dynamical systems.

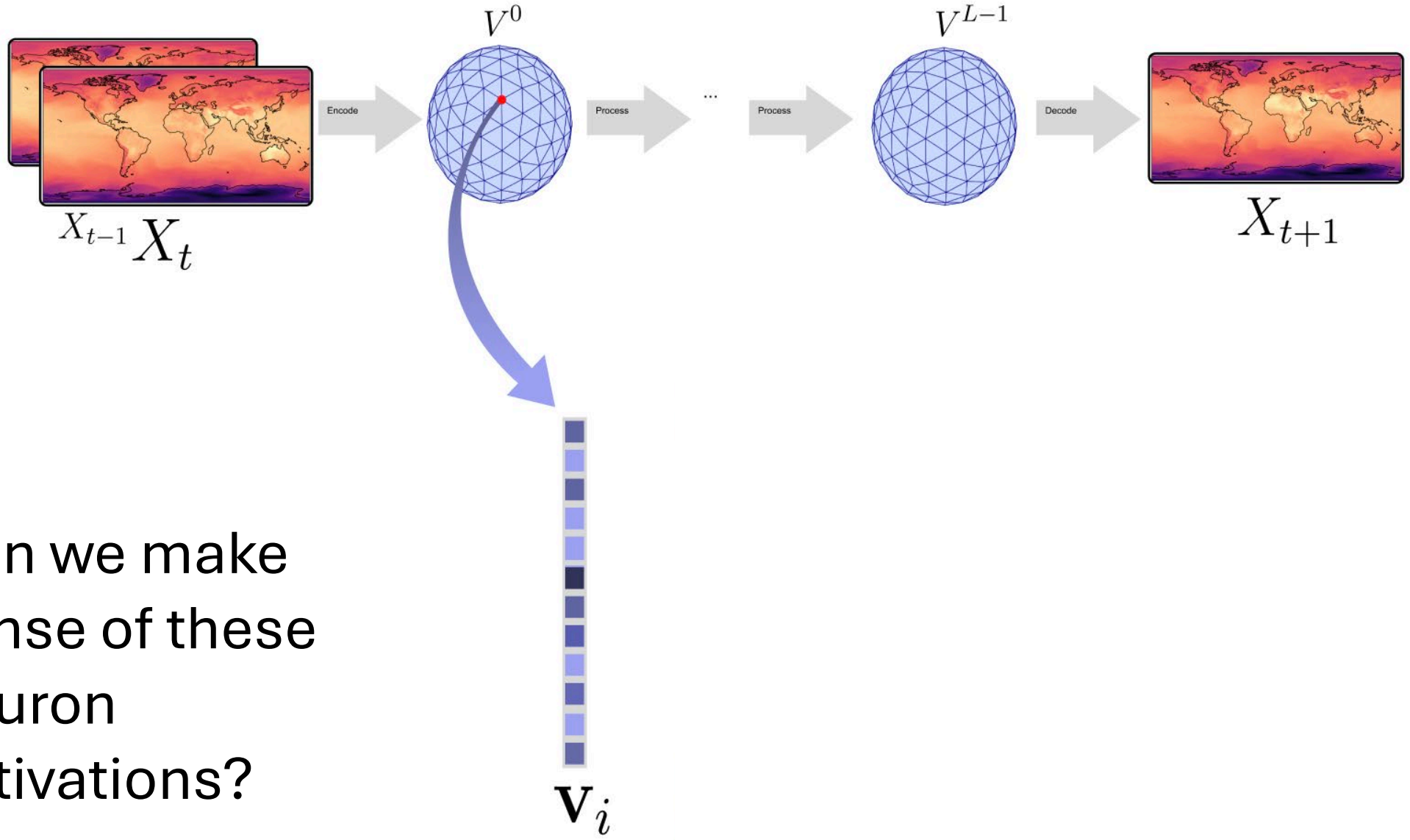
Surface variables (5)	Atmospheric variables (6)	Pressure levels (37)
2-meter temperature (2T)	**Temperature** (T)	1, 2, 3, 5, 7, 10, 20, 30, **50**, 70,
10 metre u wind component (10U)	**U component of wind** (U)	**100**, 125, **150**, 175, **200**, 225,
10 metre v wind component (10V)	**V component of wind** (V)	**250**, **300**, 350, **400**, 450, **500**,
Mean sea-level pressure (MSL)	**Geopotential** (Z)	550, **600**, 650, **700**, 750, 775,
Total precipitation (TP)	**Specific humidity** (Q)	800, 825, **850**, 875, 900, **925**,
	Vertical wind speed (W)	950, 975, **1000**

Table 1 | **Weather variables and levels modeled by GraphCast.** The numbers in parentheses in the column headings are the number of entries in the column. Boldfaced variables and levels indicates those which were included in the scorecard evaluation.



GraphCast





Can we make sense of these neuron activations?

Linear representation hypothesis

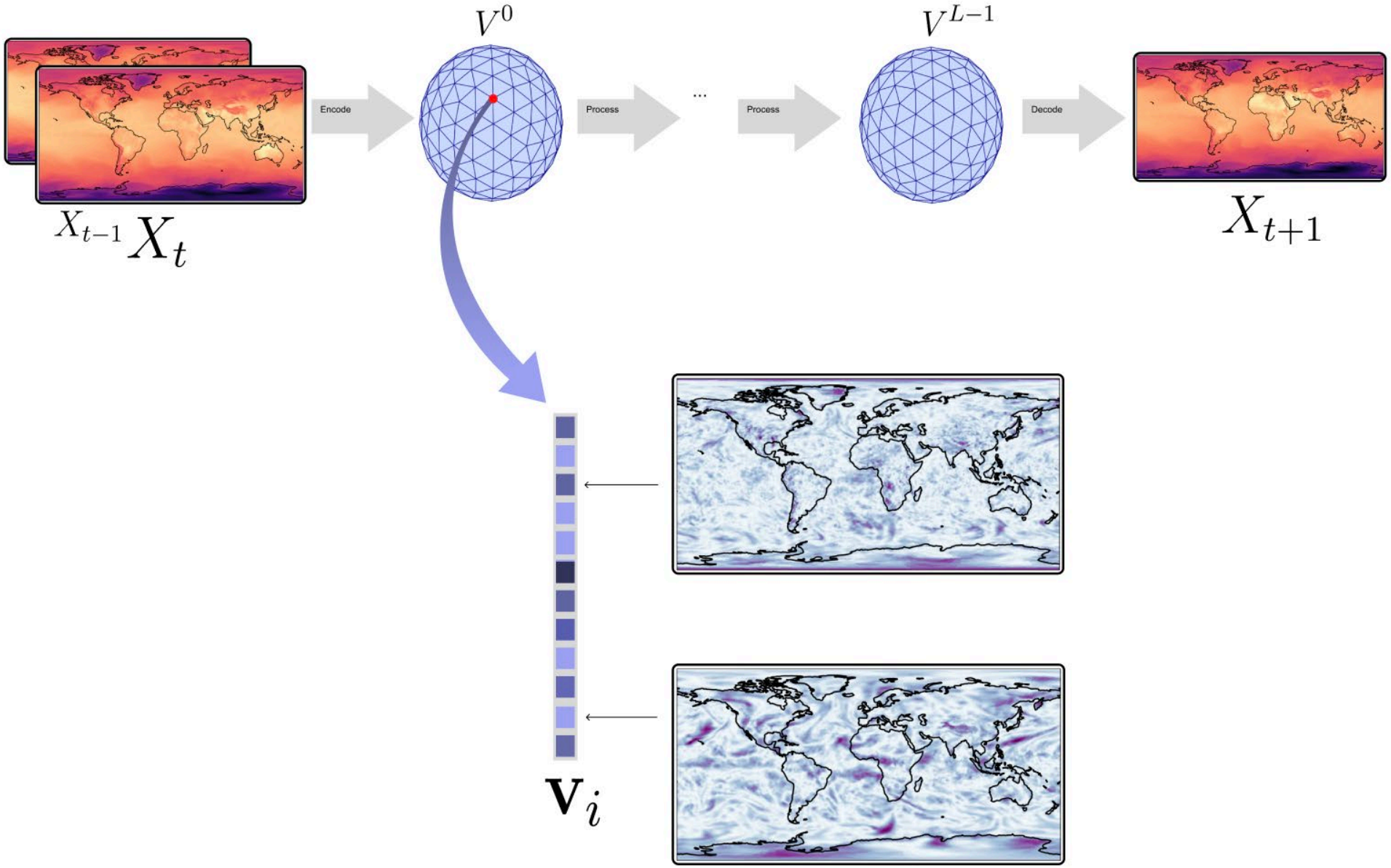
Vector of neurons

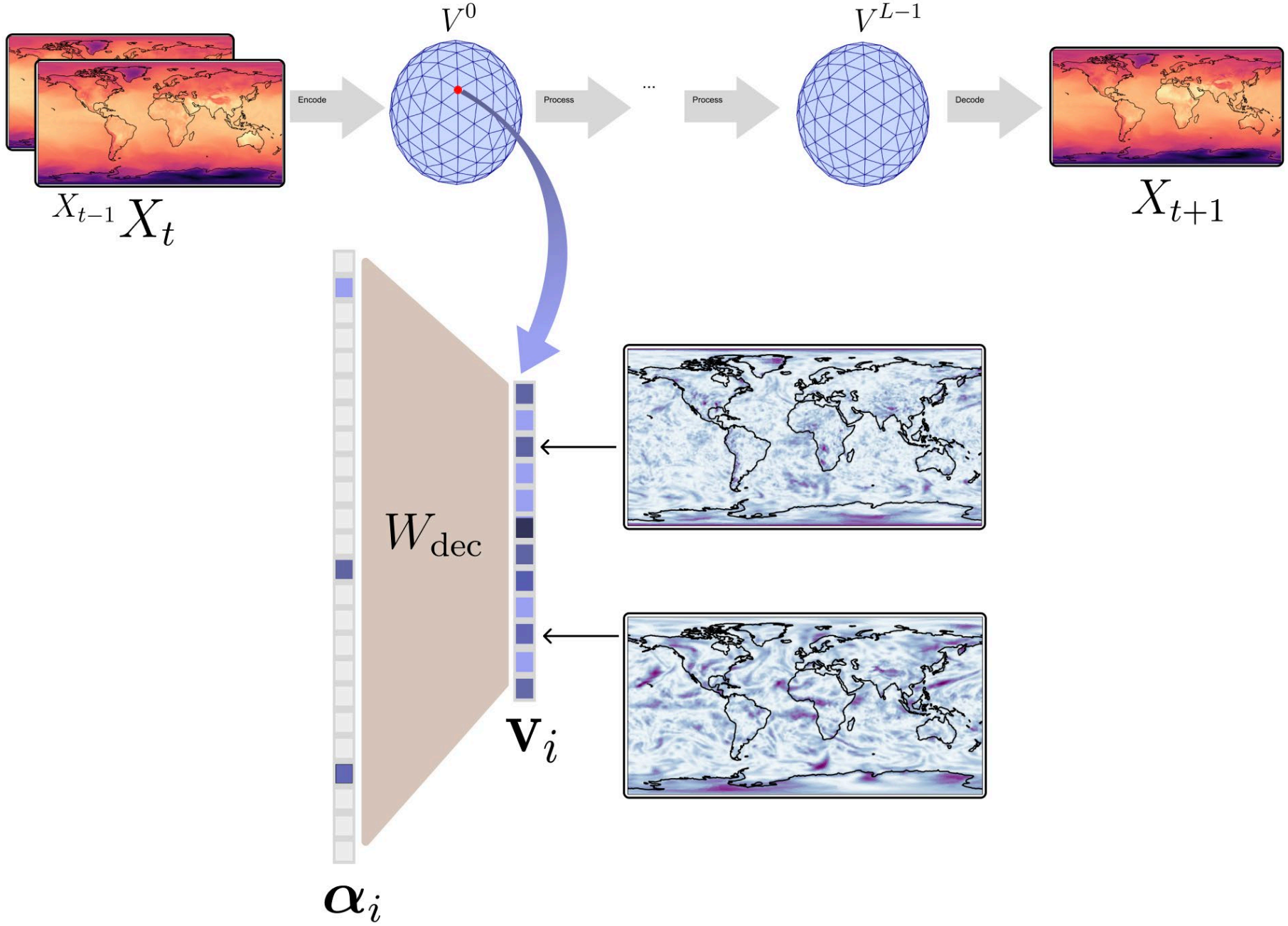


$$\mathbf{v}_i \approx W\boldsymbol{\alpha}_i + \mathbf{b}, \quad ||\boldsymbol{\alpha}_i||_0 \leq k$$

$\mathbf{v}_i$

Vector of features


V^0
V^{L-1}
Encode
Process
...
Process
Decode
X_{t-1} X_t
X_{t+1}
$\mathbf{v}_i$


V^0
V^{L-1}
Encode
Process
...
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{dec}
$\mathbf{v}_i$
$\boldsymbol{\alpha}_i$

The workhorse

Sparse Autoencoder

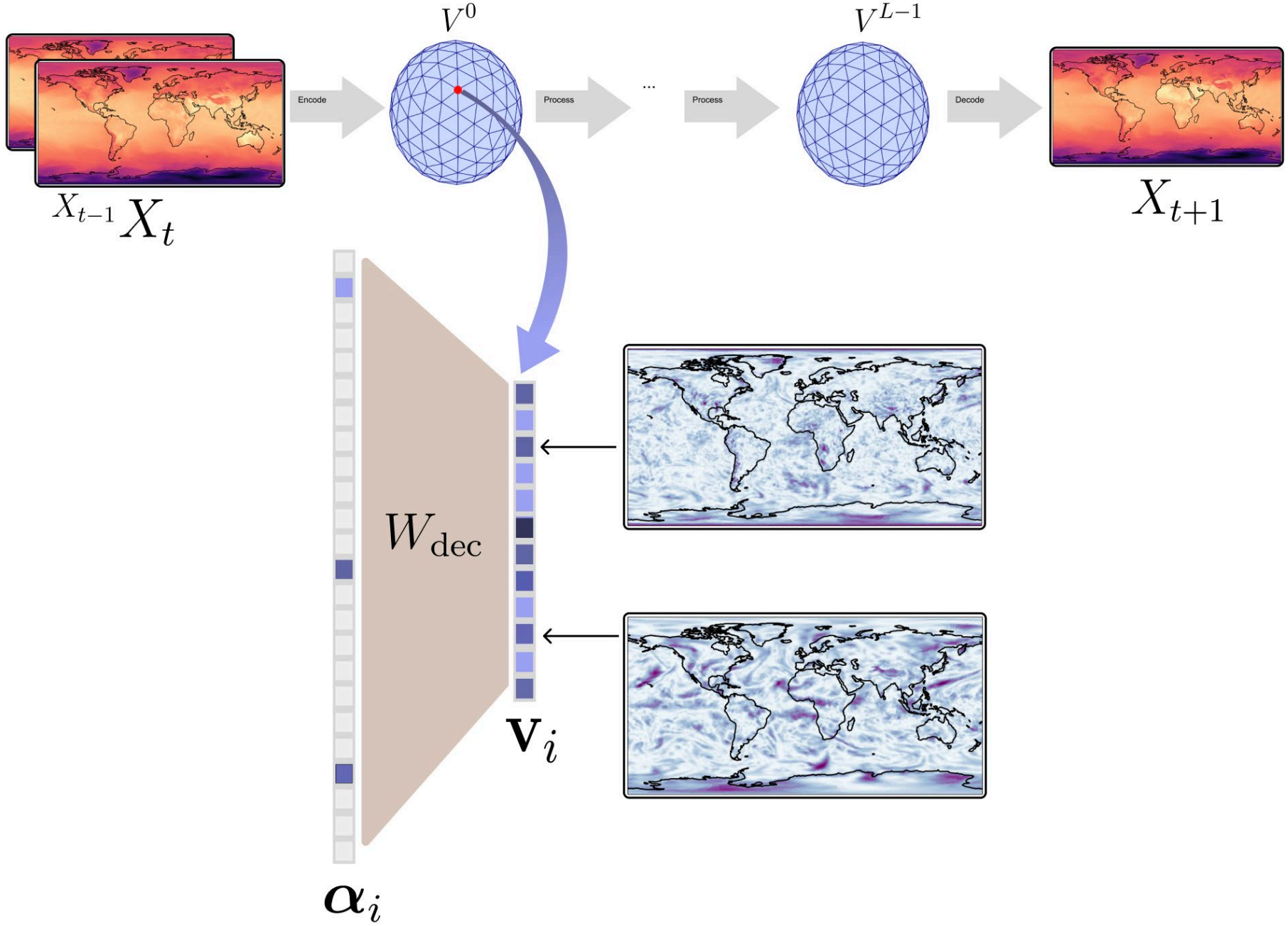
$$\boldsymbol{\alpha}_i = \mathrm{TopK}\big(W_{\mathrm{enc}}(\mathbf{v}_i - \mathbf{b})\big)$$

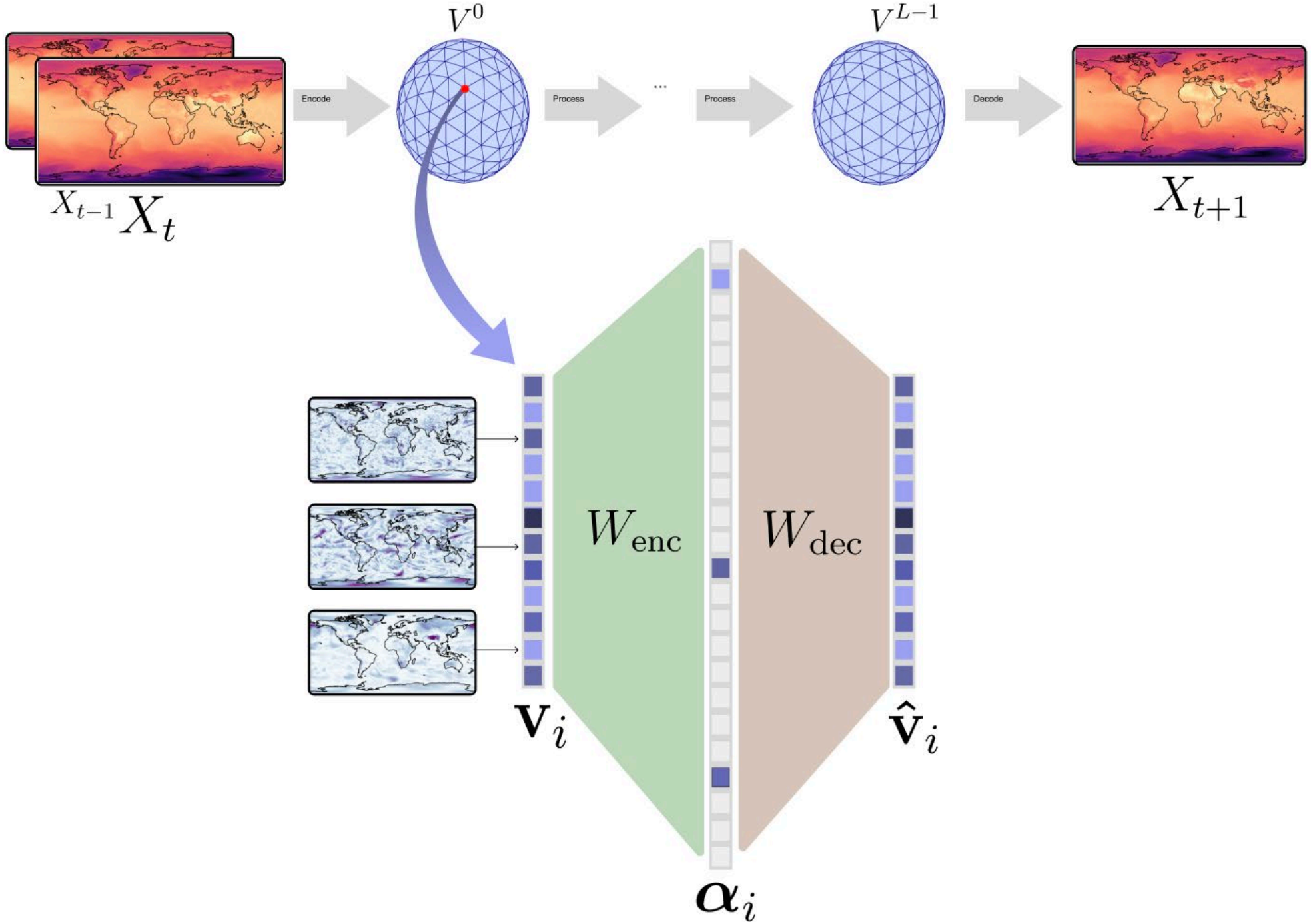
$$\hat{\mathbf{v}}_i = W_{dec}\boldsymbol{\alpha}_i + \mathbf{b}$$

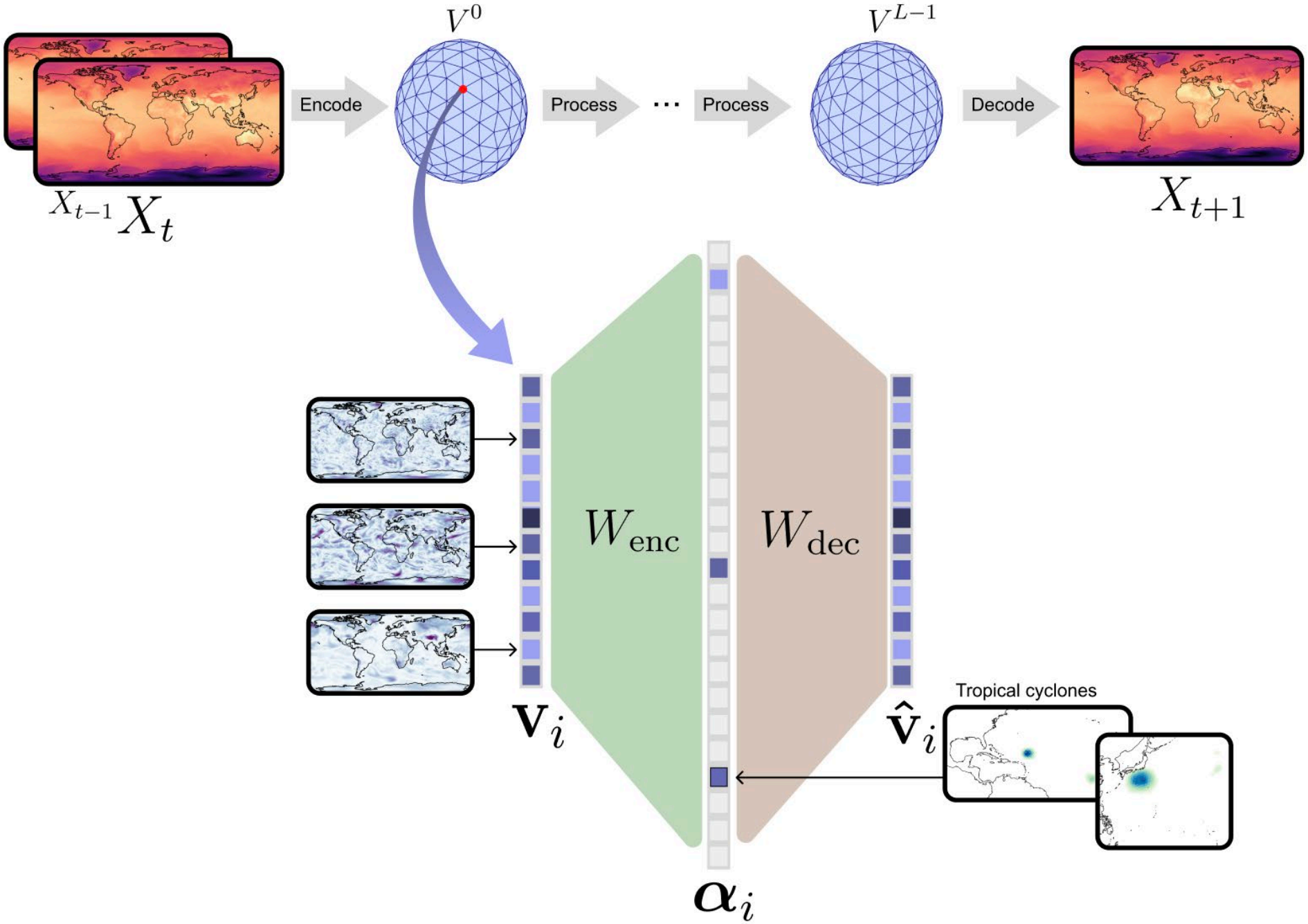
$$\mathbf{v}_i \in \mathbb{R}^d,\ \boldsymbol{\alpha}_i \in \mathbb{R}^f,\ ||\boldsymbol{\alpha}_i||_0 \leq k$$

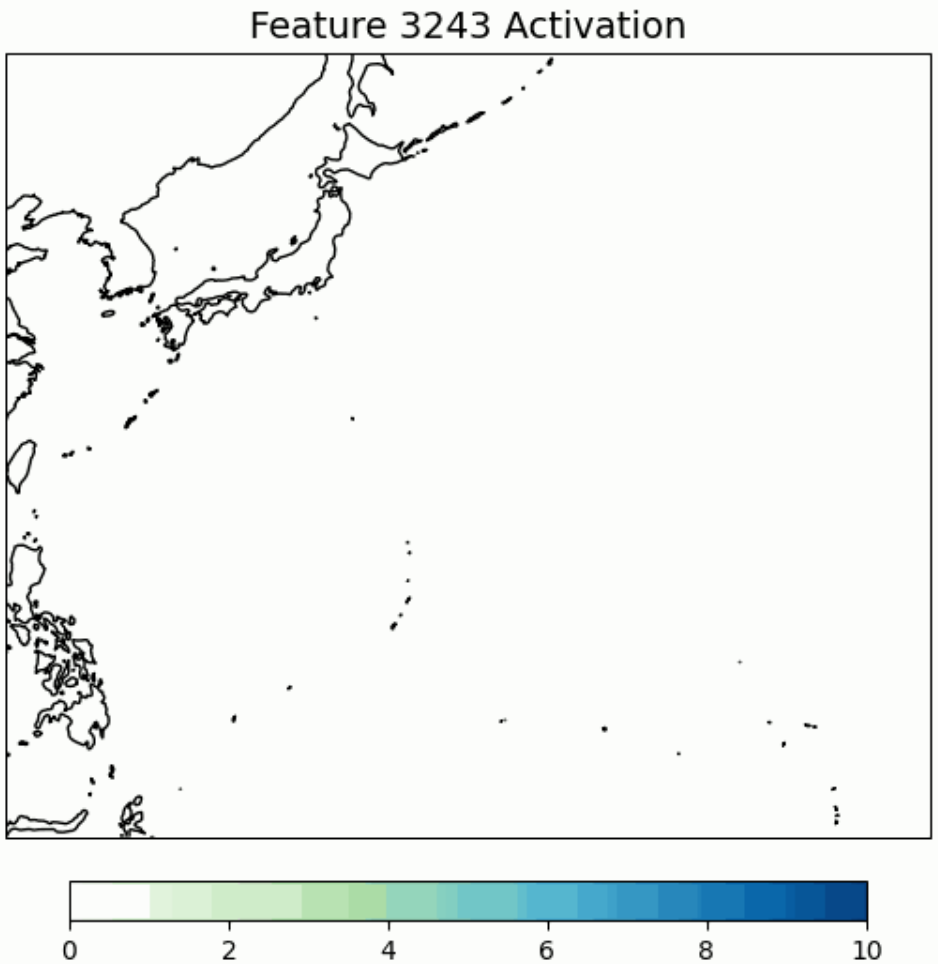
Typical values in GraphCast could be

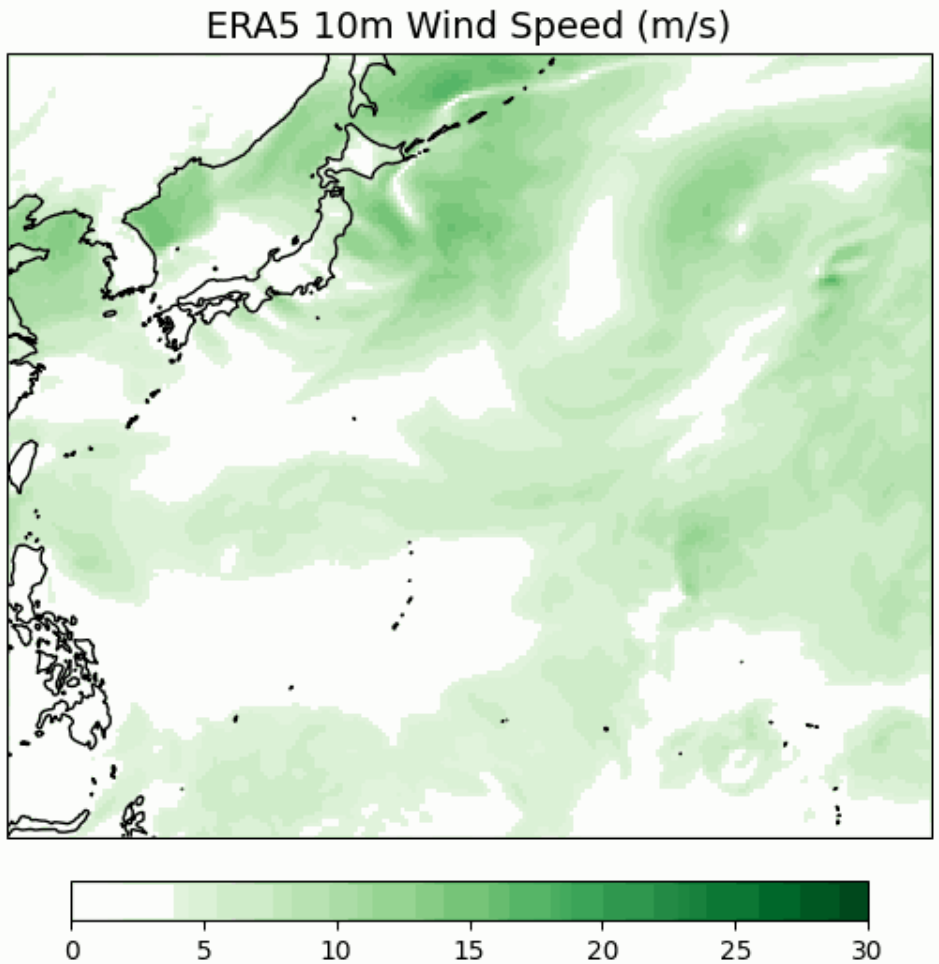
$$d = 512 \qquad \mathrm{f} = 8192 \qquad \mathrm{k} = 32$$


V^0
V^{L-1}
Encode
Process
...
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{dec}
$\mathbf{v}_i$
$\boldsymbol{\alpha}_i$


V^0
V^{L-1}
Encode
Process
...
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$

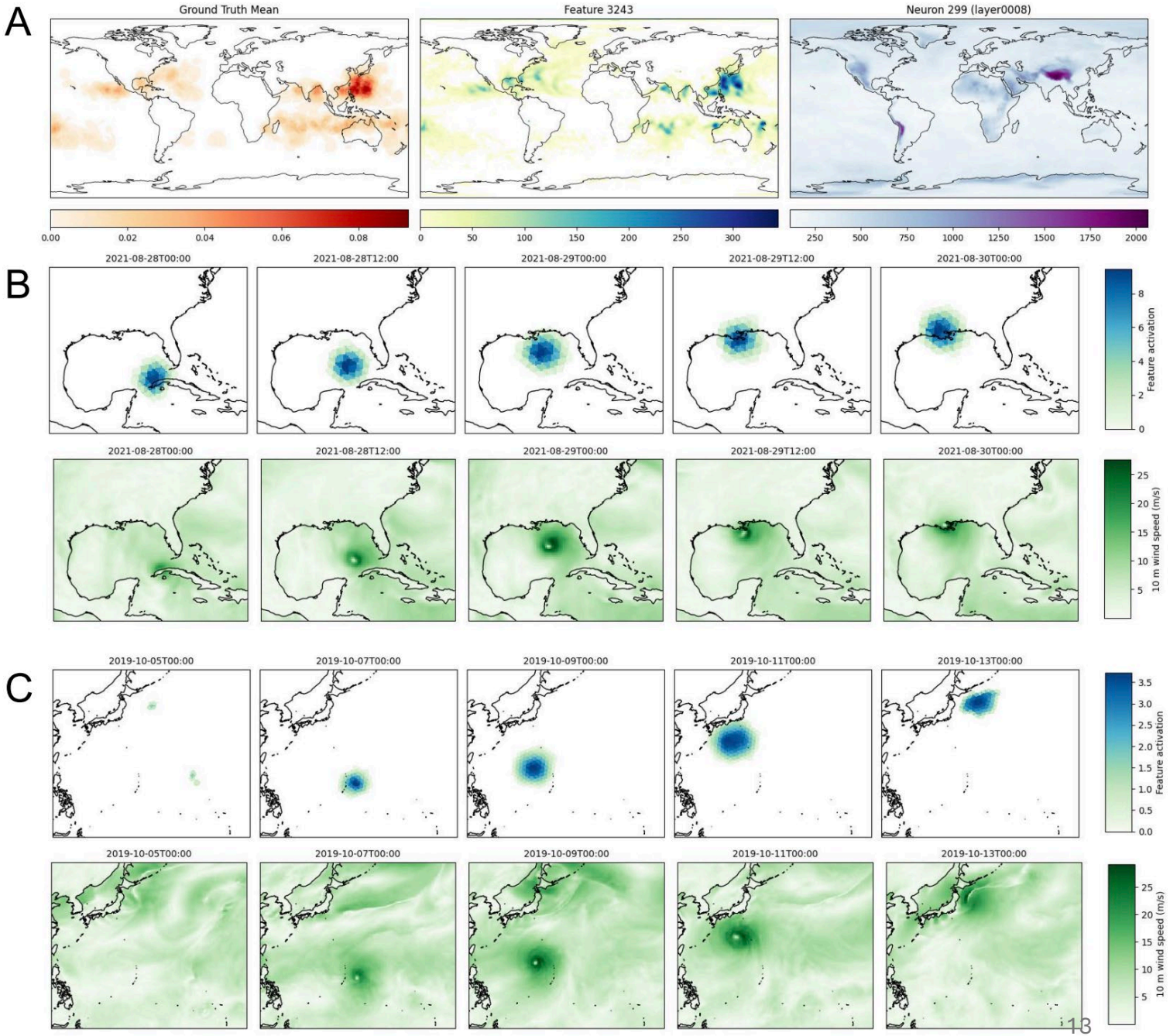

V^0
V^{L-1}
Encode
Process
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$
Tropical cyclones


Feature 3243 Activation
0
2
4
6
8
10


ERA5 10m Wind Speed (m/s)
0
5
10
15
20
25
30

Feature 3243 activates in the presence of tropical cyclones

But what is it *really*?



Causal Steering

Sparse Autoencoder

$$\boldsymbol{\alpha}_i = \mathrm{TopK}\big(W_{\mathrm{enc}}(\mathbf{v}_i - \mathbf{b})\big)$$
$$\hat{\mathbf{v}}_i = W_{dec}\boldsymbol{\alpha}_i + \mathbf{b}$$

1. Decompose residual stream

$$\mathbf{v}_i^l = \hat{\mathbf{v}}_i^l + (\mathbf{v}_i^l - \hat{\mathbf{v}}_i^l)$$

2. Apply encoder, modify feature with steering vector

$$\tilde{\alpha}_i^l = \mathbf{s}_{f,\gamma} \odot \alpha_i^l$$

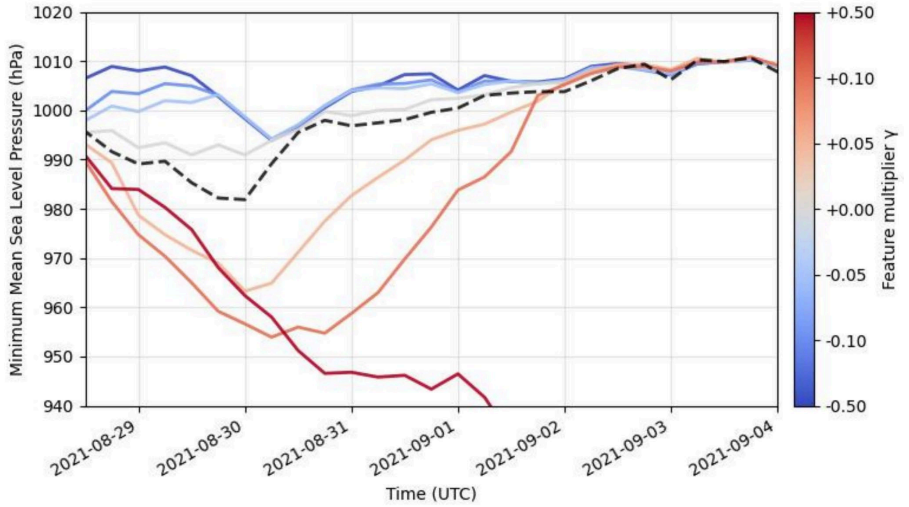
3. Decode modified features, to get modified residual stream

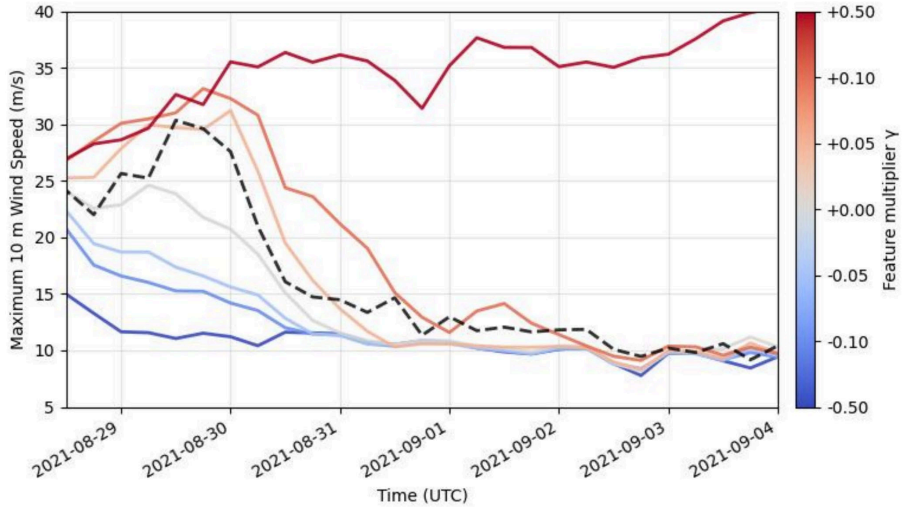
$$\tilde{\mathbf{v}}_i^l = W_{dec}\tilde{\alpha}_i^l + \mathbf{b}_{pre}$$

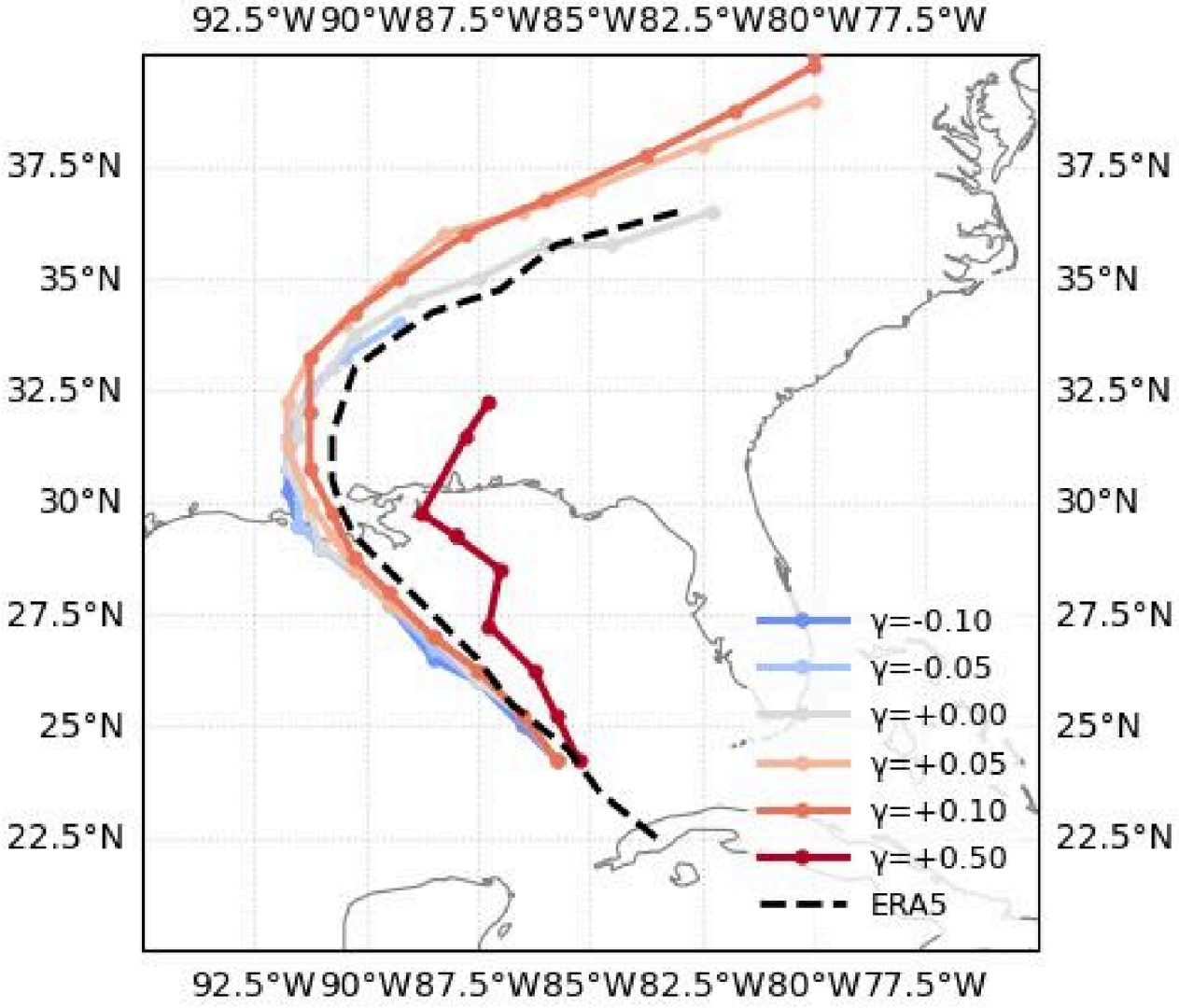
4. Add back

$$\mathbf{v}_{i,mod}^l = \tilde{\mathbf{v}}_i^l + (\mathbf{v}_i^l - \hat{\mathbf{v}}_i^l)$$

Feature 3243 controls the strength of evolving hurricanes





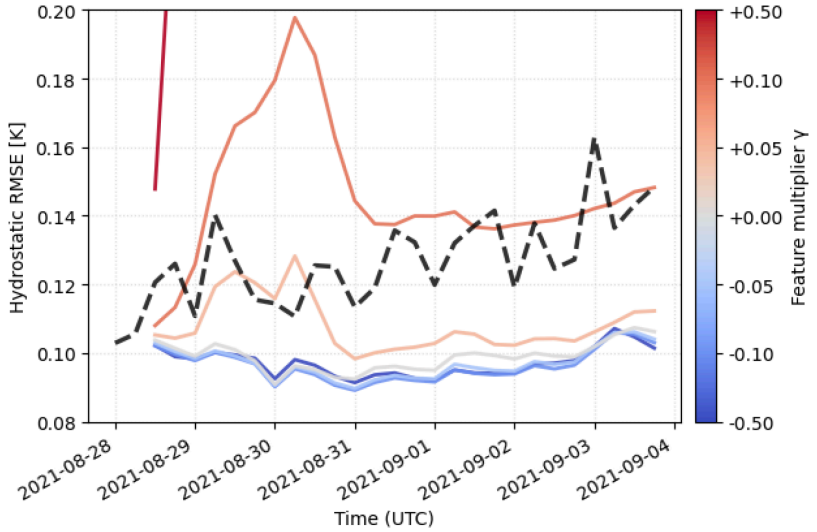


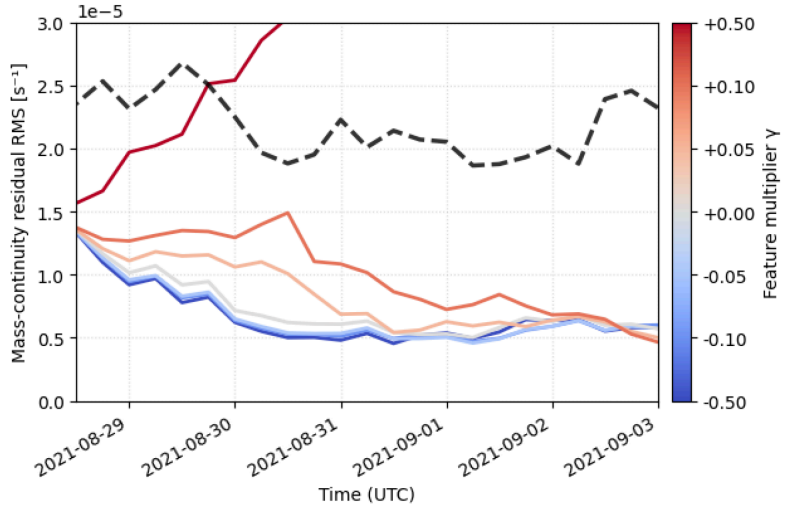
Can test physicality of resulting fields

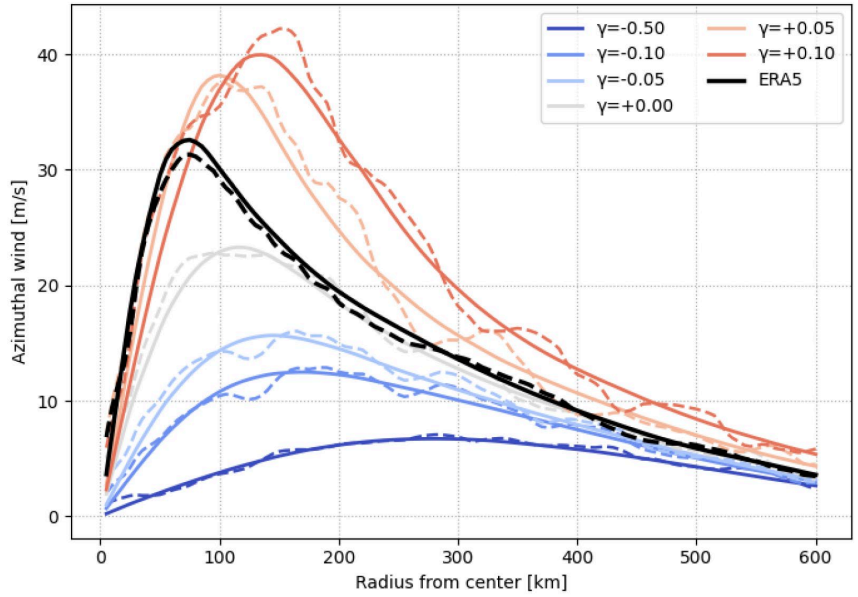
1. Hydrostatic balance

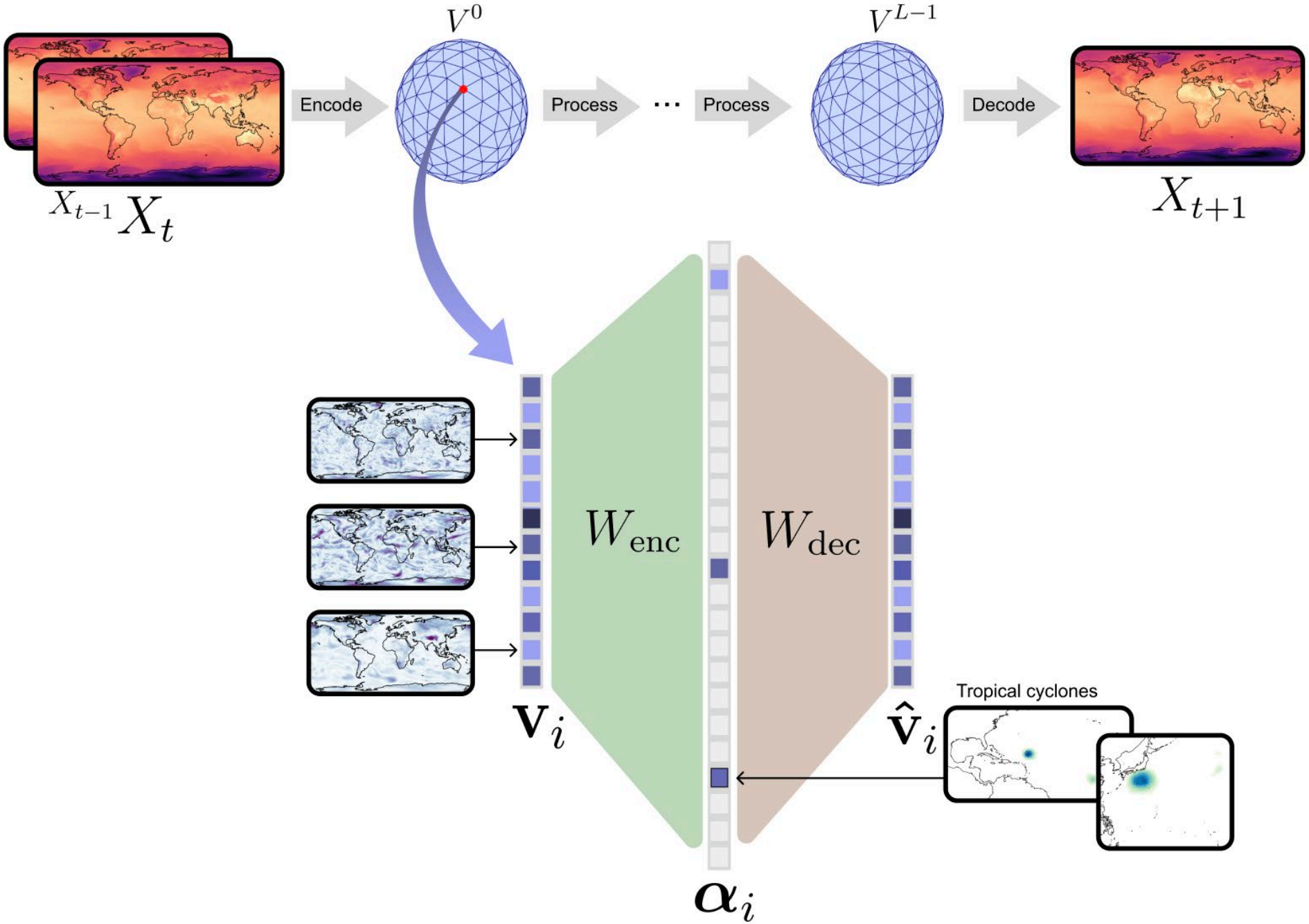
2. Mass continuity

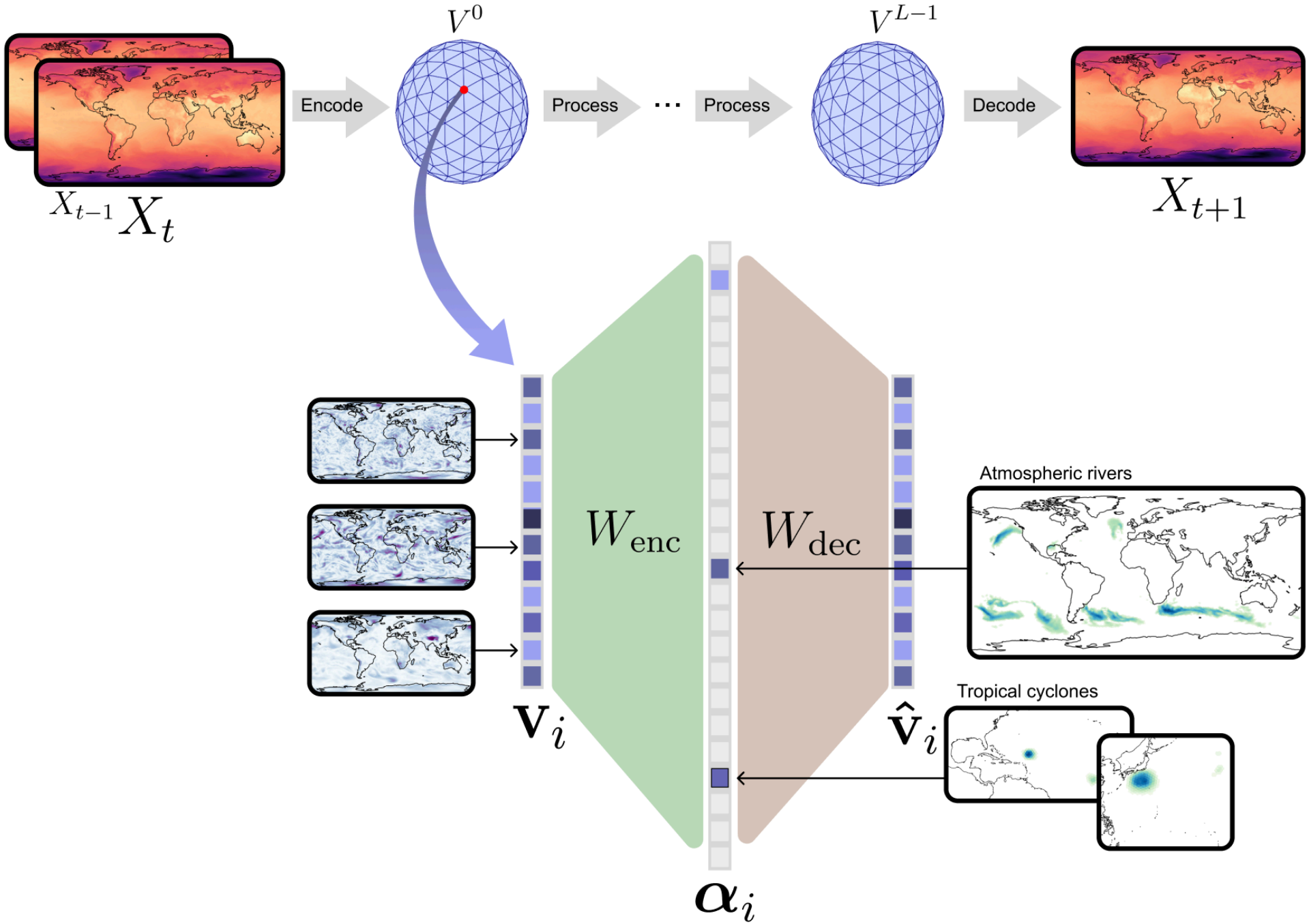
3. Gradient wind balance



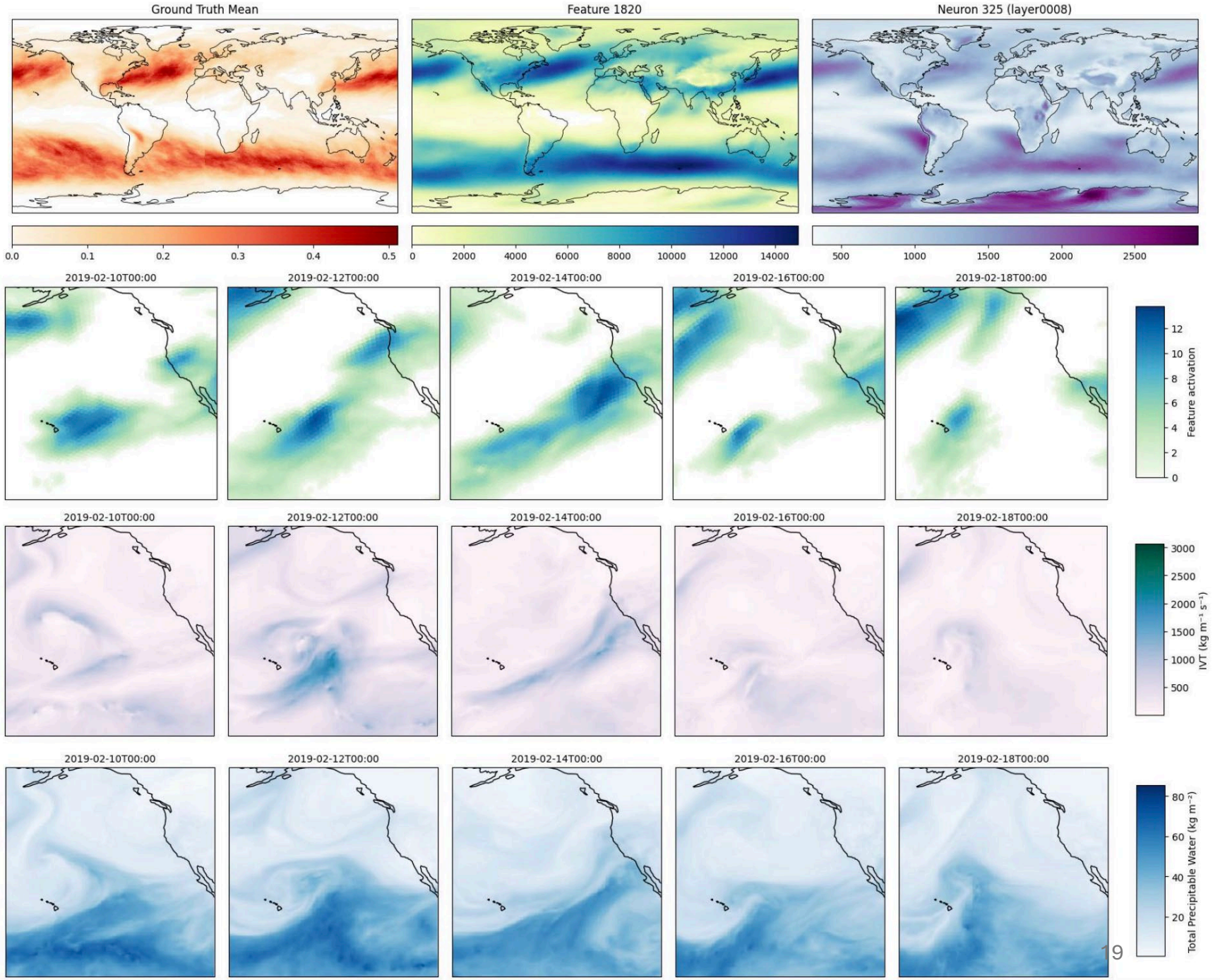


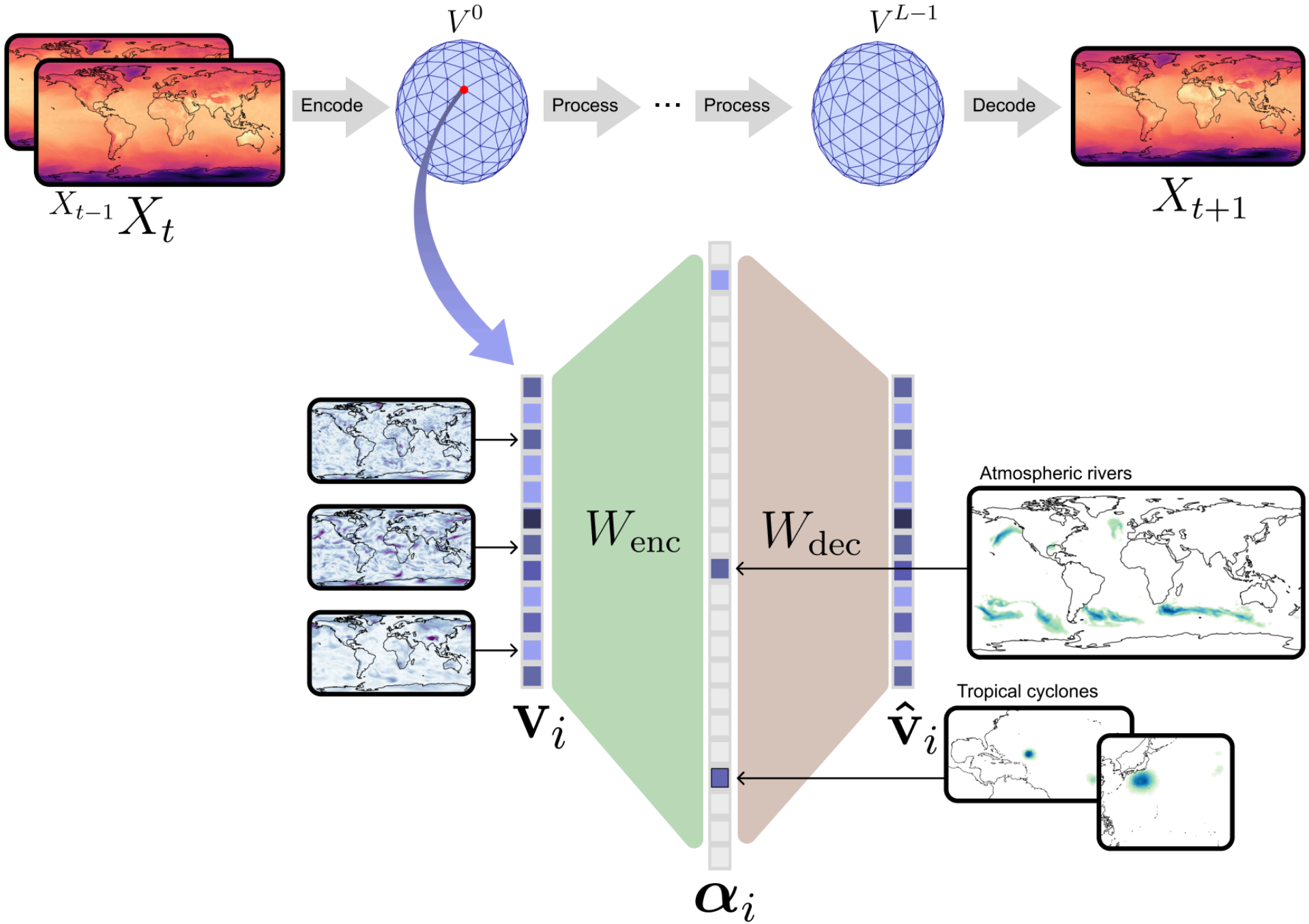


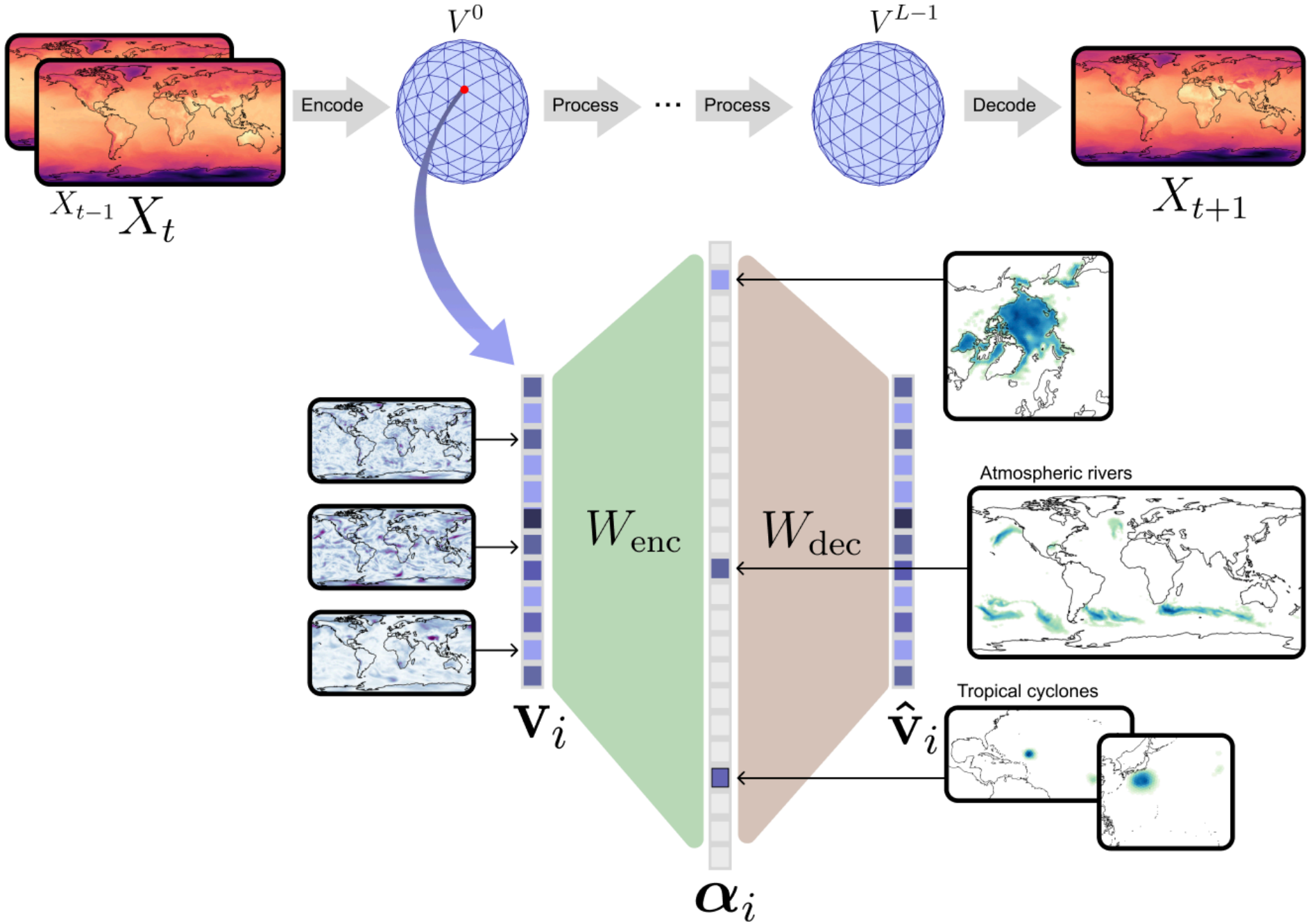

V^0
V^{L-1}
Encode
Process
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$
Tropical cyclones


V^0
V^{L-1}
Encode
Process
···
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$
Atmospheric rivers
Tropical cyclones

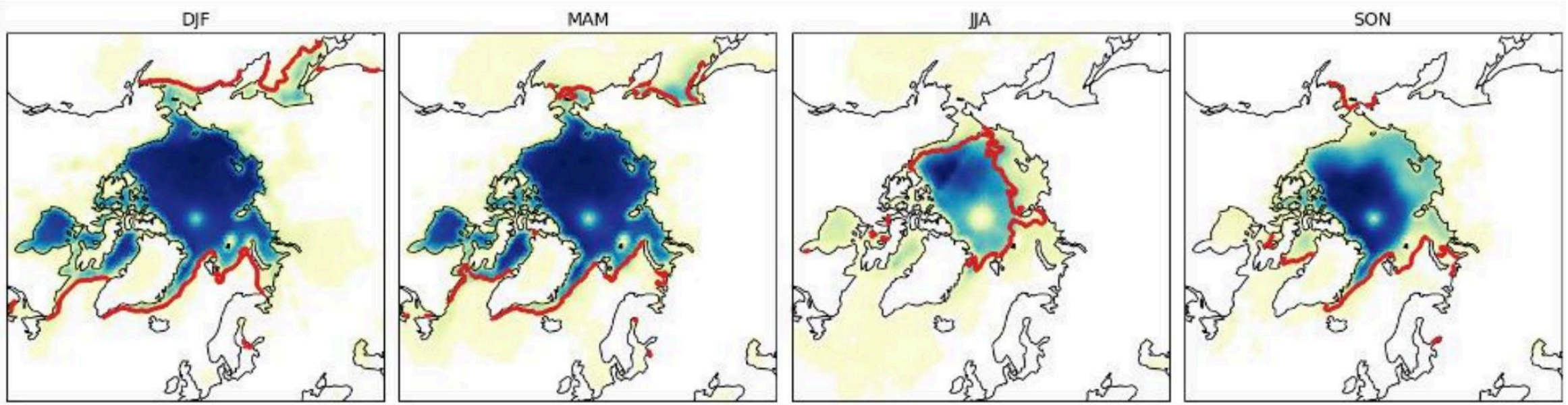
Feature 1820 activates in the presence of atmospheric rivers



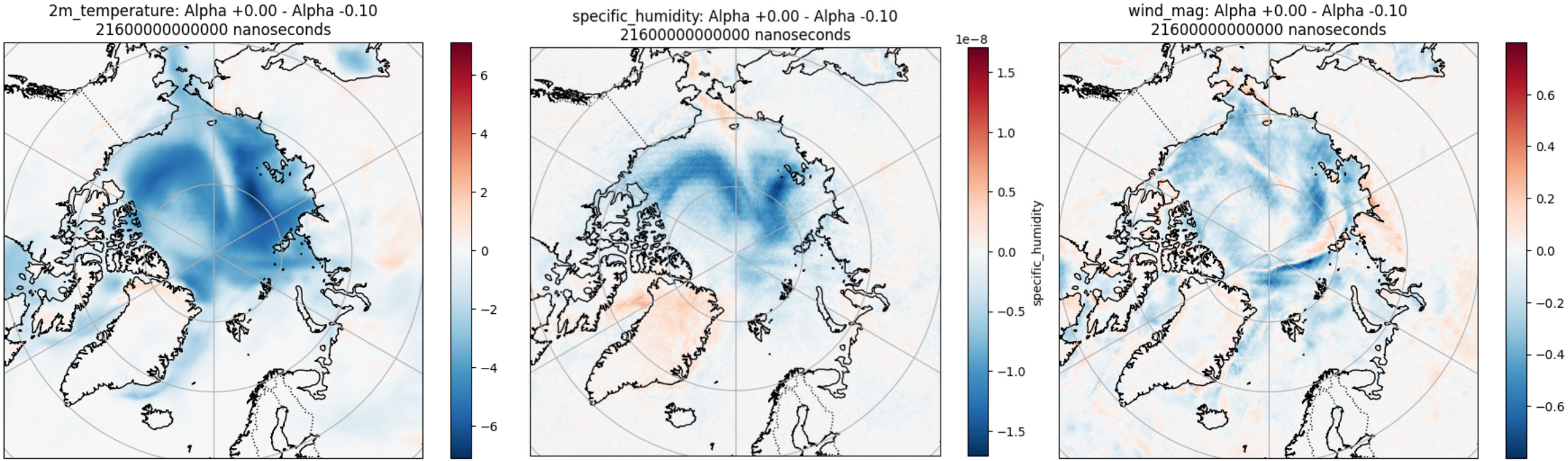

V^0
V^{L-1}
Encode
Process
···
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$
Atmospheric rivers
Tropical cyclones


V^0
V^{L-1}
Encode
Process
Process
Decode
X_{t-1} X_t
X_{t+1}
W_{enc}
W_{dec}
$\mathbf{v}_i$
$\hat{\mathbf{v}}_i$
$\boldsymbol{\alpha}_i$
Atmospheric rivers
Tropical cyclones

Feature 1710 activates in the presence of sea ice

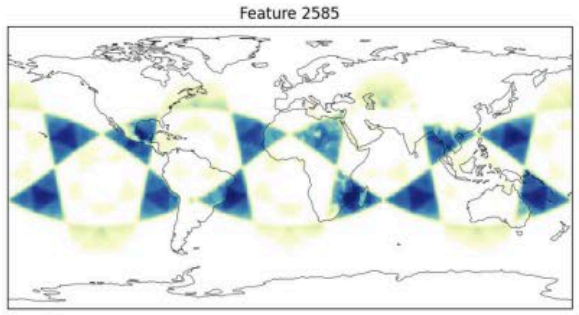


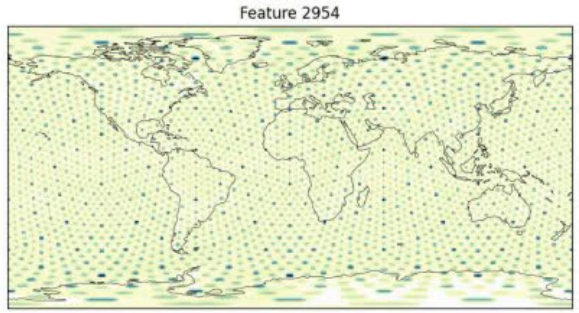
Experiment: what happens when we decrease the strength of feature 1710?

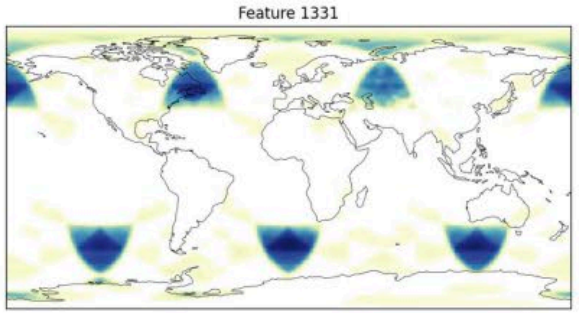


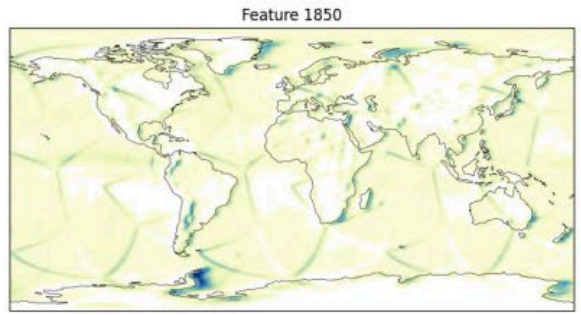
Temperature, specific humidity, and wind magnitude all **increase**

Other features are clearly not physical…









What about *dynamics*?

Article

A foundation model for the Earth system

https://doi.org/10.1038/s41586-025-09005-y

Received: 28 May 2024

Accepted: 9 April 2025

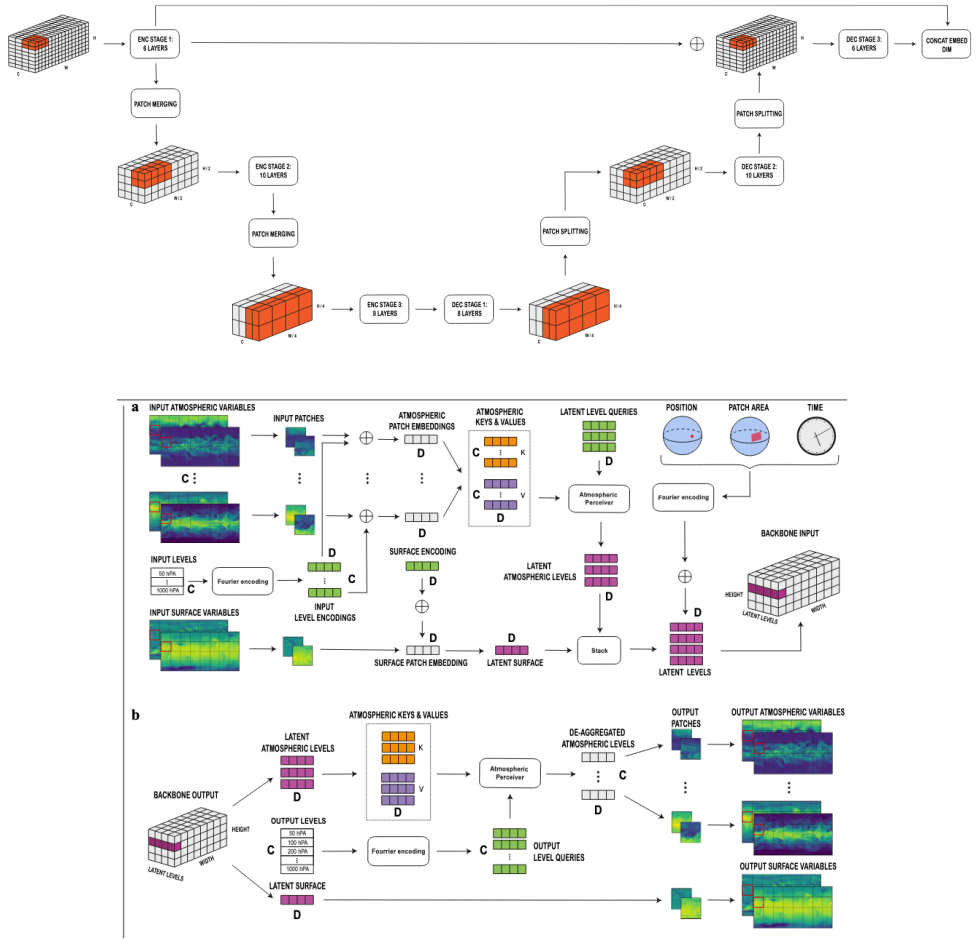
Published online: 21 May 2025

Open access

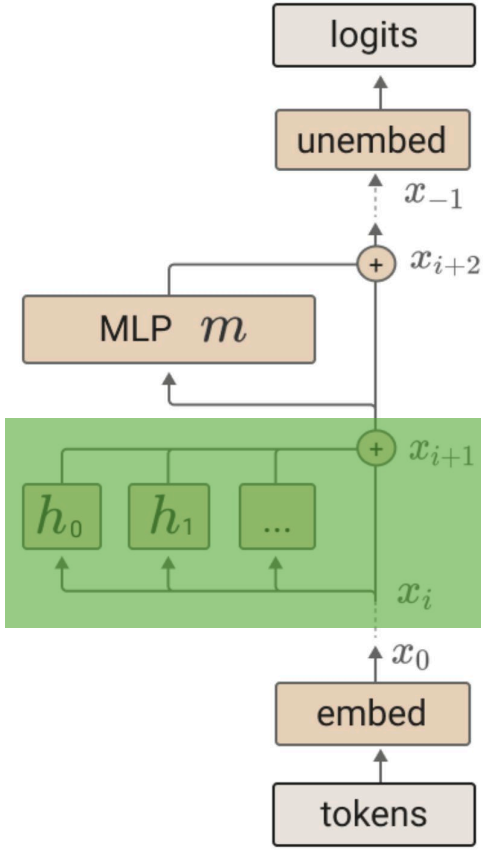
Check for updates

Cristian Bodnar[1,2,11], Wessel P. Bruinsma[1,11], Ana Lucic[1,3,11], Megan Stanley[1,11], Anna Allen[4,11], Johannes Brandstetter[1,5], Patrick Garvan[1], Maik Riechert[1], Jonathan A. Weyn[6], Haiyu Dong[6], Jayesh K. Gupta[2,7], Kit Thambiratnam[6], Alexander T. Archibald[4], Chun-Chieh Wu[8], Elizabeth Heider[1], Max Welling[1,3], Richard E. Turner[1,4,9] & Paris Perdikaris[1,10]✉

Reliable forecasting of the Earth system is essential for mitigating natural disasters and supporting human progress. Traditional numerical models, although powerful, are extremely computationally expensive[1]. Recent advances in artificial intelligence (AI) have shown promise in improving both predictive performance and efficiency[2,3], yet their potential remains underexplored in many Earth system domains. Here we introduce Aurora, a large-scale foundation model trained on more than one million hours of diverse geophysical data. Aurora outperforms operational forecasts in predicting air quality, ocean waves, tropical cyclone tracks and high-resolution weather, all at orders of magnitude lower computational cost. With the ability to be fine-tuned for diverse applications at modest expense, Aurora represents a notable step towards democratizing accurate and efficient Earth system predictions. These results highlight the transformative potential of AI in environmental forecasting and pave the way for broader accessibility to high-quality climate and weather information.



Weather transformer



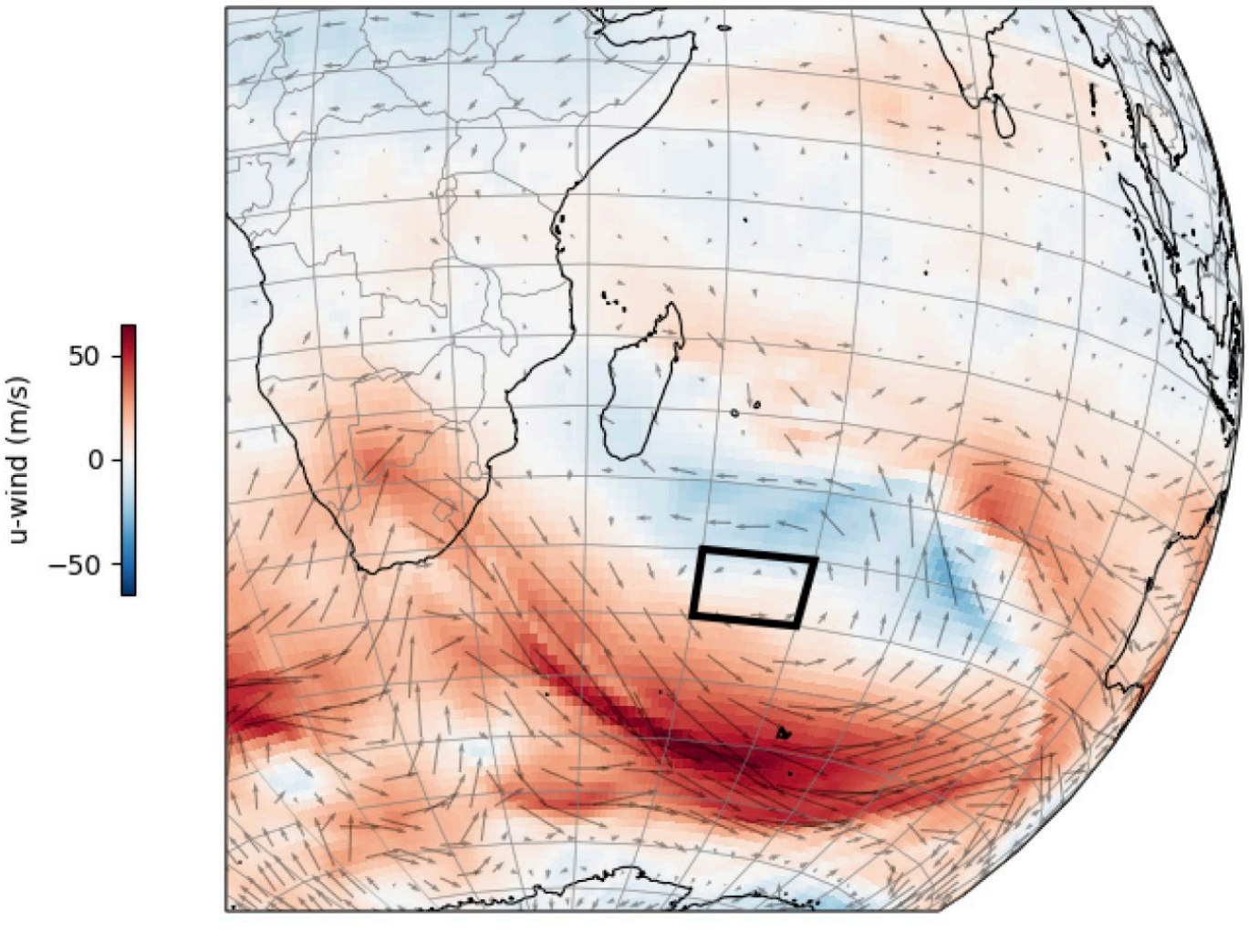
Attention

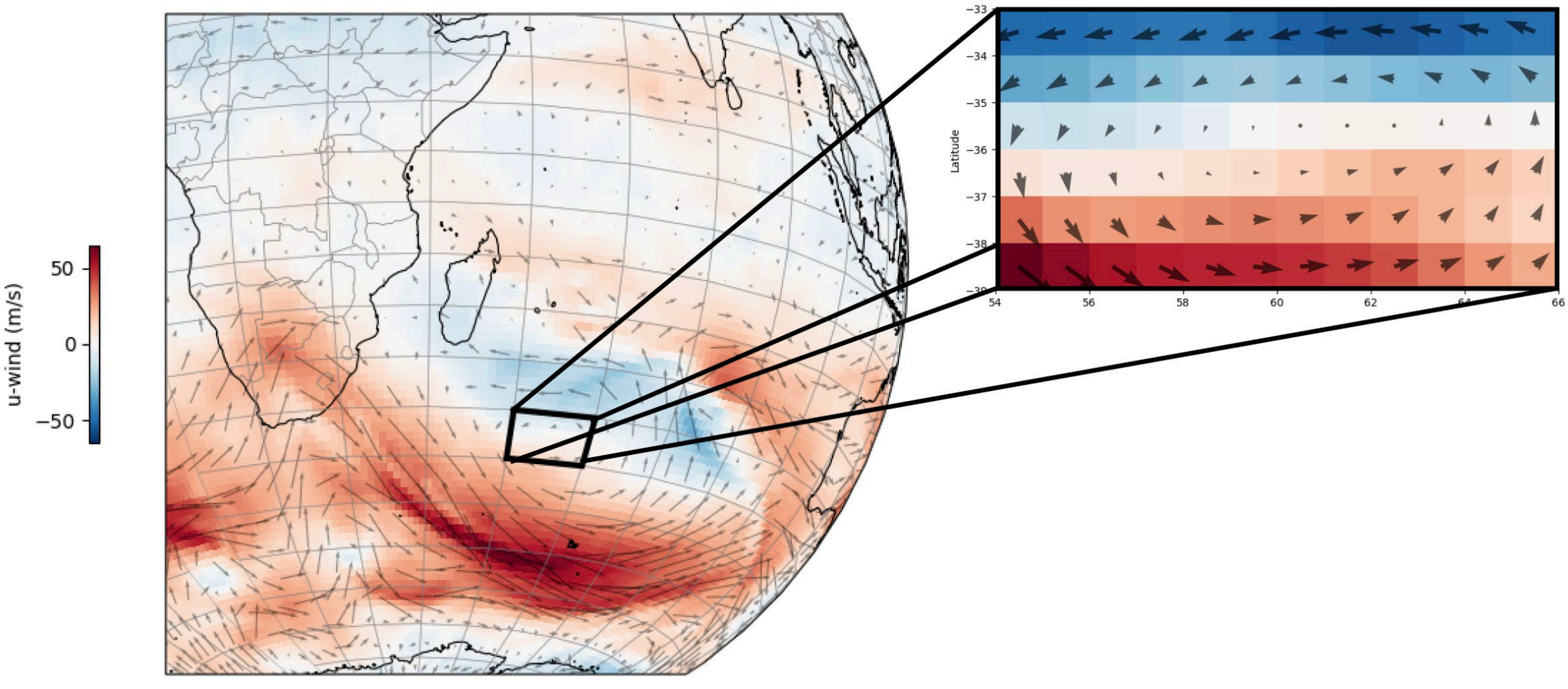
$$a_{ij} = \text{softmax}_j\left((x_i^T \mathbf{W}_Q^T \mathbf{W}_K x_j)/\sqrt{d_{\text{head}}}\right)$$

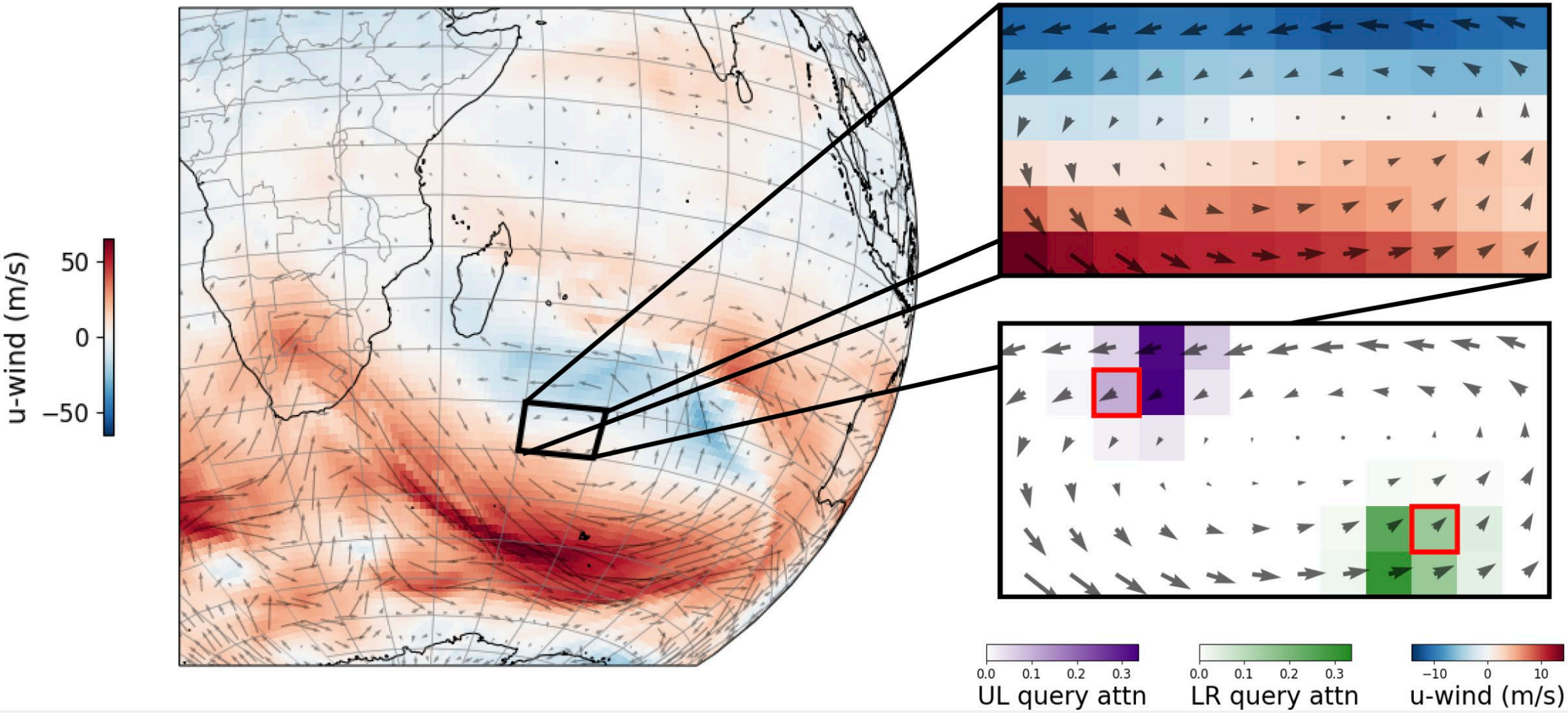
QK: which patches talks to patch i?

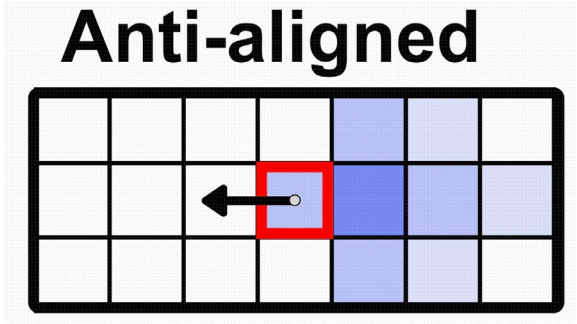
$$h_i = \sum_j a_{ij}(\mathbf{W}_O \mathbf{W}_V x_j)$$

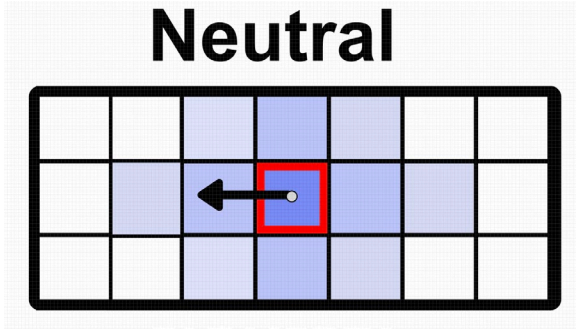
OV: what do they write to patch i?

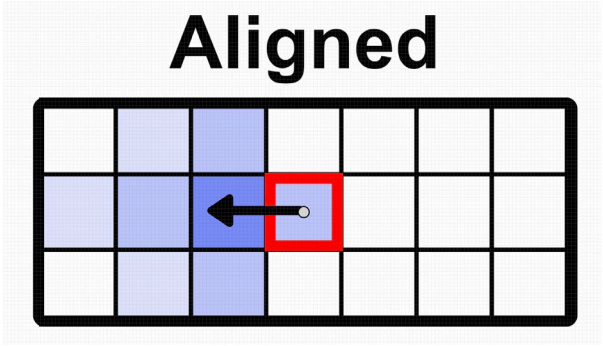

u-wind (m/s)
50
0
−50

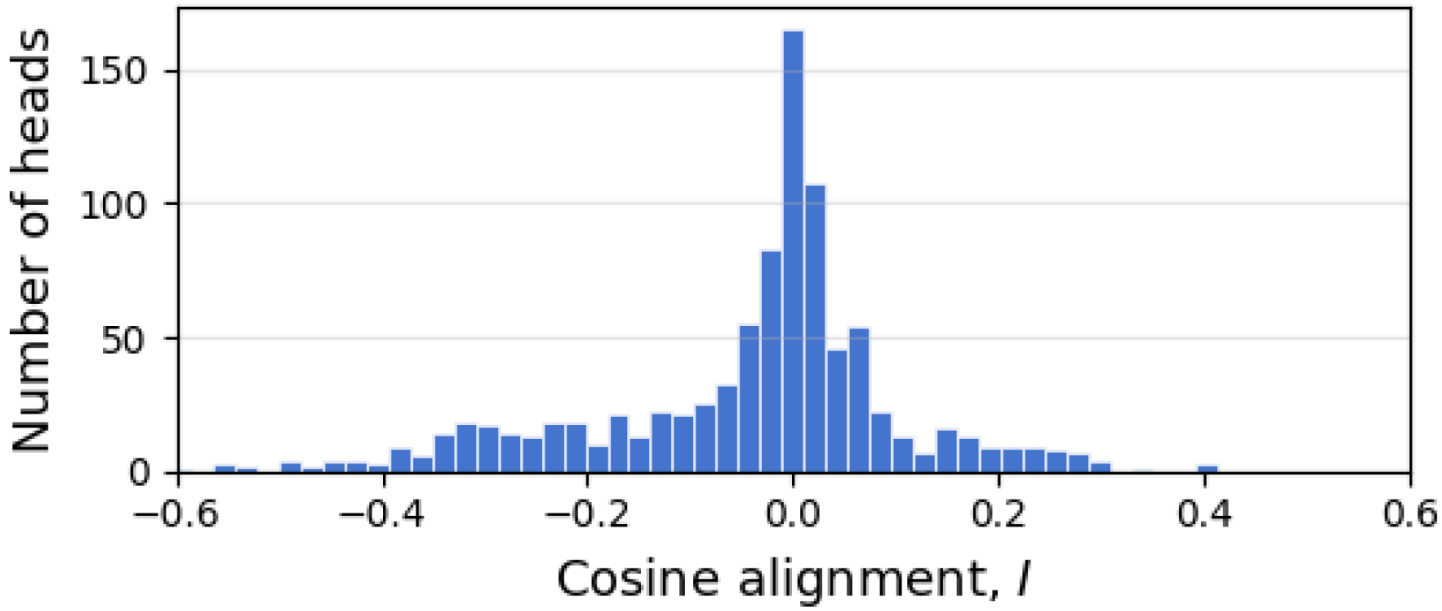

u-wind (m/s)
50
0
−50
Latitude
−33
−34
−35
−36
−37
−38
54
56
58
60
62
66


u-wind (m/s)
50
0
−50
0.0 0.1 0.2 0.3
UL query attn
0.0 0.1 0.2 0.3
LR query attn
−10 0 10
u-wind (m/s)





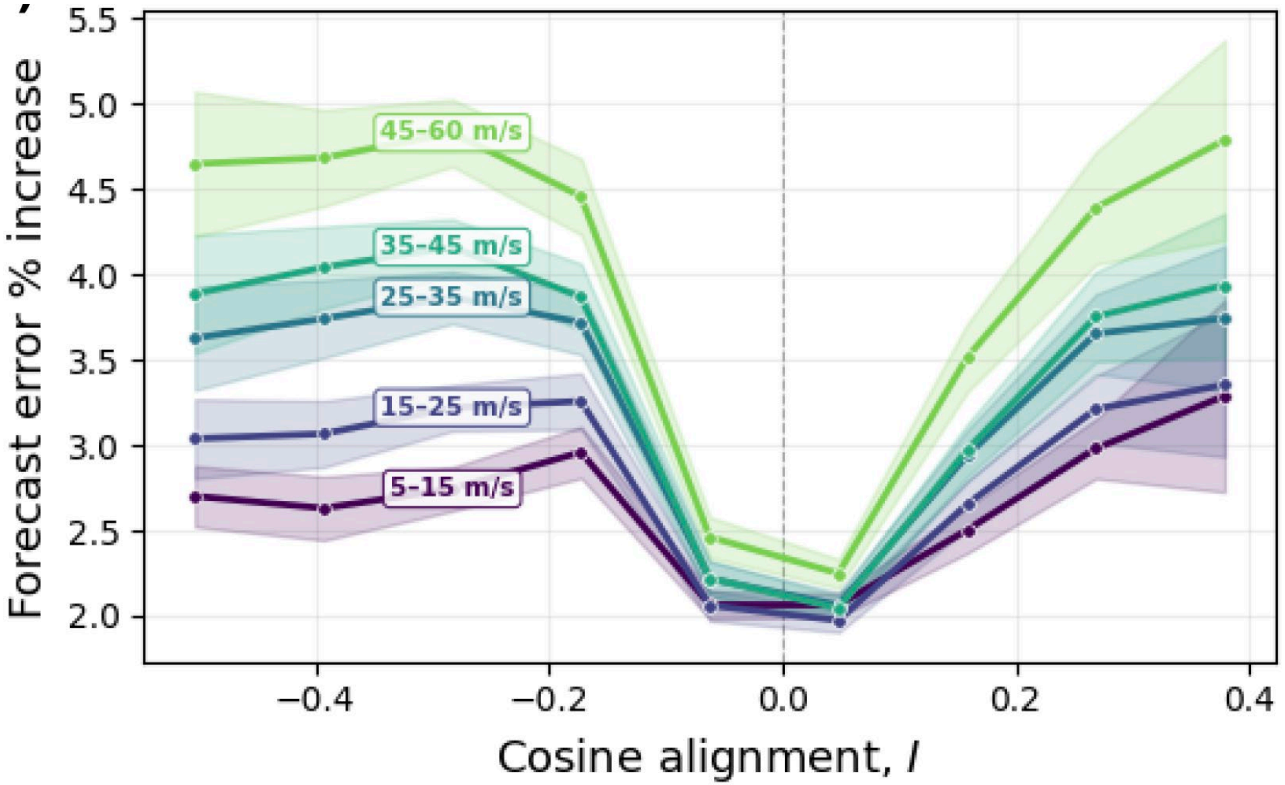




Significant fraction of attention heads are aligned with the velocity field

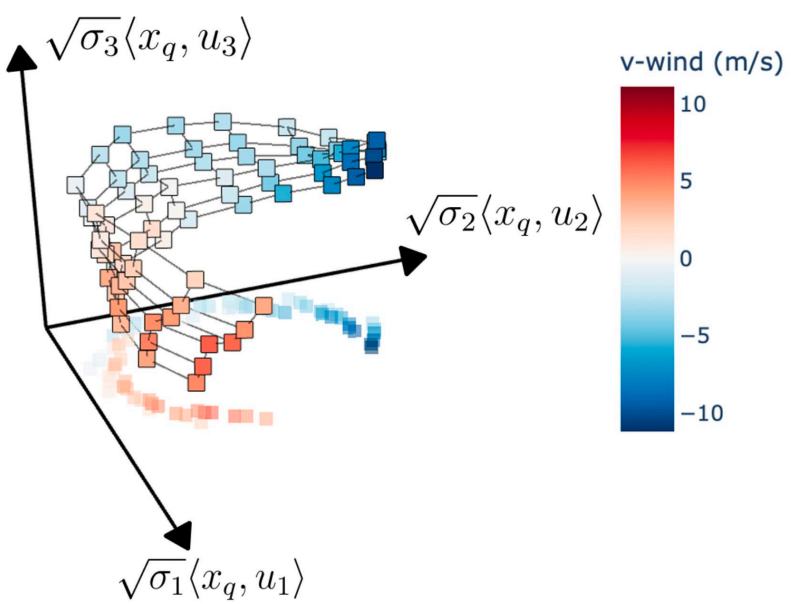
Do they “do” advection?

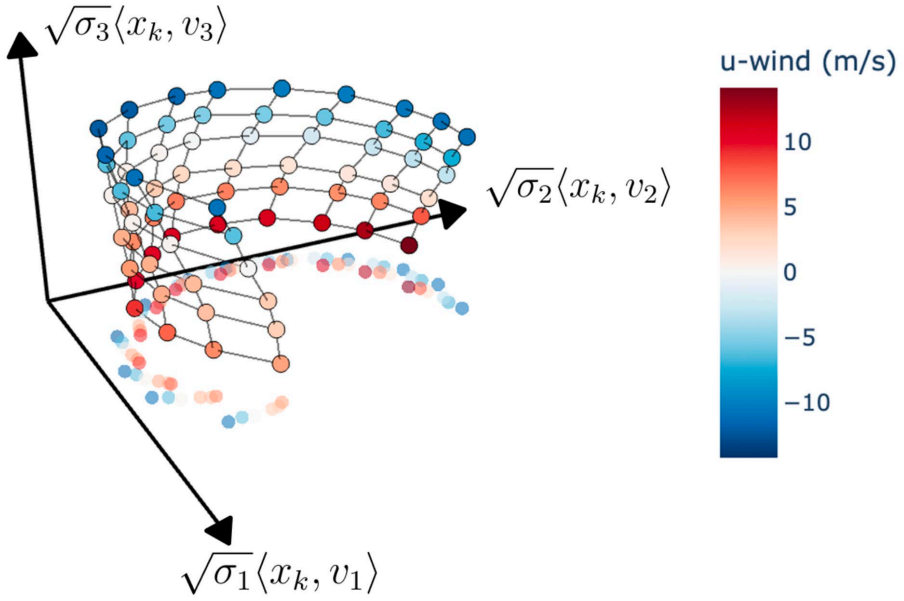
If we remove the heads from the forward pass, can the model still advect?



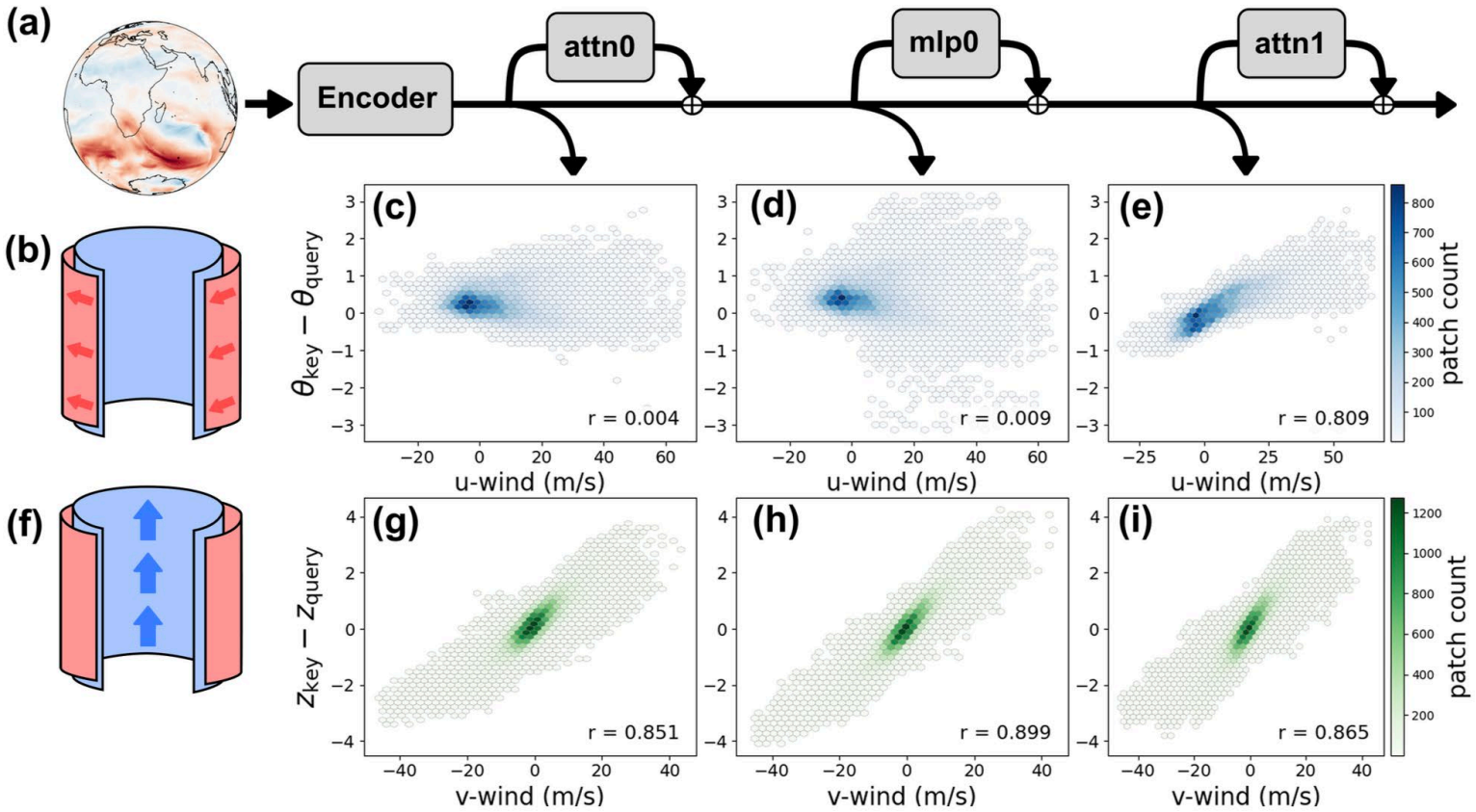
Model struggles in high-wind regions without them

How does the model align its attention?





We can locate where these twists/stretches happen



Conclusion

Mostly toy answers to complicated problems

Circuit tracing is a natural next step

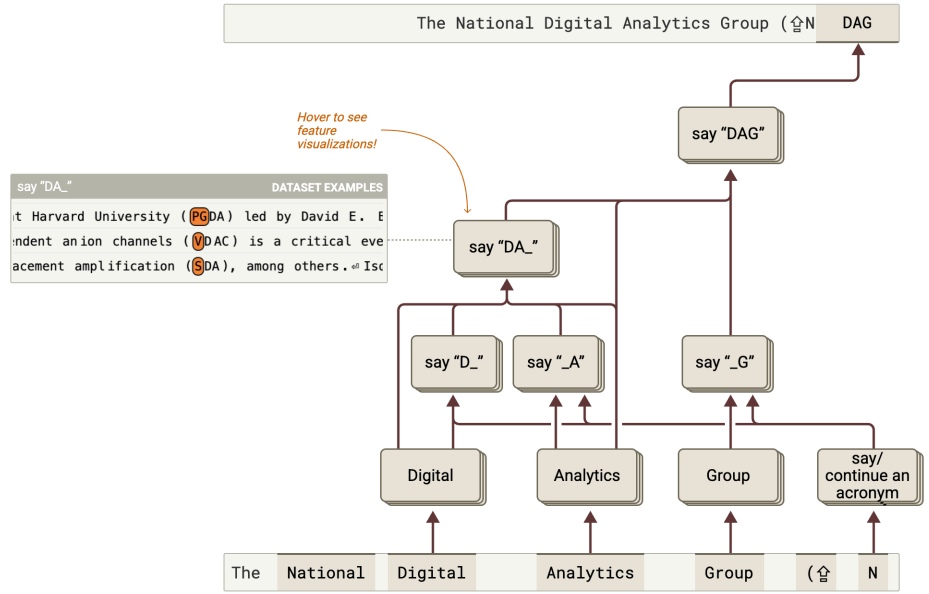


Figure 5: A simplified diagram of the attribution graph for 18L completing a fictional acronym.

Read more:

T. MacMillan and N.T. Ouellette. Towards mechanistic understanding in a data-driven weather model: internal activations reveal interpretable physical features. 12 2025. URL http://arxiv.org/abs/2512. 24440

T. MacMillan, E. Huseby, S. Takatori, and N.T. Ouellette. Advection Heads in an Atmosphere Foundation Model. PAI 2026.

Reach out! theo@getcausal.ai We're hiring :)