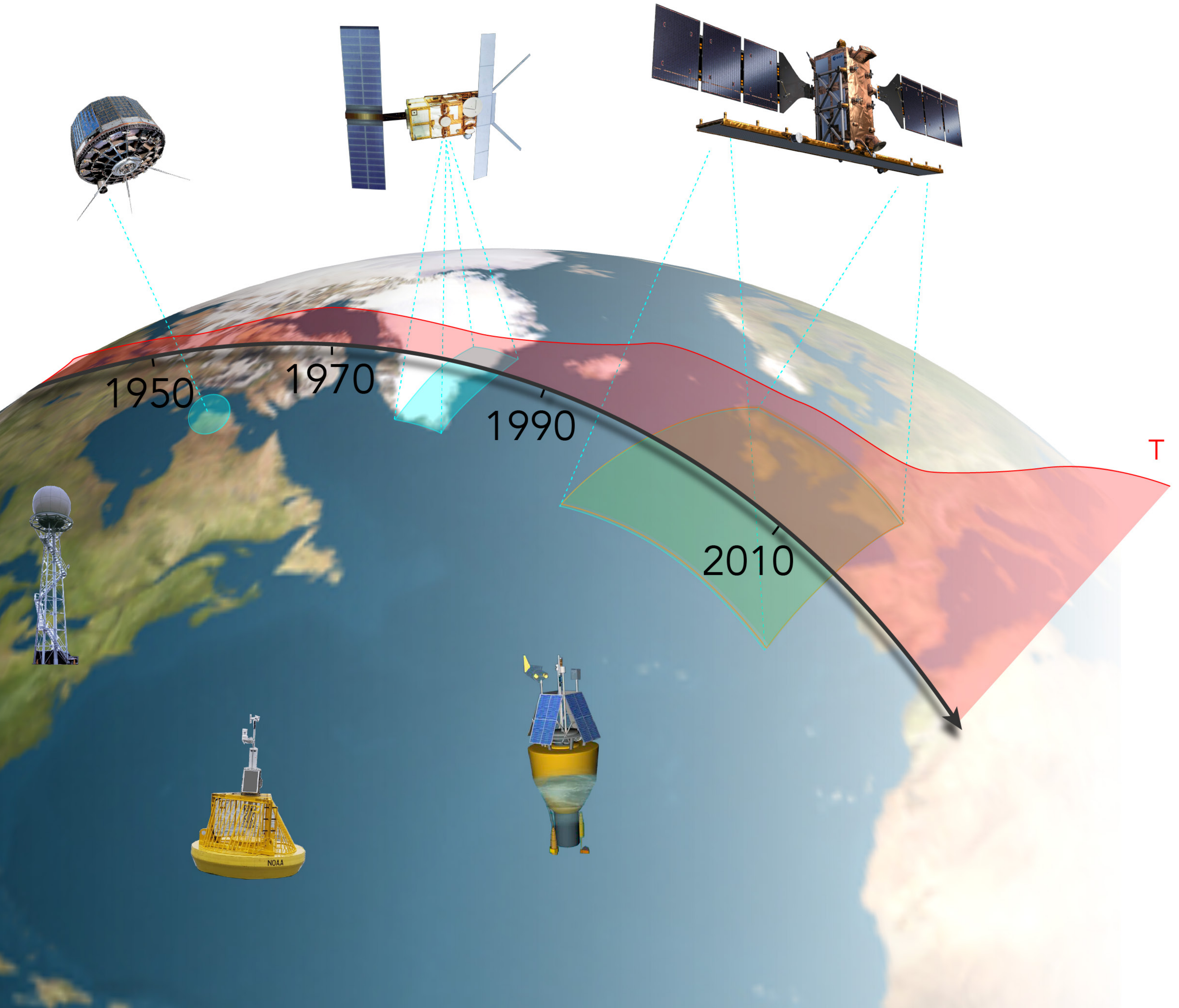
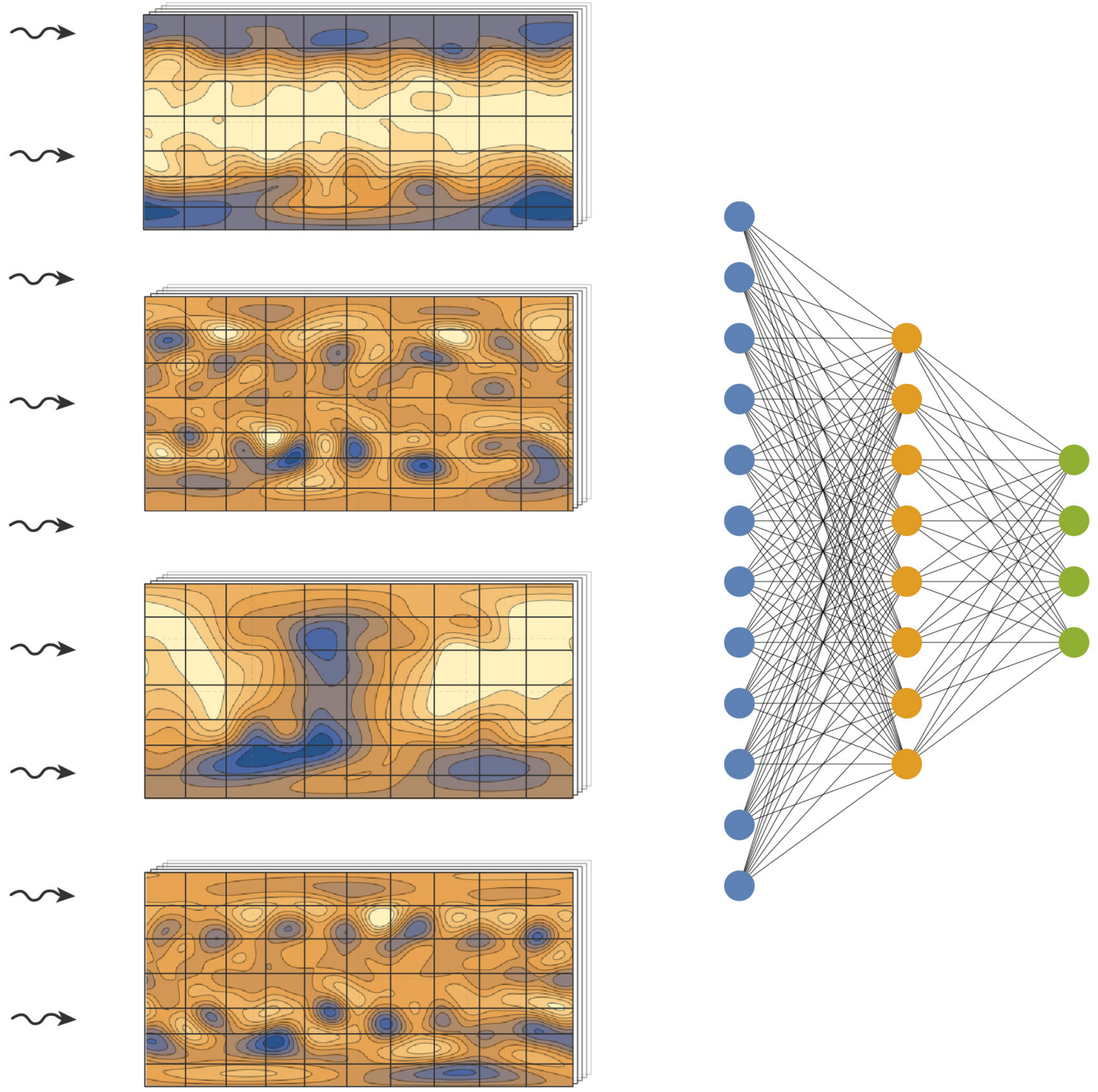
Representation learning for the Earth Sciences

Christian Lessig, Otto-von-Guericke-Universität Magdeburg





scientific insight

improved applications

Motivation

- Large amounts of data available in the Earth sciences:
  - ERA5: ≈6 PB
  - CMIP6: ≈100 PB
  - MetOp-SG: 8 x 864 GB/day (80 Mbit/s)
  - OCEAN5: ≈4 PB

Motivation

- Large amounts of data available in the Earth sciences:
  - ERA5: ≈6 PB
  - CMIP6: ≈100 PB
  - MetOp-SG: 8 x 864 GB/day (80 Mbit/s)
  - OCEAN5: ≈4 PB

growing fast

Motivation

- Large amounts of data available in the Earth sciences:
  - ERA5: ≈6 PB
  - CMIP6: ≈100 PB
  - MetOp-SG: 8 x 864 GB/day (80 Mbit/s)
  - OCEAN5: ≈4 PB

  growing fast

- Observational or quasi-observational data with effects and phenomena not captured in, e.g., analytic models

Motivation

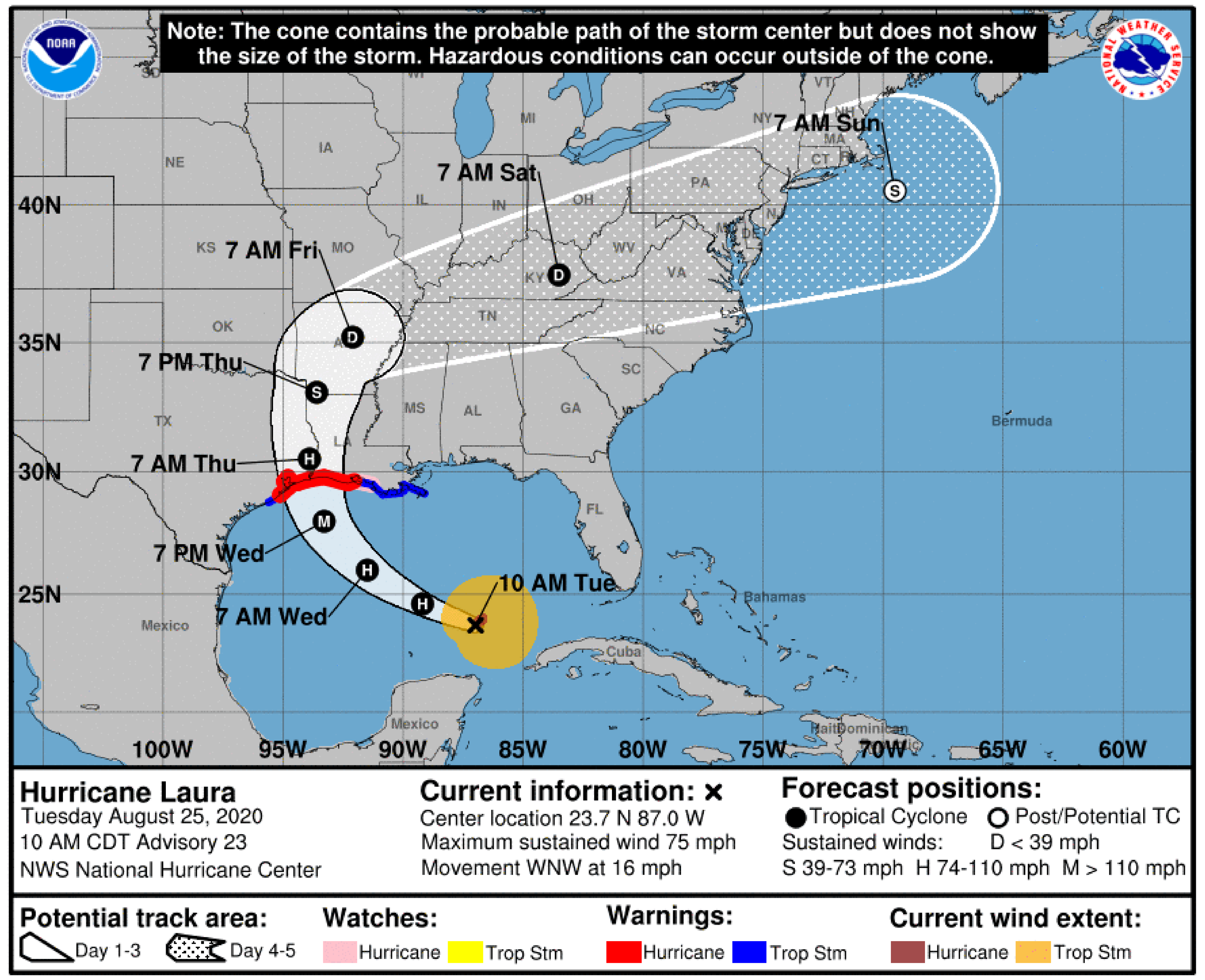
- How to use this data for machine learning in the Earth sciences?

Motivation

- How to use this data for machine learning in the Earth sciences?
  - Most data is unlabeled
  - Super-computing infrastructure required for storing and processing
  - Unclear how to ensure that learned models are physically consistent

Motivation

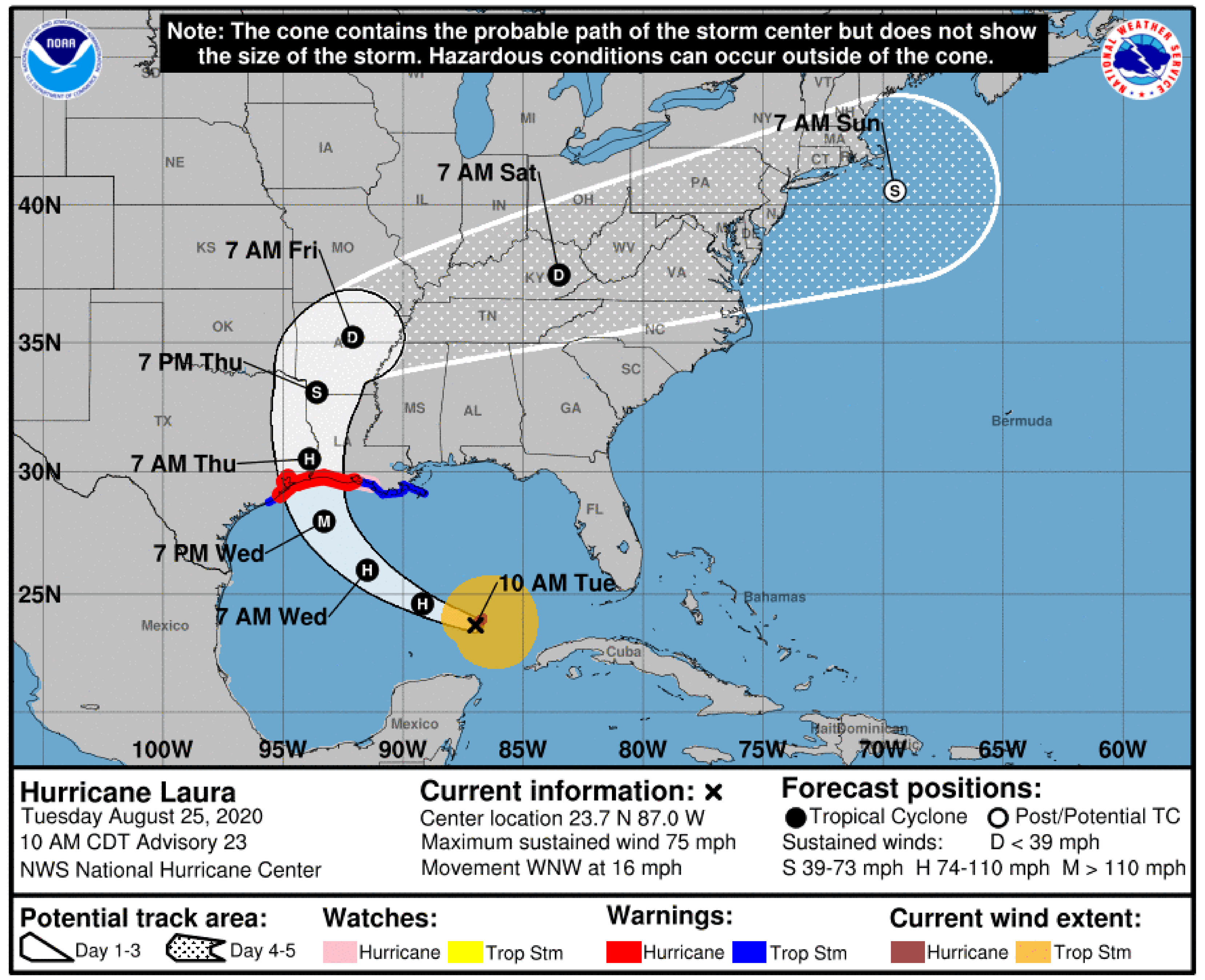
- Example: hurricane tracking
  - Great importance for risk assessment and climate projections



https://www.nhc.noaa.gov/aboutnhcgraphics.shtml

Motivation

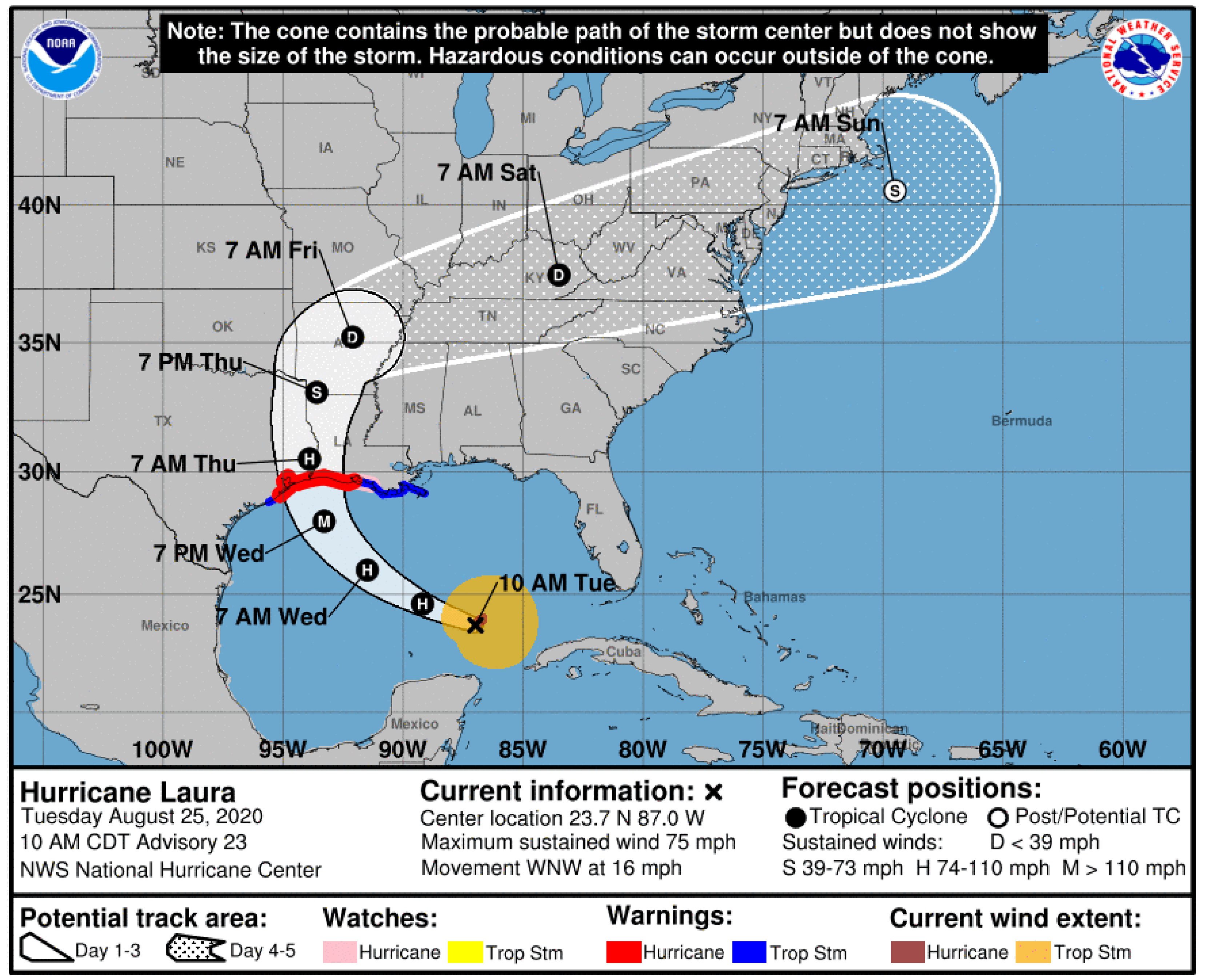
- Example: hurricane tracking
  - Large importance for immediate effects and climate projections
  - NOAA HURDAT2 Atlantic hurricane database: 6.5 MB



https://www.nhc.noaa.gov/aboutnhcgraphics.shtml

Motivation

- Example: hurricane tracking
  - Large importance for immediate effects and climate projections
  - NOAA HURDAT2 Atlantic hurricane database: 6.5 MB



https://www.nhc.noaa.gov/aboutnhcgraphics.shtml

$\Rightarrow$ Can we use unlabeled ERA5 (or similar) to augment the very small amounts of labeled hurricane tracking data?

Motivation

- Similar situation for other applications:
  - Ozone and air pollution prediction
  - Observations with missing data
  - Classification of extreme events
  - Prediction of extreme events
  - …

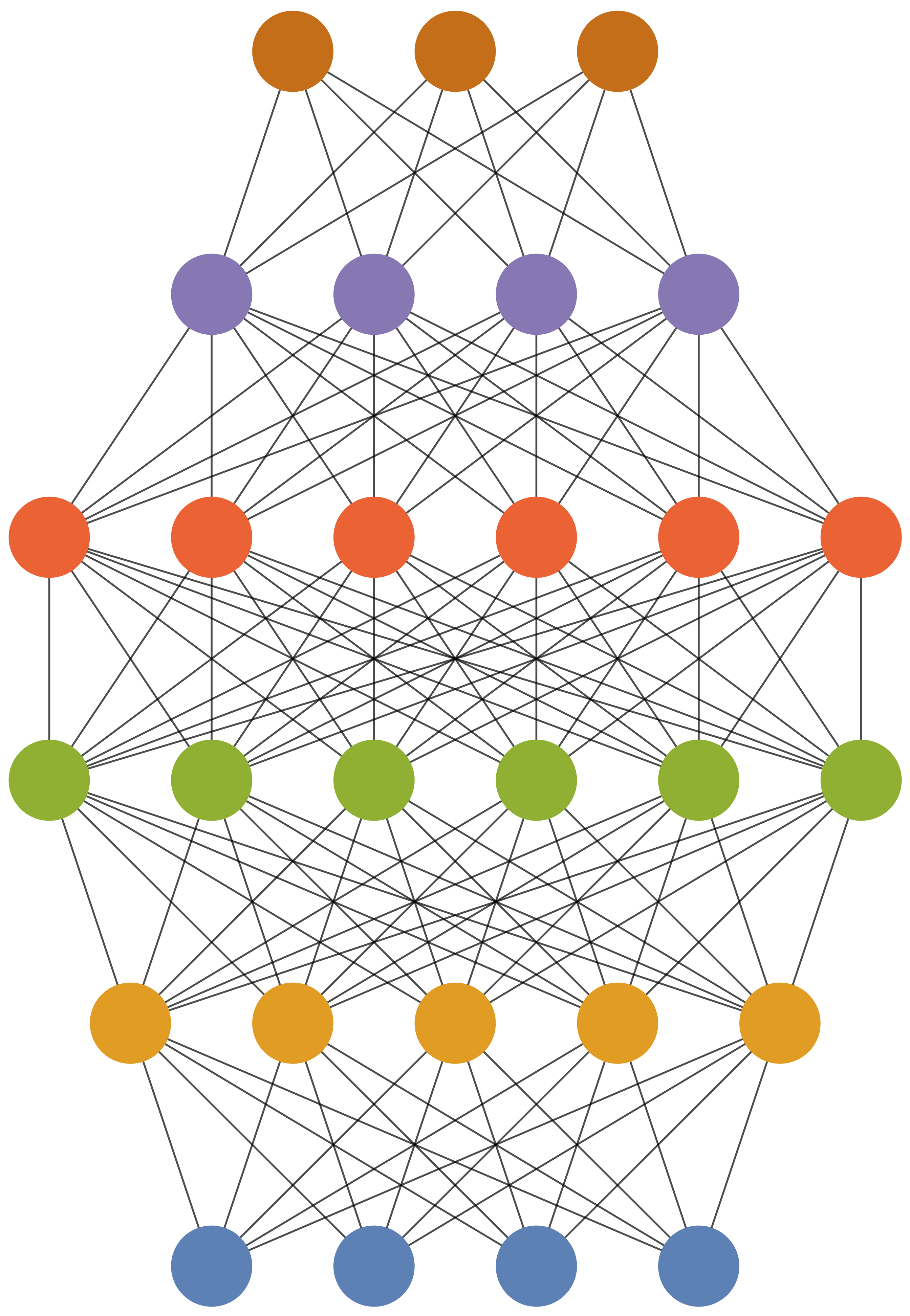
Self-supervised representation learning

- Yoshua Bengio: “Humans develop representations and abstractions to enable problem-solving and reasoning; our computers should do the same.”[1]

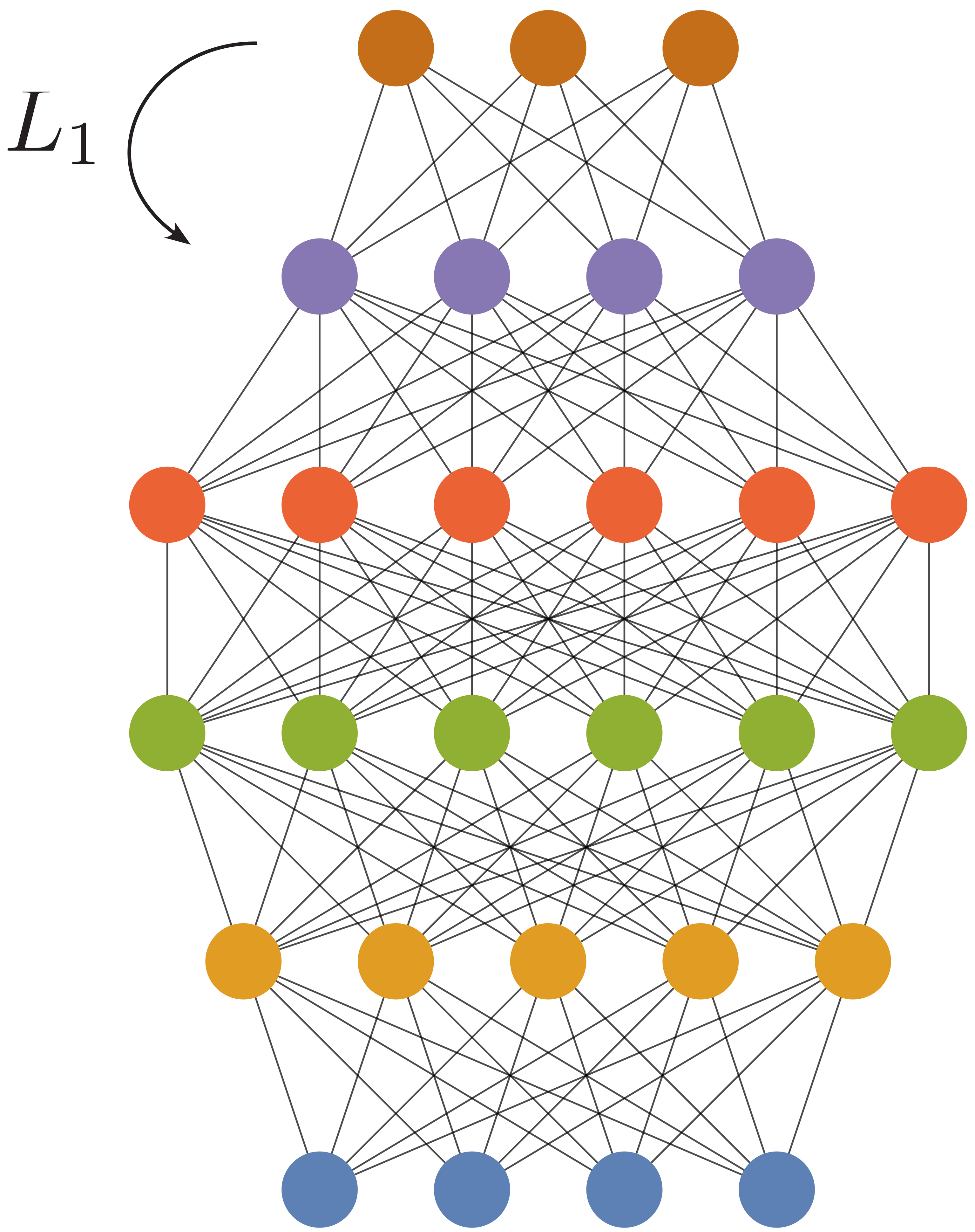
- Yann LeCun: “Self-supervised learning: The dark matter of intelligence“[2]

[1] http://www.iro.umontreal.ca/~bengioy/talks/icml2012-YB-tutorial.pdf
[2] https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

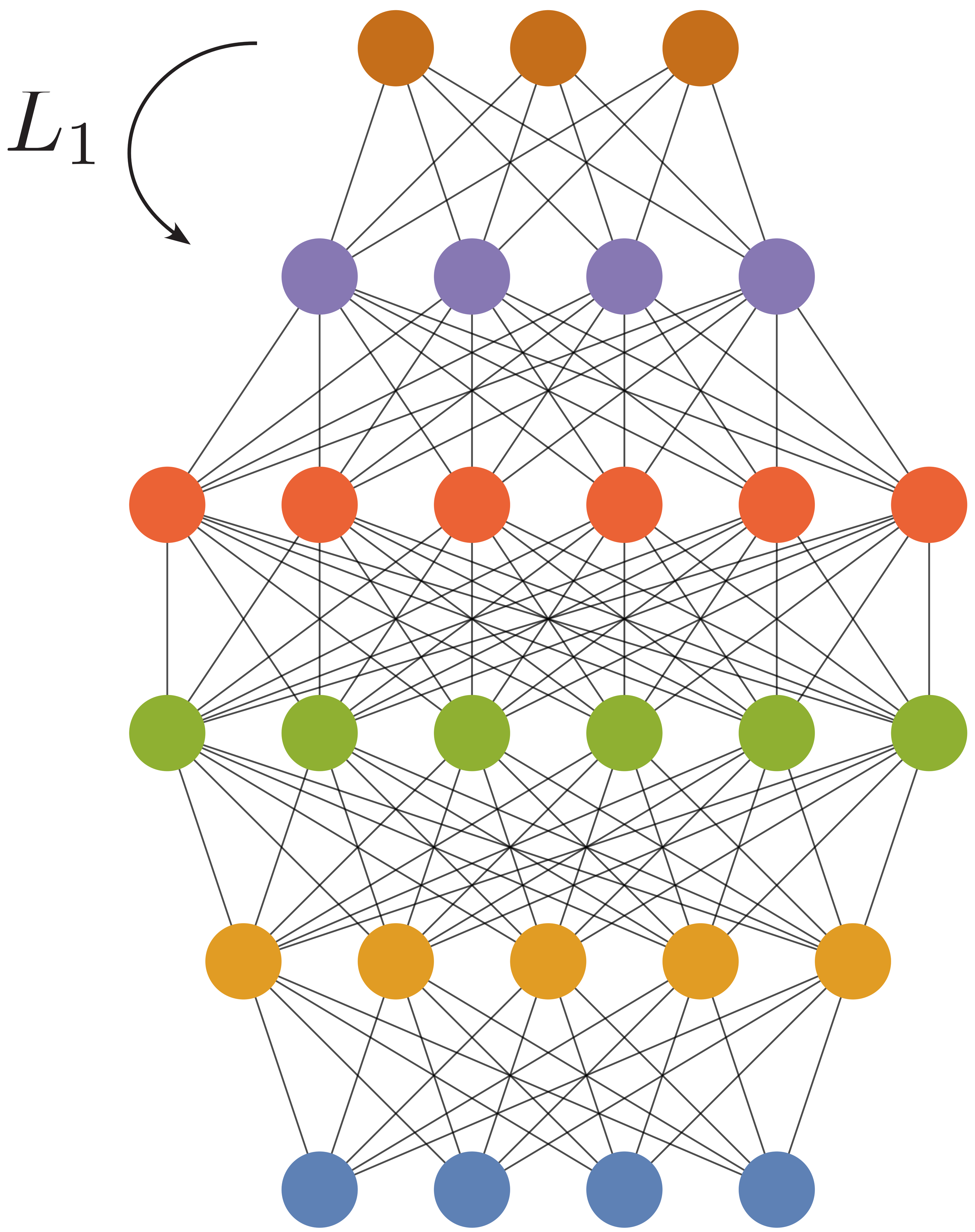
Self-supervised representation learning



Self-supervised representation learning

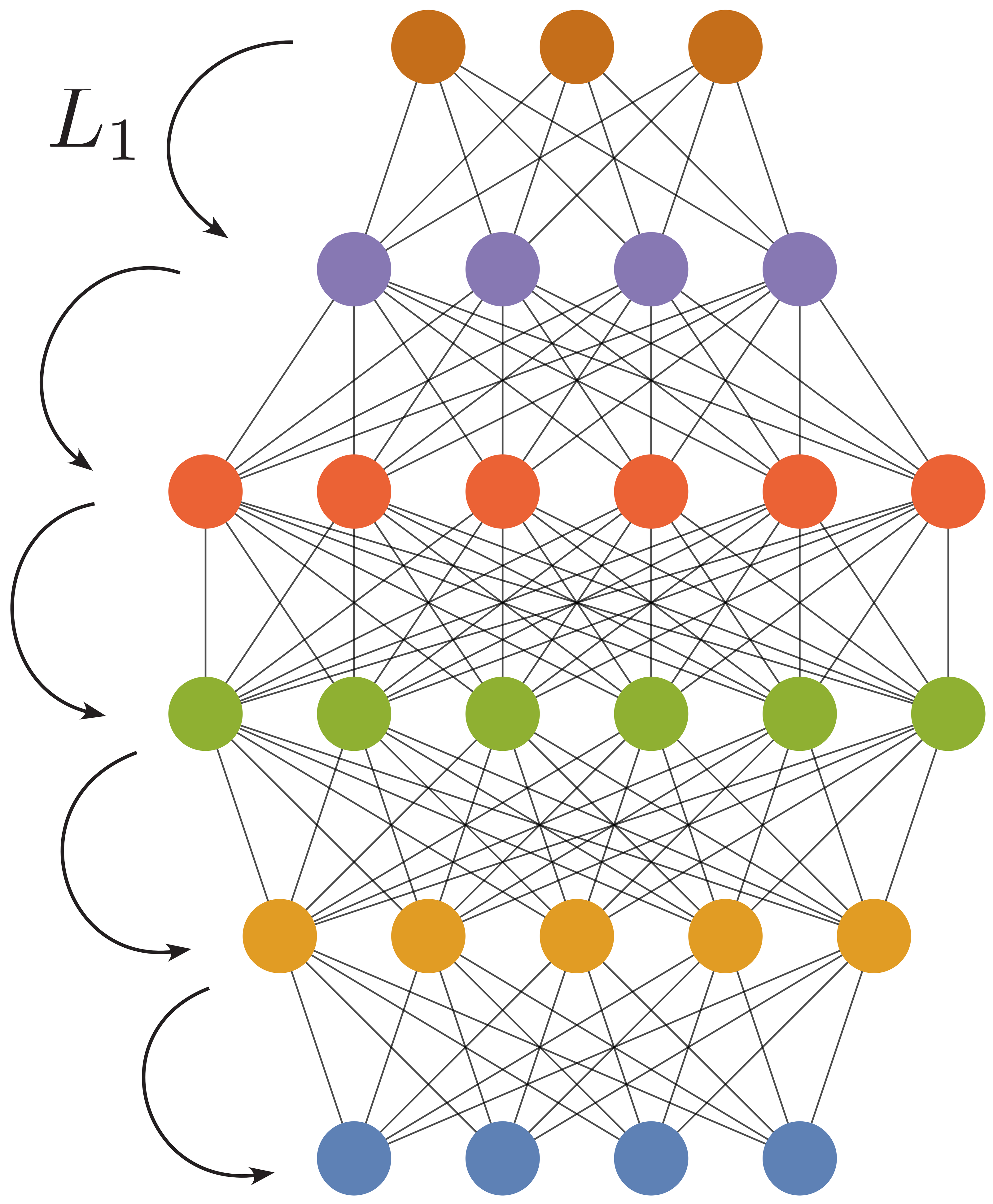


Self-supervised representation learning



$$L_1 : \mathbb{R}^n \to \mathbb{R}^m$$

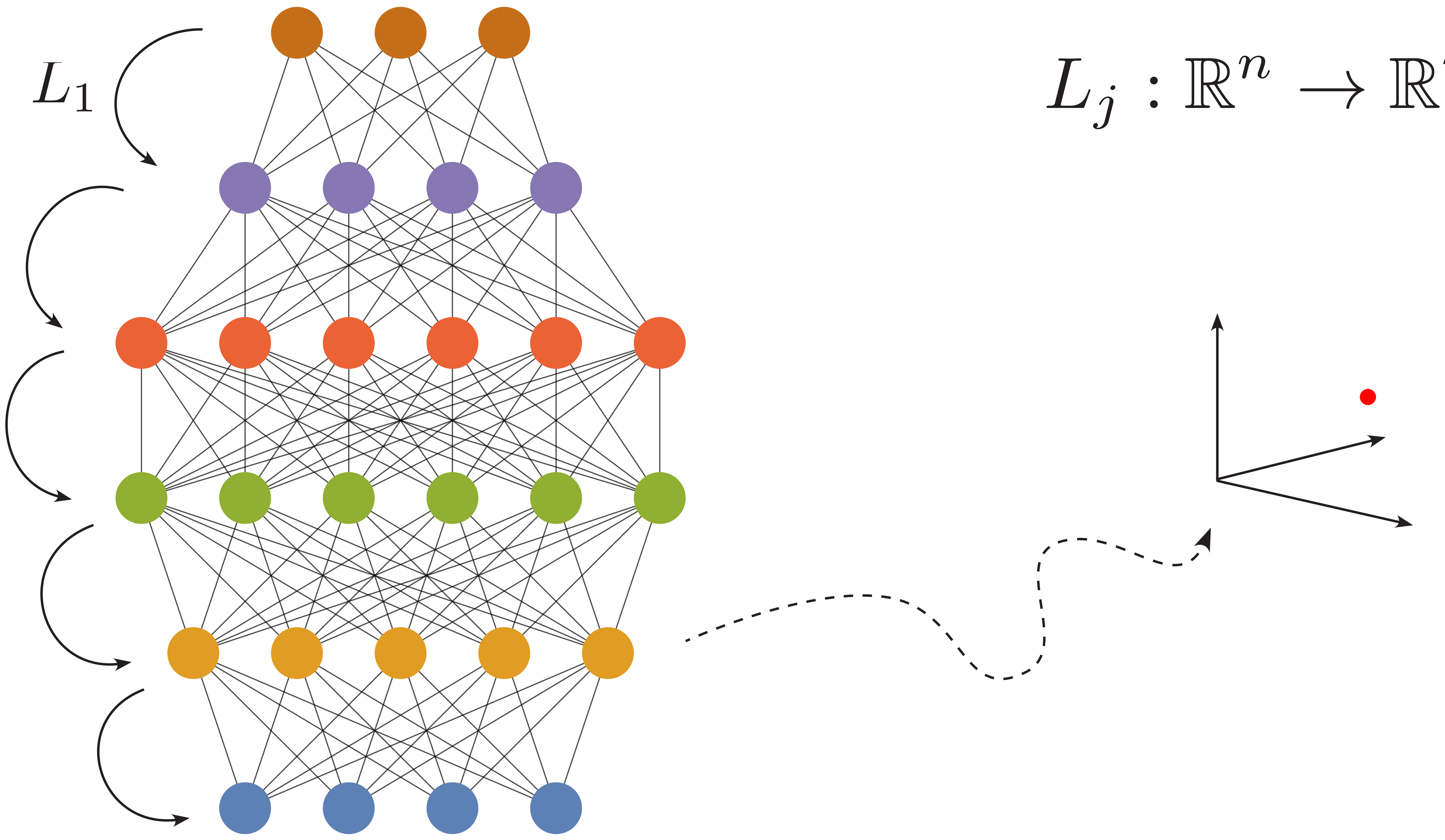
Self-supervised representation learning



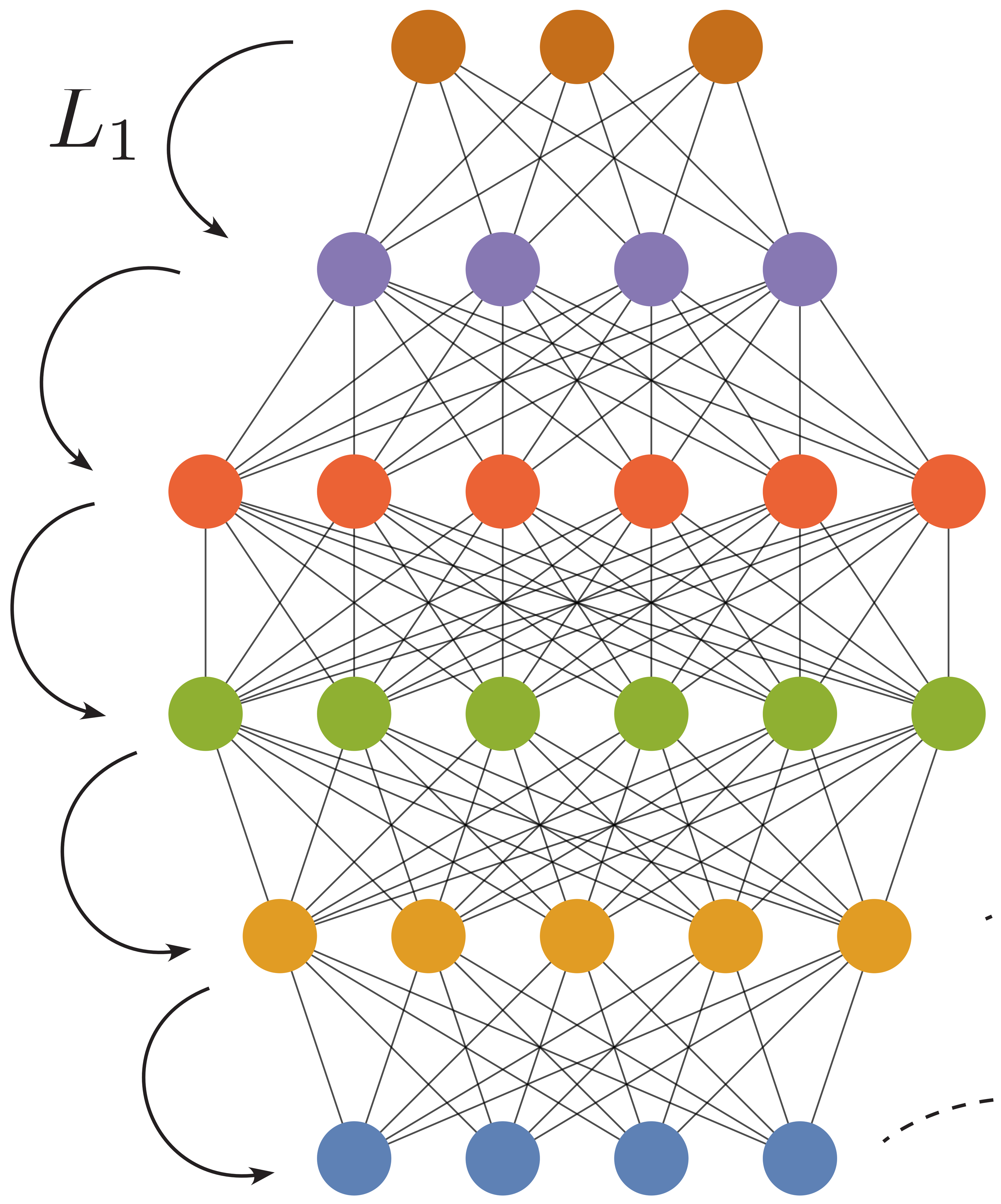
$$L_j : \mathbb{R}^n \to \mathbb{R}^m$$

Self-supervised representation learning

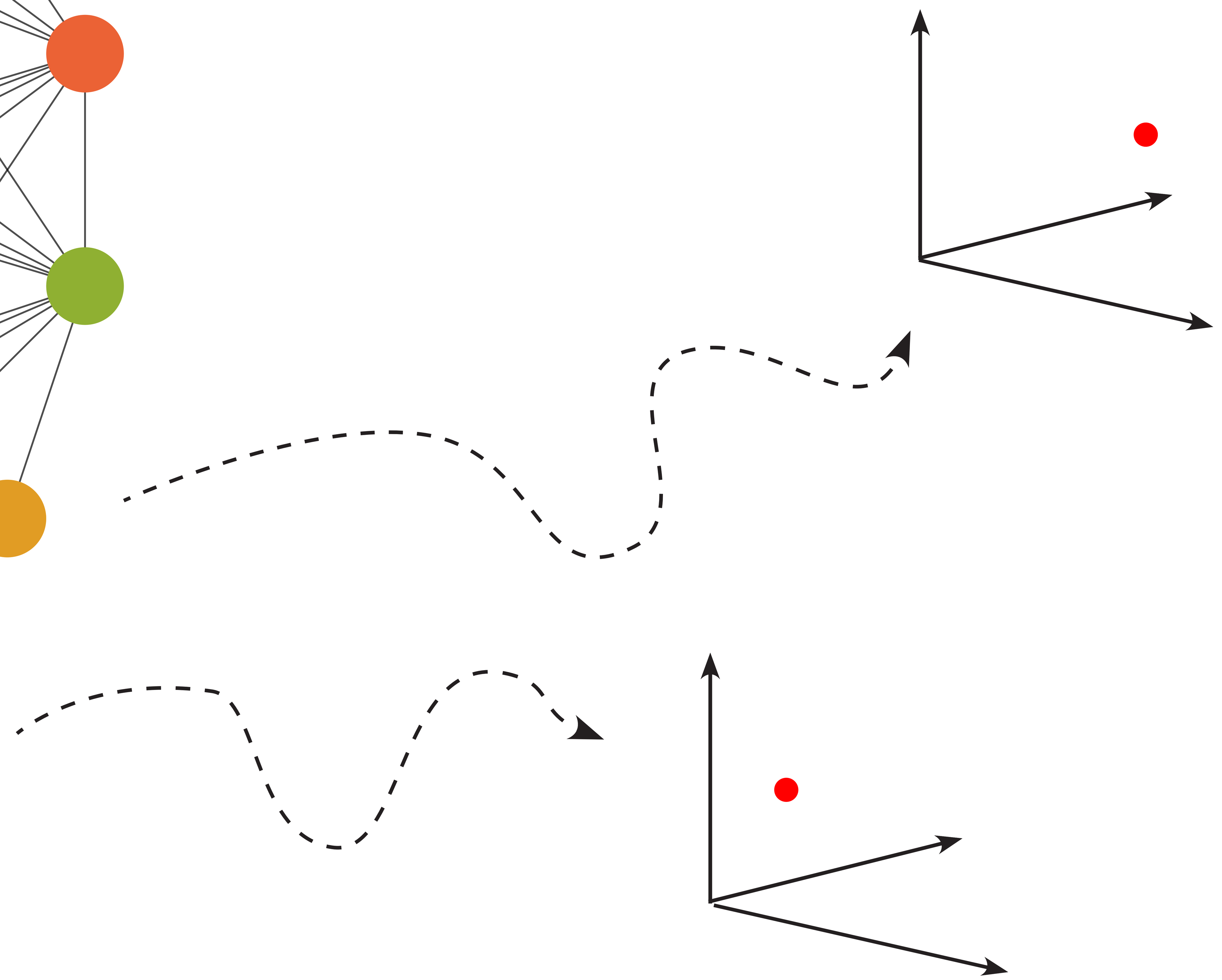
$$L_j : \mathbb{R}^n \to \mathbb{R}^m$$



Self-supervised representation learning



$$L_j : \mathbb{R}^n \to \mathbb{R}^m$$

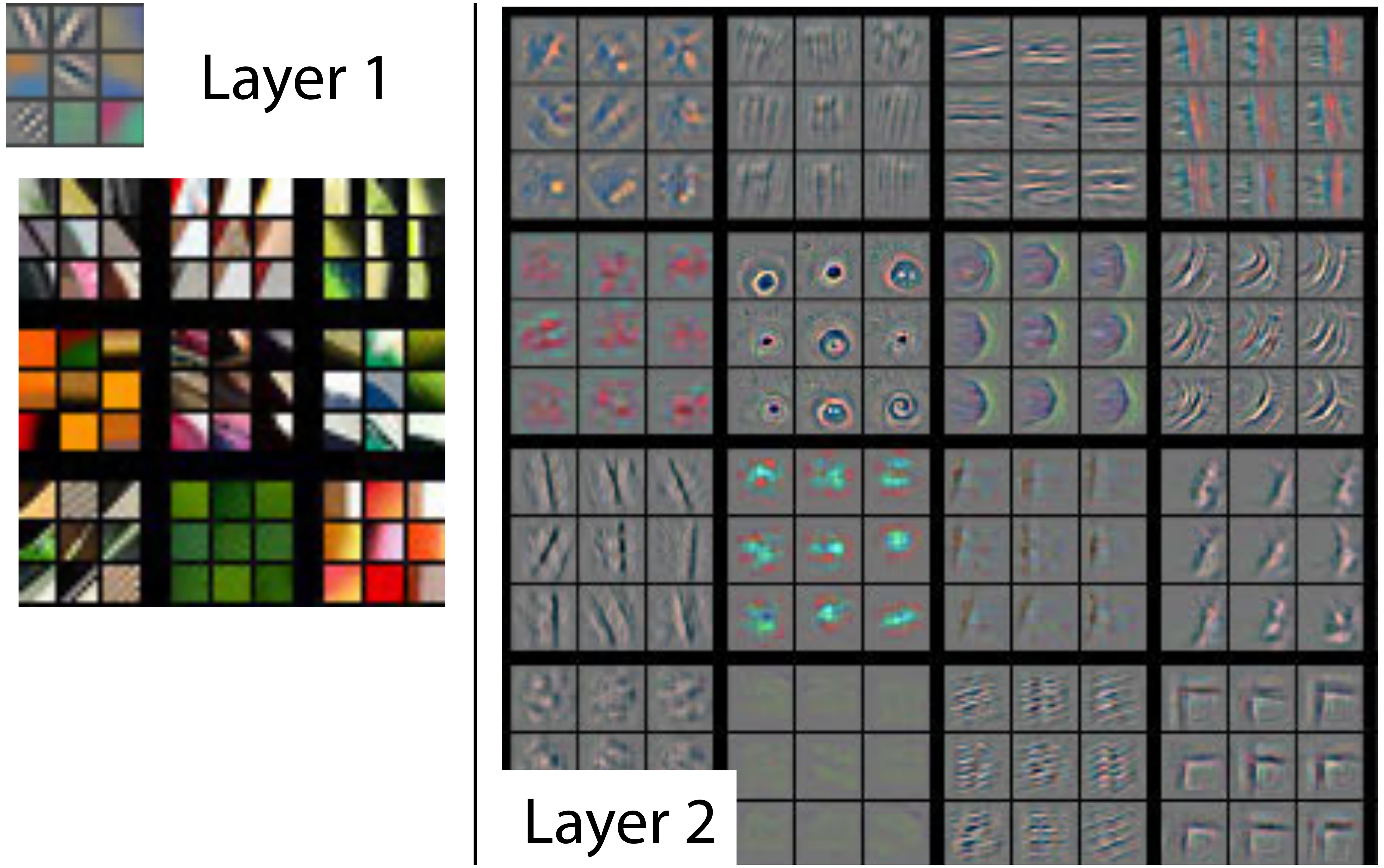


Self-supervised representation learning

- Representation learning
  - Learn a task-independent representation of the data in the *feature space* of the neural network

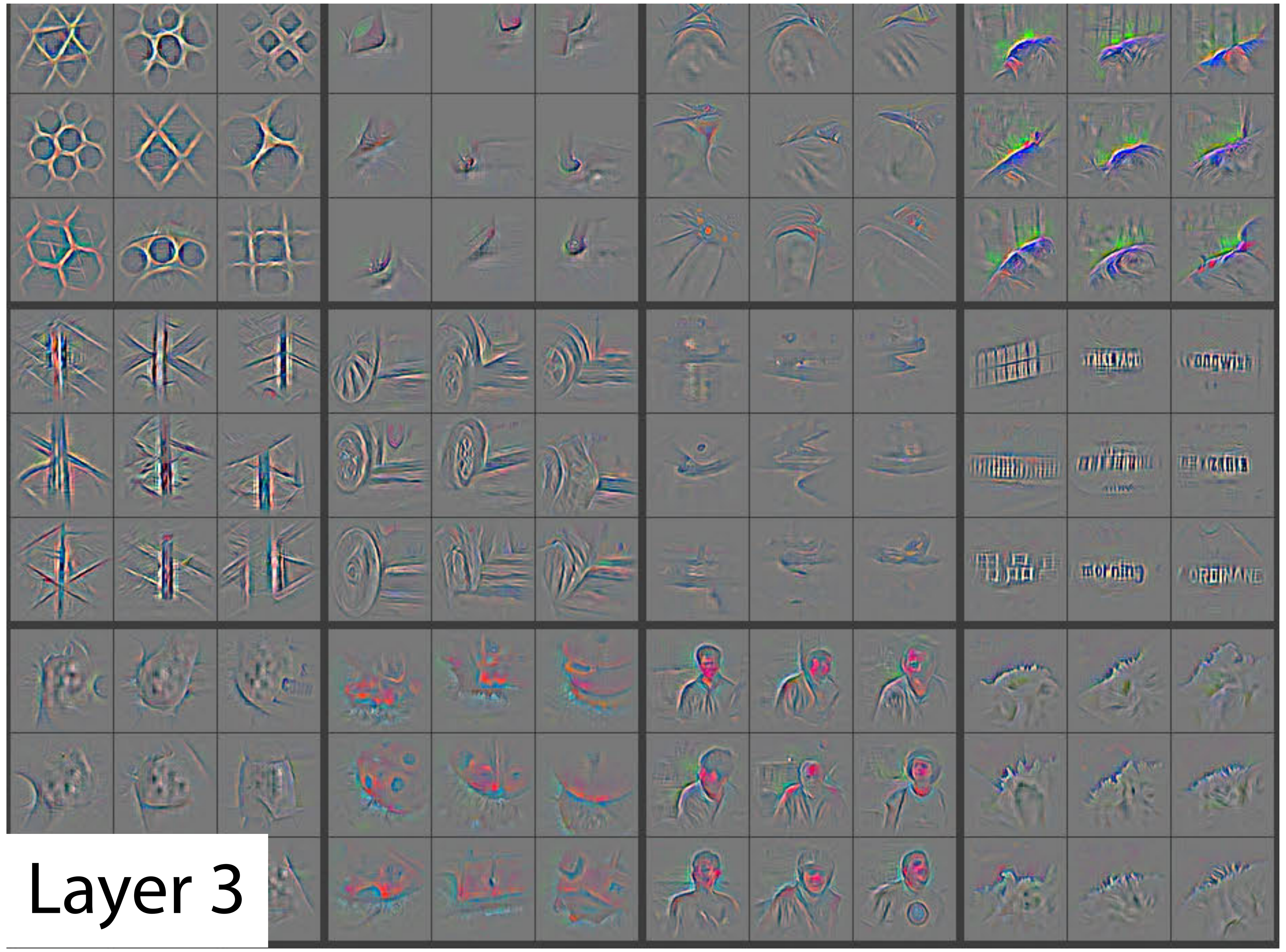
$$\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^{m_1} \to \cdots \to \mathbb{R}^{m_{j-1}} \to \mathbb{R}^m$$

Self-supervised representation learning



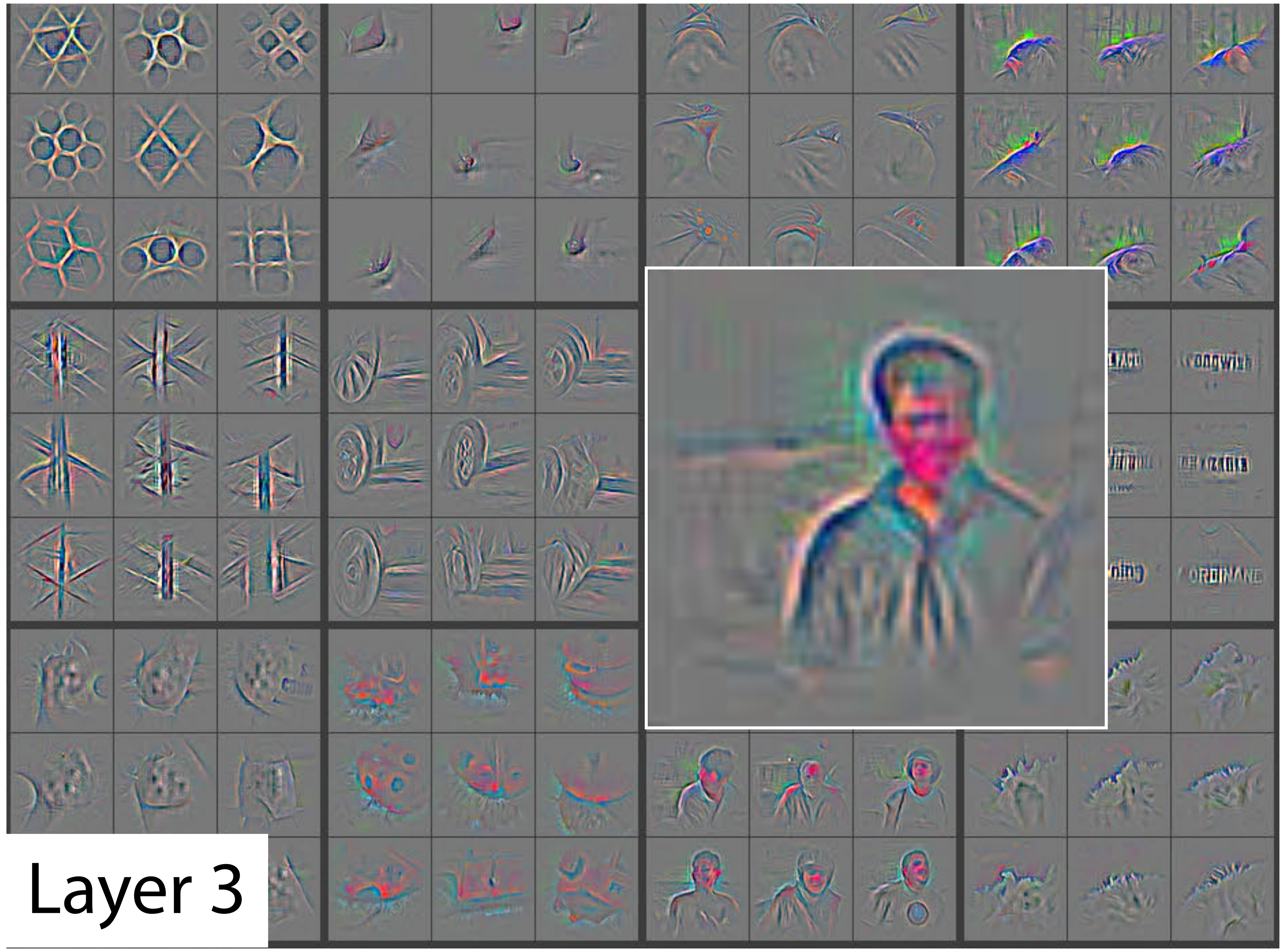
From M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing.

Self-supervised representation learning



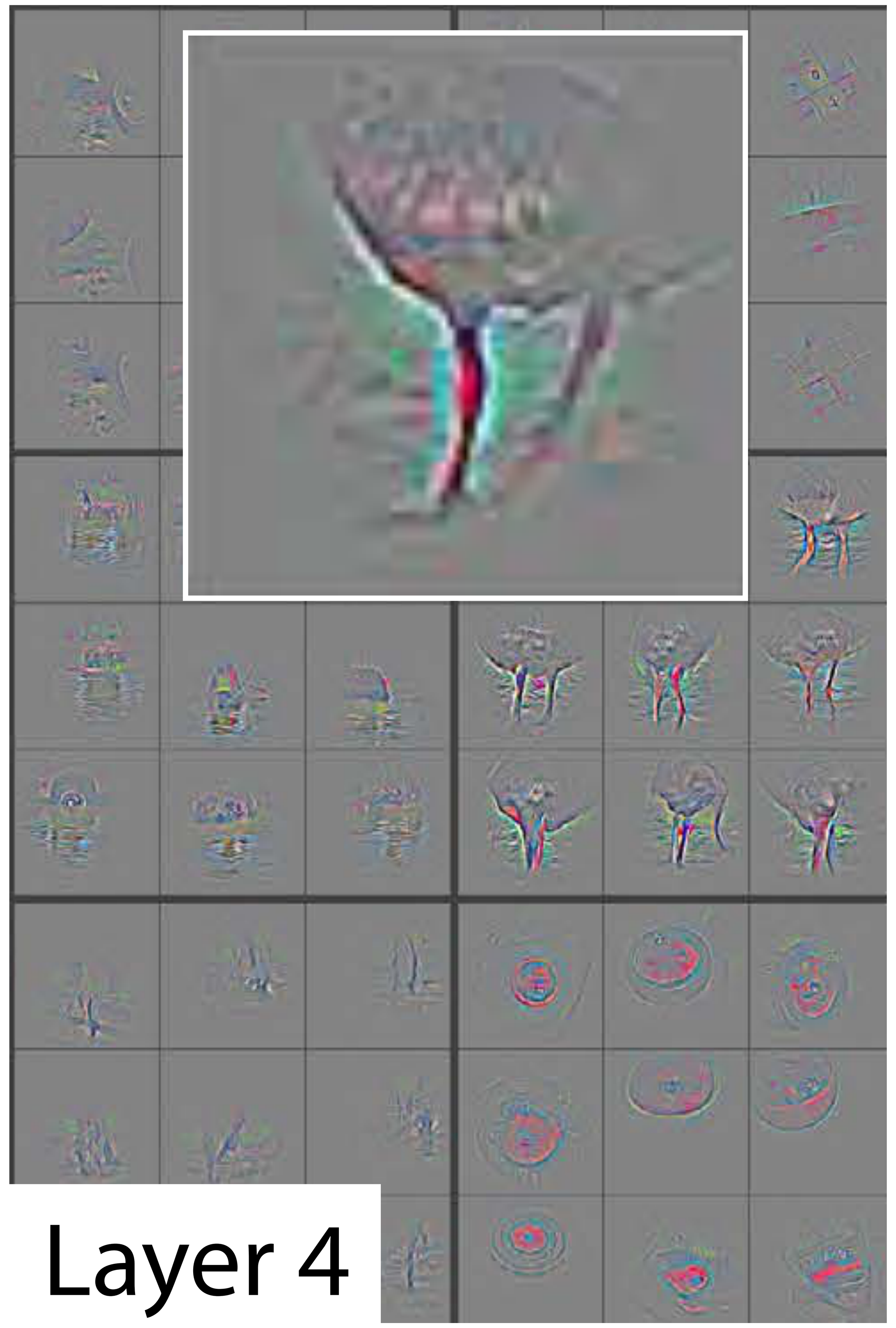
From M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing.

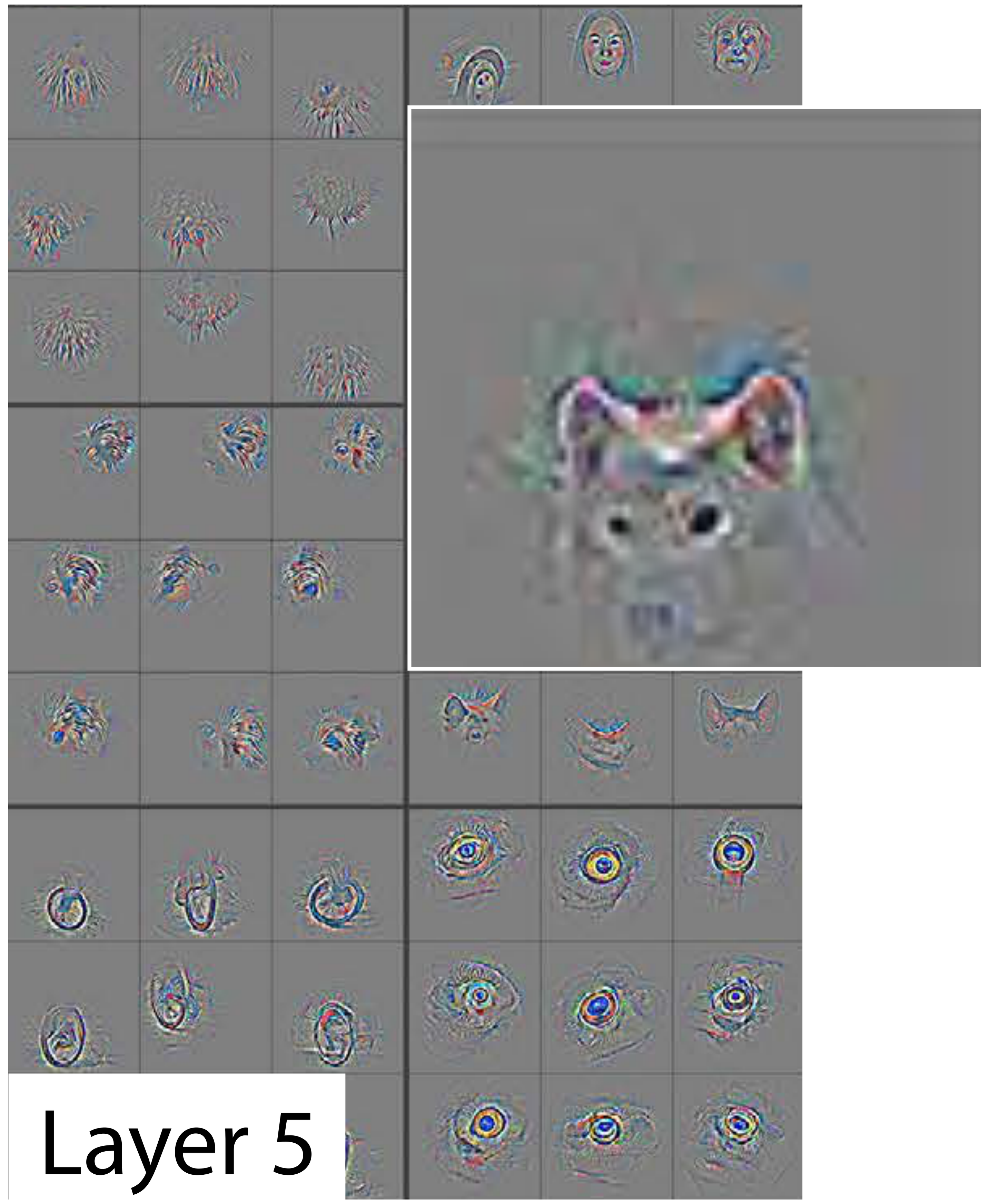
Self-supervised representation learning



From M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing.

Self-supervised representation learning





From M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing.

Self-supervised representation learning

- Representation learning
  - Learn a task-independent representation of the data in the *feature space* of the neural network

$$\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^{m_1} \to \cdots \to \mathbb{R}^{m_{j-1}} \to \mathbb{R}^m$$

Self-supervised representation learning

- Representation learning
  - Learn a task-independent representation of the data in the *feature space* of the neural network

$$\mathcal{N} : \mathbb{R}^n \rightarrow \mathbb{R}^{m_1} \rightarrow \cdots \rightarrow \mathbb{R}^{m_{j-1}} \rightarrow \mathbb{R}^m$$

- Self-supervised training
  - Train with “labels” intrinsic to the data

Self-supervised representation learning

- Self-supervised pretext task:

Self-supervised representation learning

- Self-supervised pretext task: inpainting of randomly deleted image parts[1]

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Self-supervised representation learning

- Self-supervised pretext task: inpainting of randomly delet-ed image parts[1]

(a) Input context

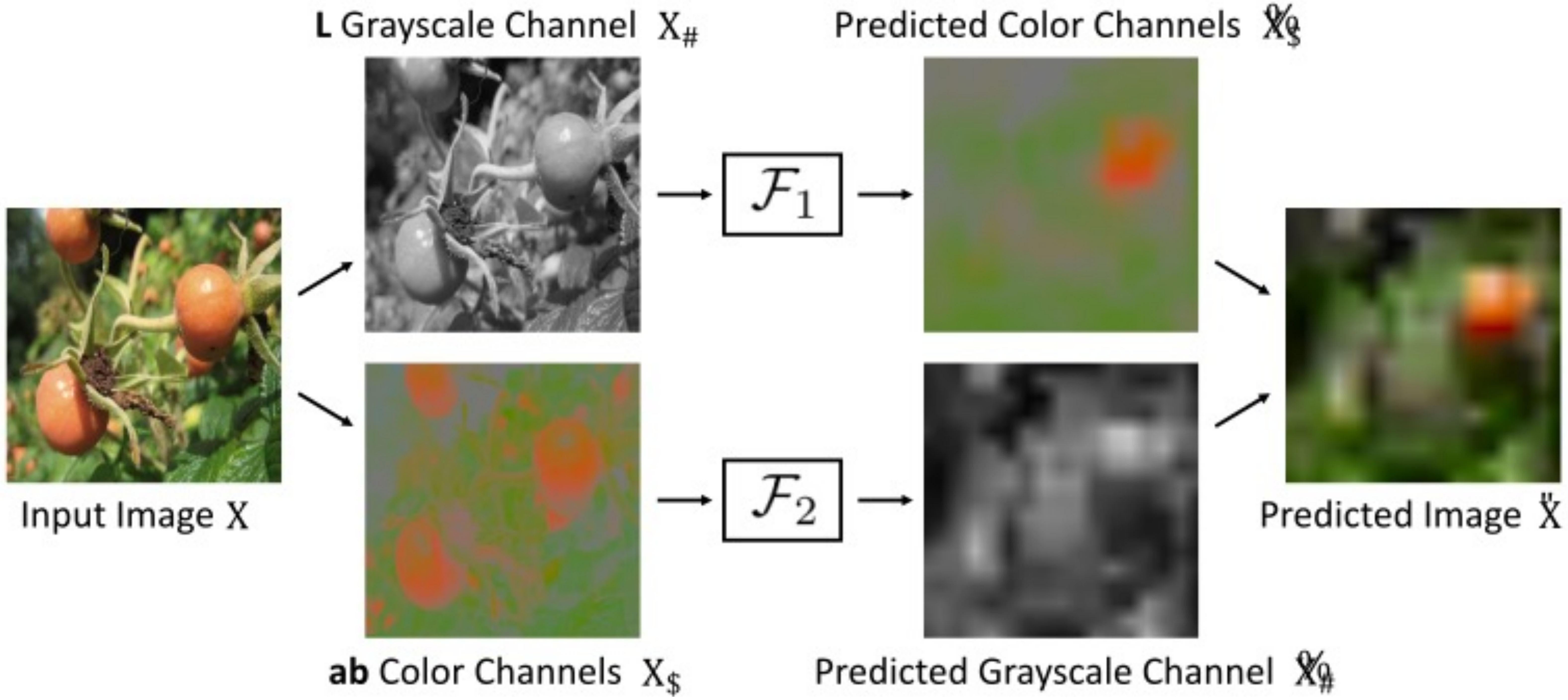
(c) Context Encoder (L 2 loss)

(d) Context Encoder (L 2 +Adversarial loss)

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encod-ers: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Self-supervised representation learning

- Self-supervised pretext task: predicting deleted color and gray scale channels[1]



[1] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

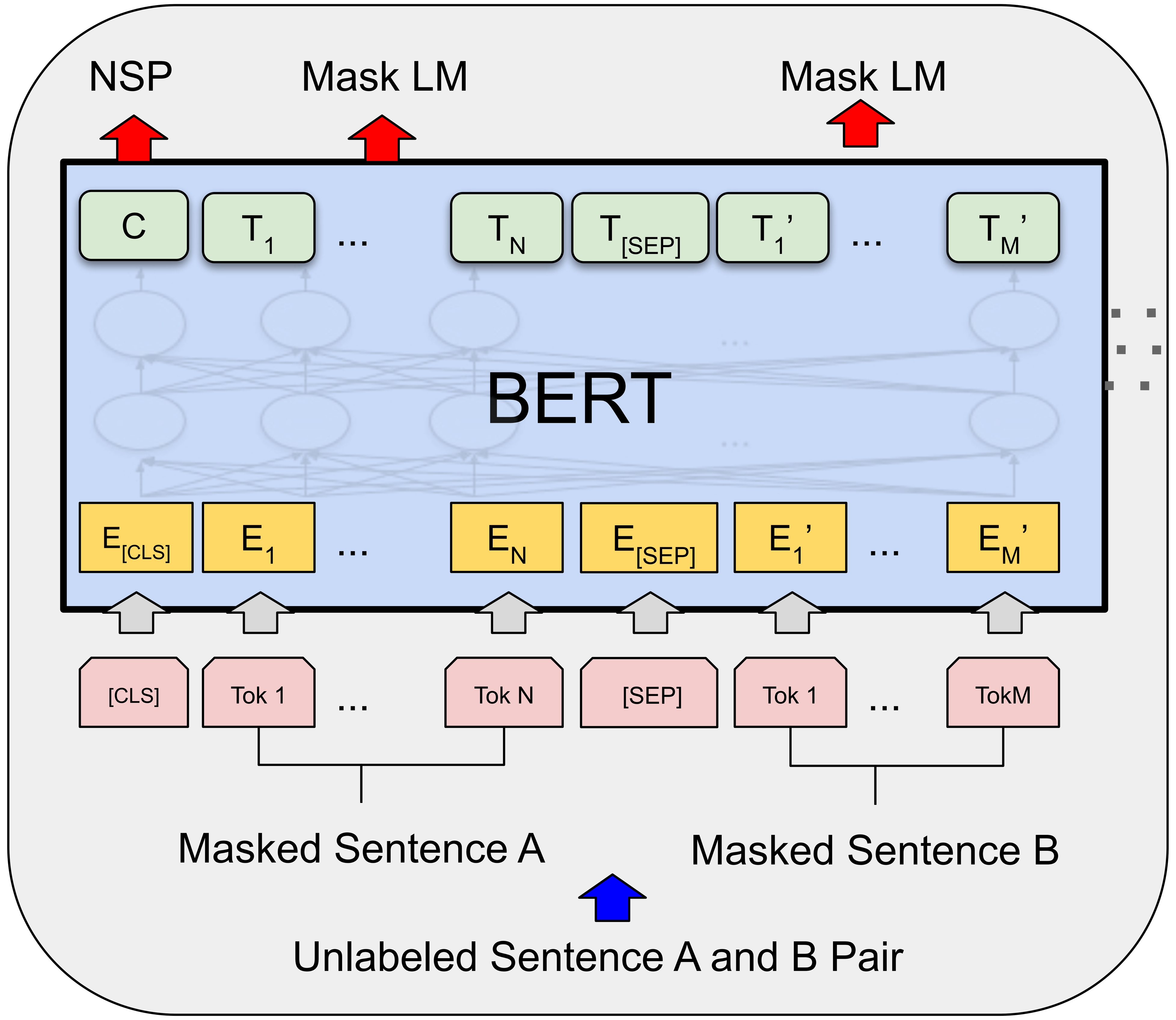
Self-supervised representation learning

- BERT[1]
  - Self-supervised representation learning for natural language processing (NLP)
  - Very large transformer neural network with billions of parameters
  - Self-supervised training essentially only feasible option at this scale

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

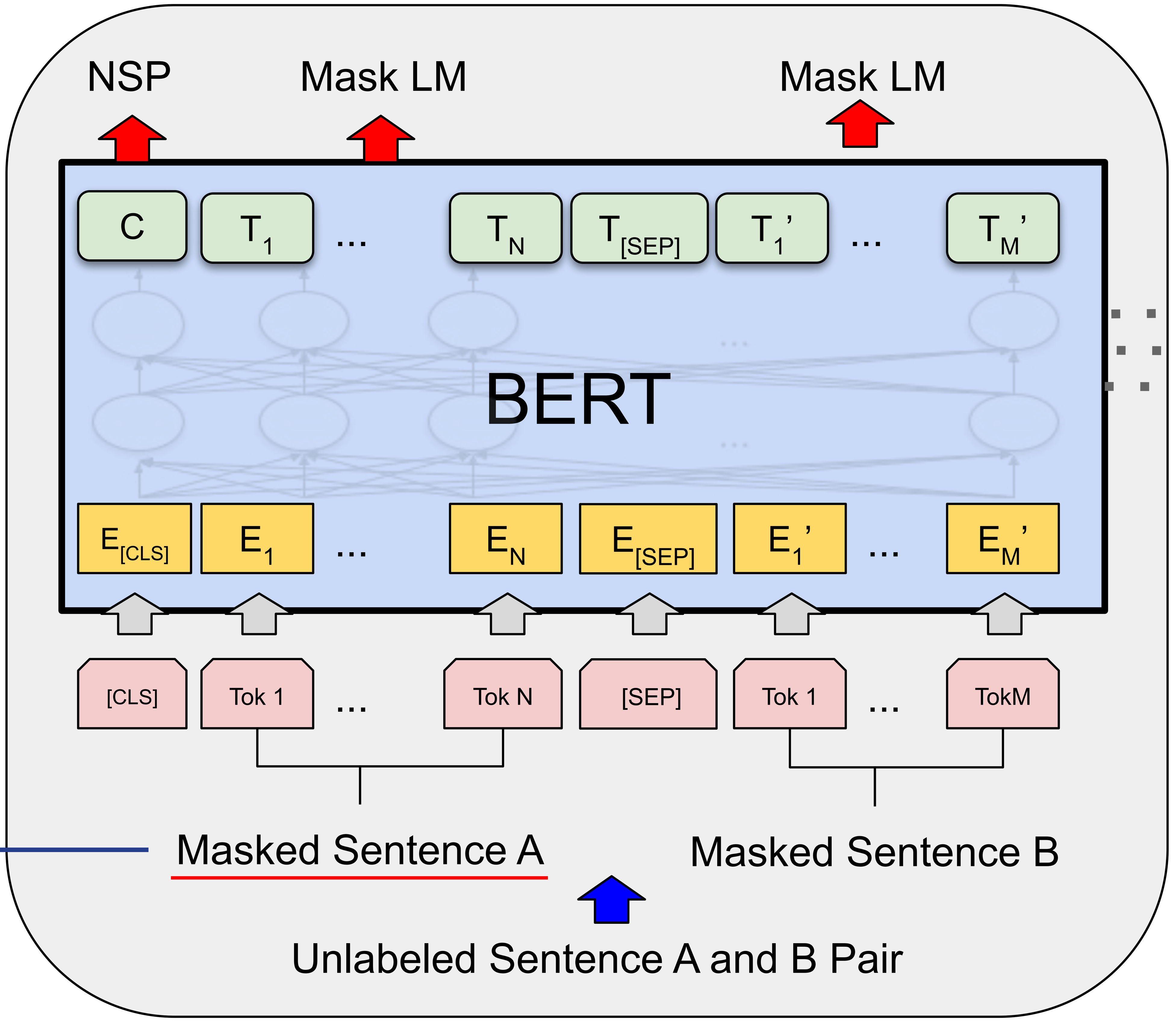
- BERT[1]



[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

- BERT[1]



The sun was shining ~~bright~~.

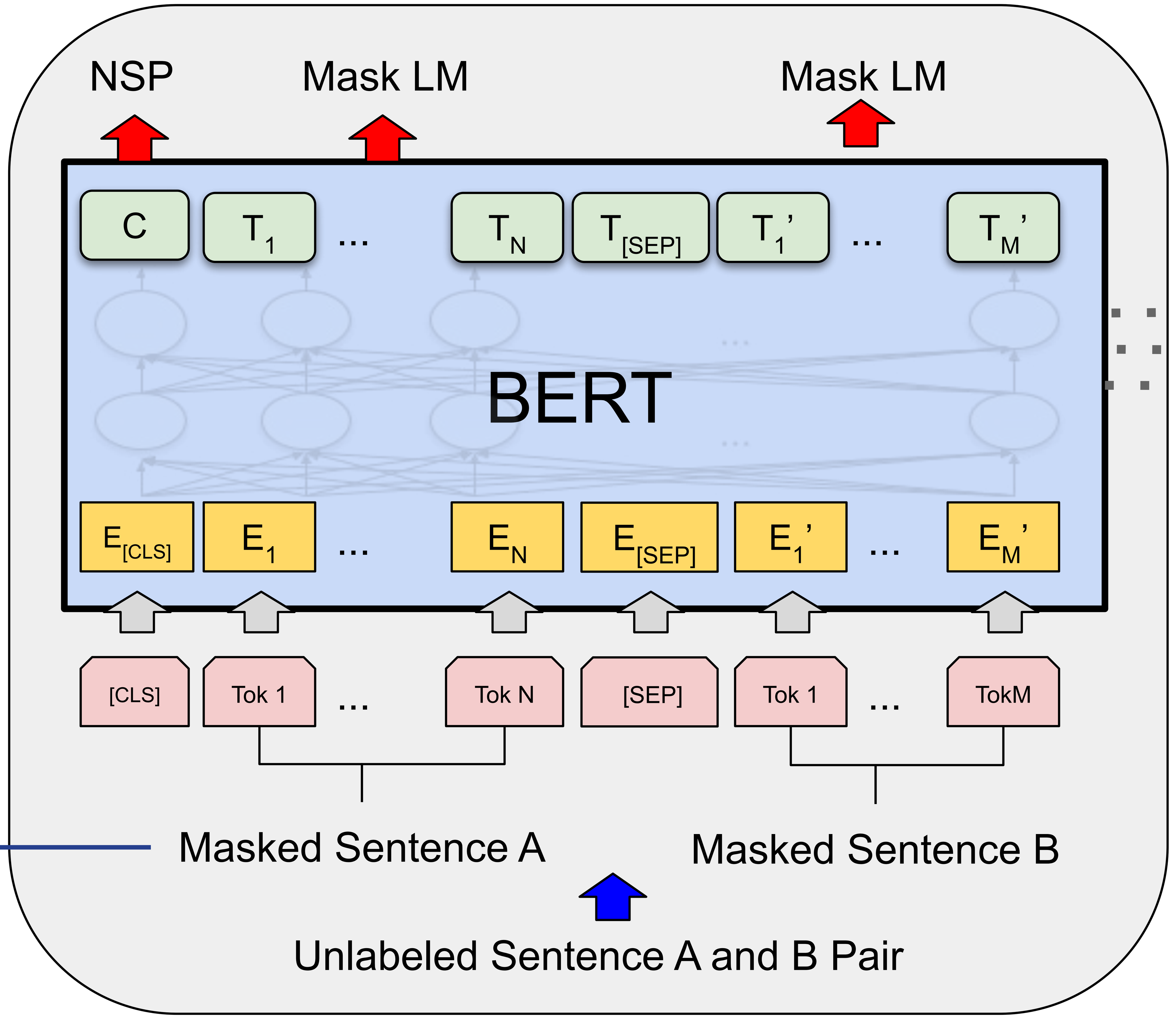
[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

- BERT[1]

Network predicts deleted word

The sun was shining ~~bright~~.

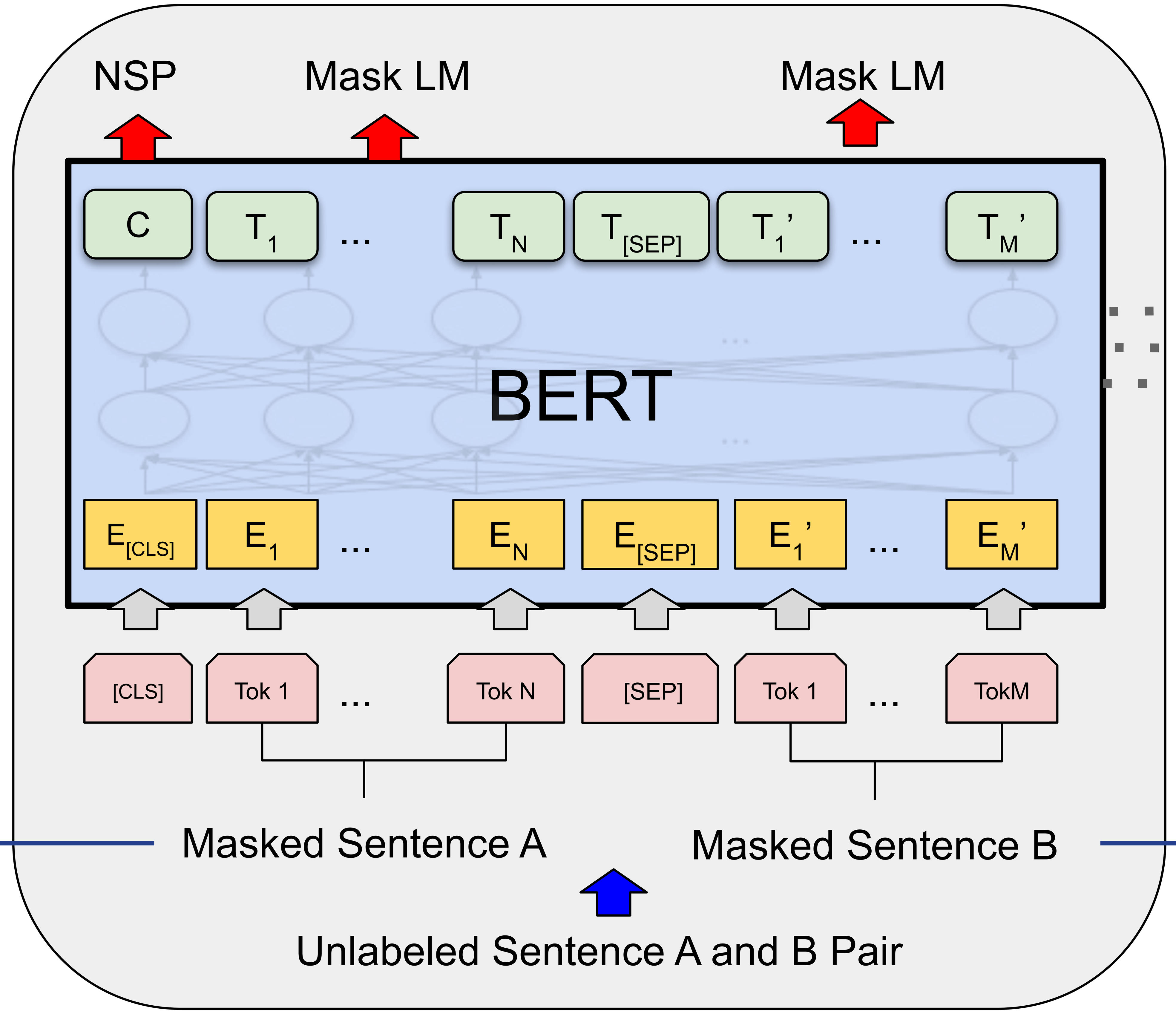


[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

- BERT[1]

Network predicts deleted word



Network predicts deleted word

The sun was shining ~~bright~~.

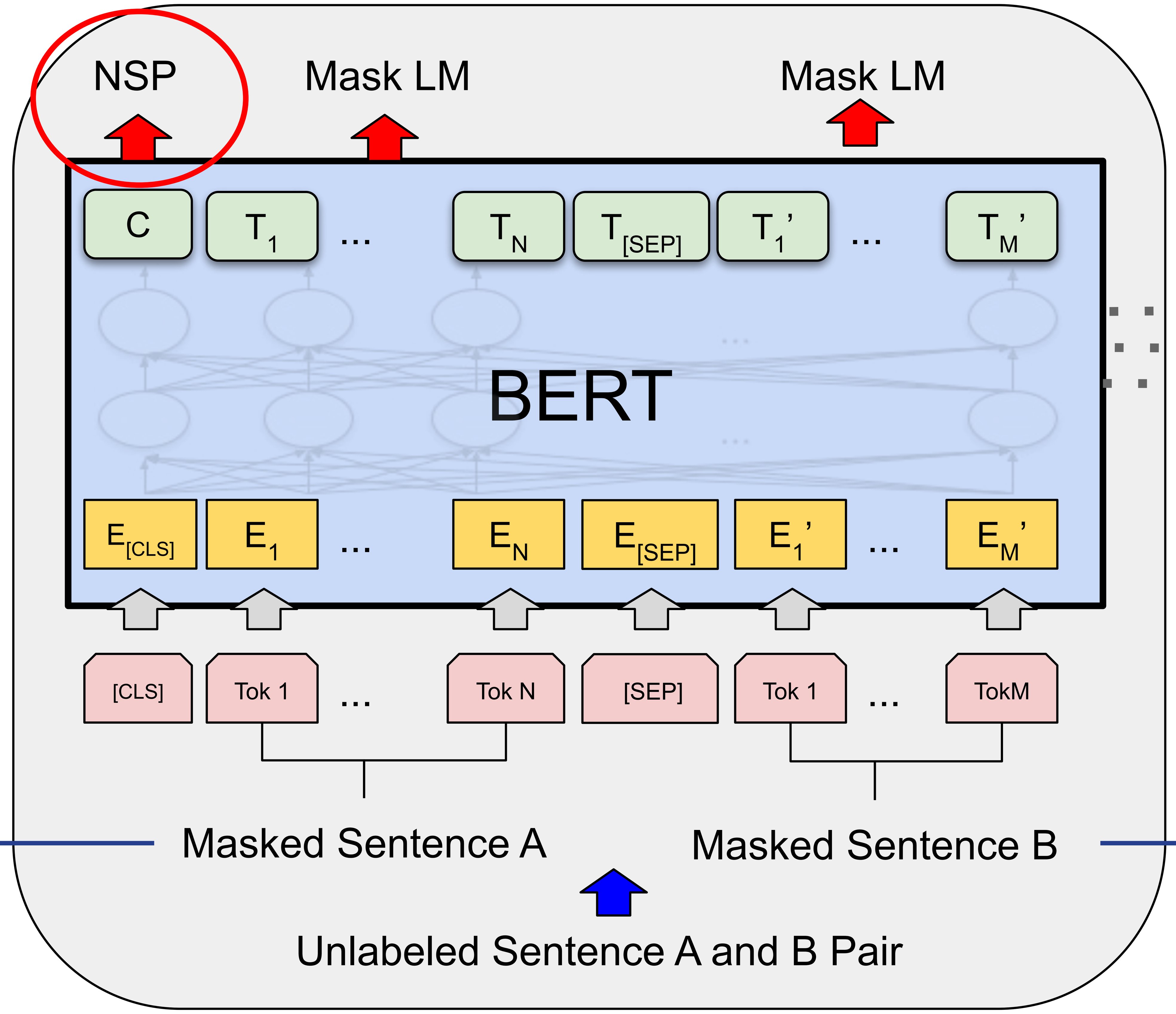
She is ~~wearing~~ a gray shirt

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

- BERT[1]

binary next sentence prediction



The sun was shining ~~bright~~.

She is ~~wearing~~ a gray shirt

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

Self-supervised representation learning

- BERT[1]

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD 1.1 results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

Performance of fine-tuned model on question-answer benchmark

Transformer neural networks

◦ Scale well to highly parallel training and billions of parameters, especially for sequential data

Transformer neural networks

- Scale well to highly parallel training and billions of parameters, especially for sequential data

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	—
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	—
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	—
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	—
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	—
TPUv3-core-days	$\boxed{\text{2.5k}}$	0.68k	0.23k	$\boxed{\text{9.9k}}$	12.3k

very large transformer

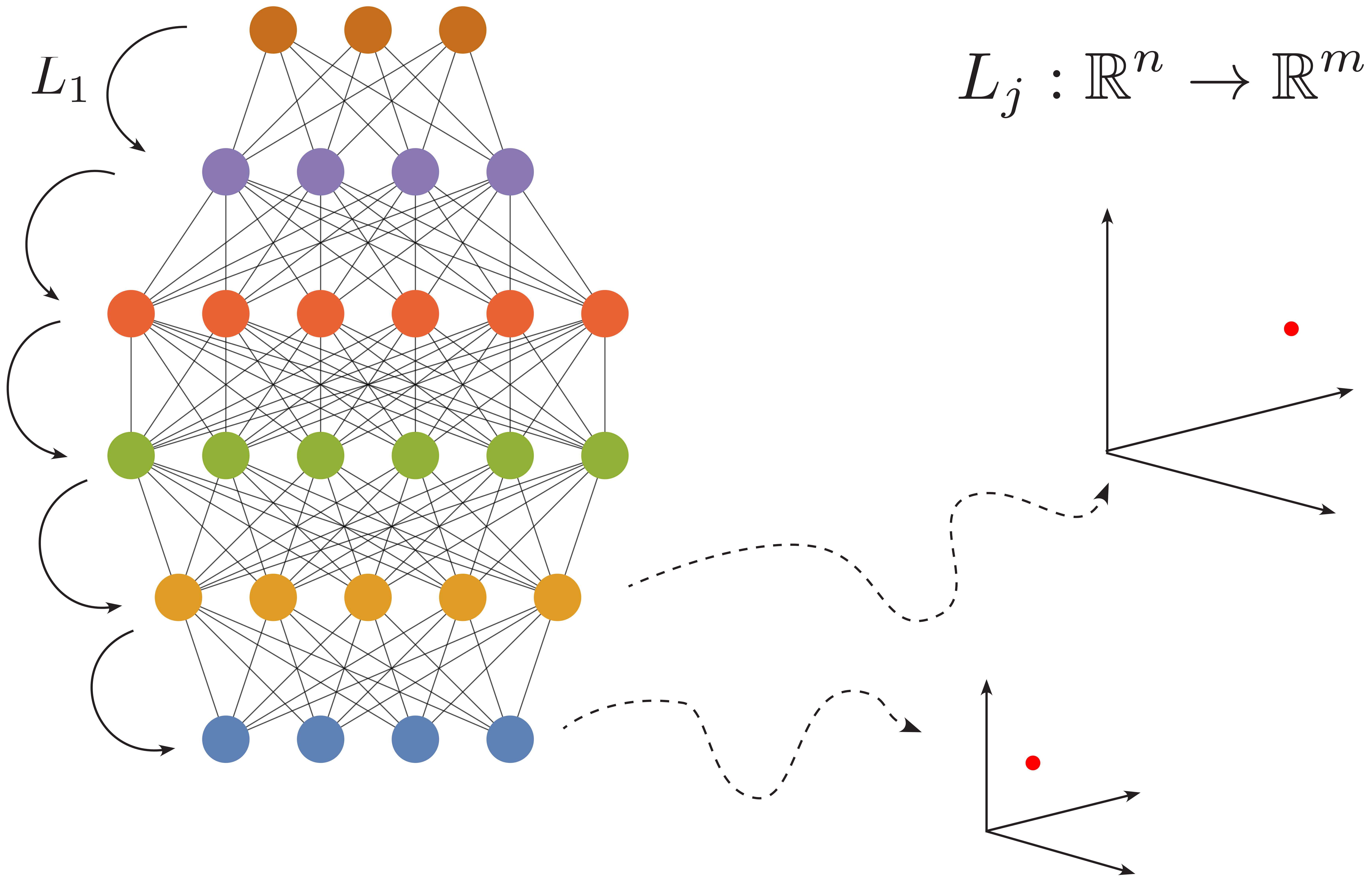
large convolutional network

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

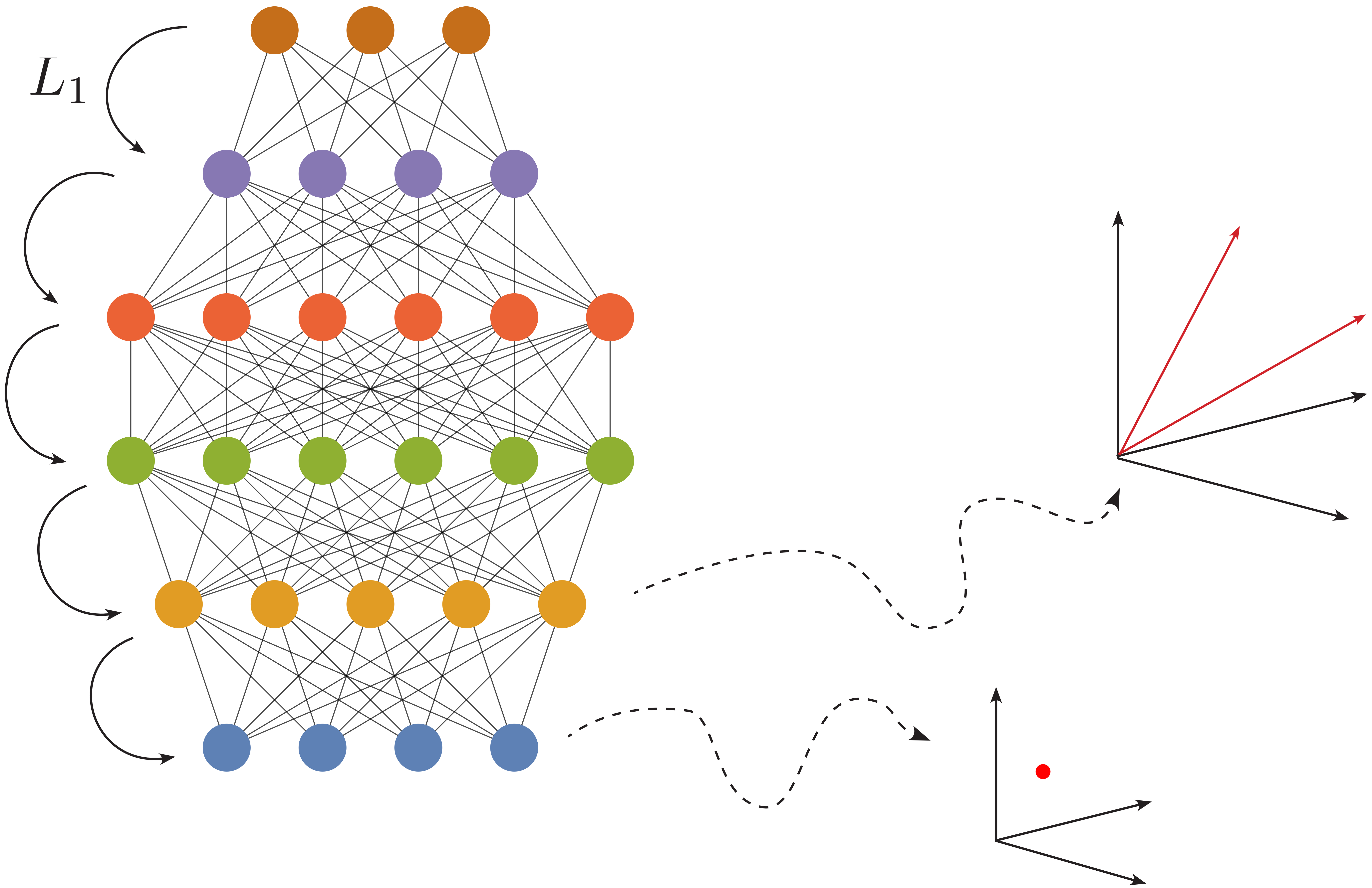
Transformer neural networks

- Scale well to highly parallel training and billions of parameters, especially for sequential data
- Attention mechanism allows to model complex dependencies in data

Transformer neural networks

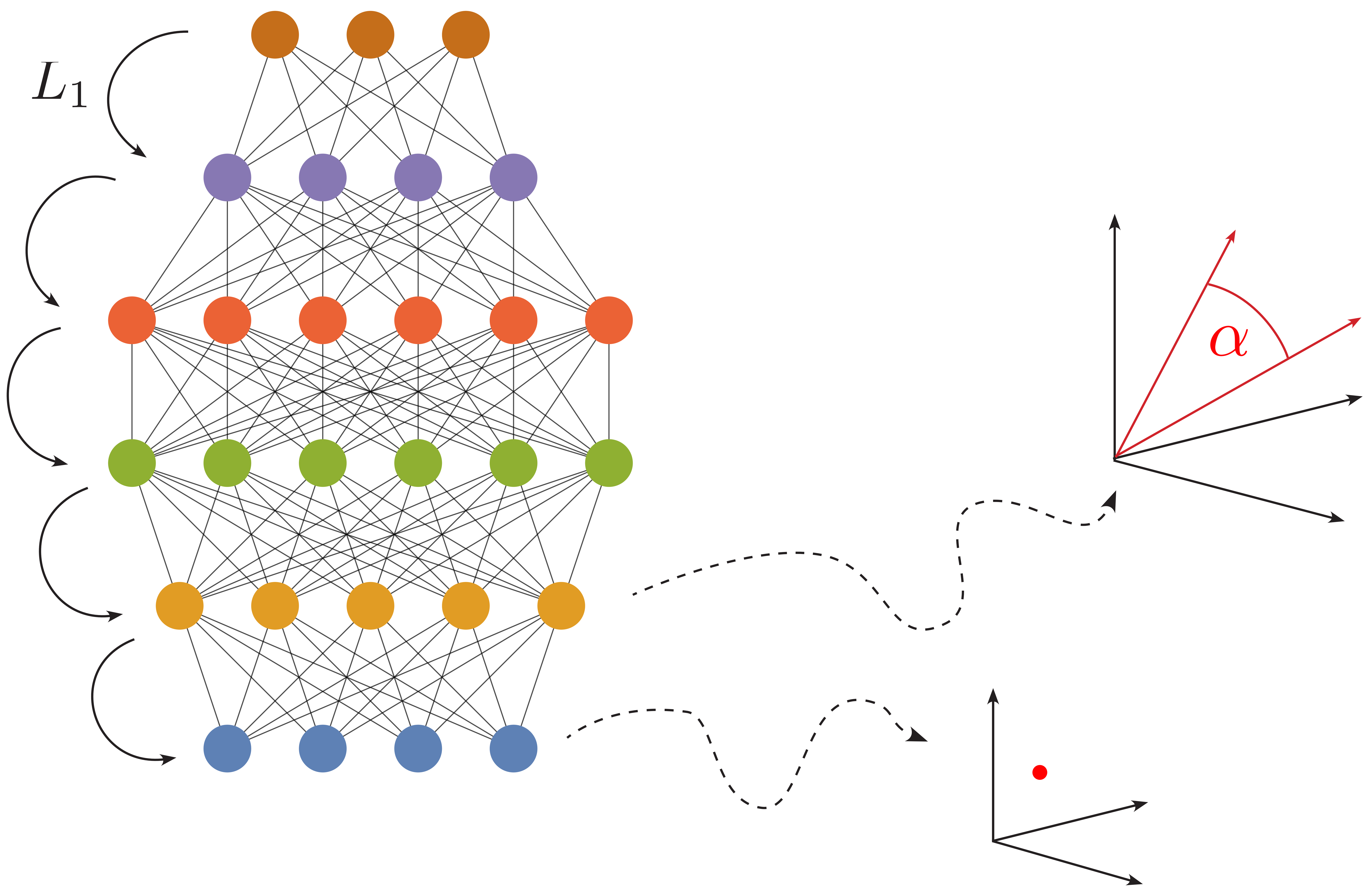


Transformer neural networks



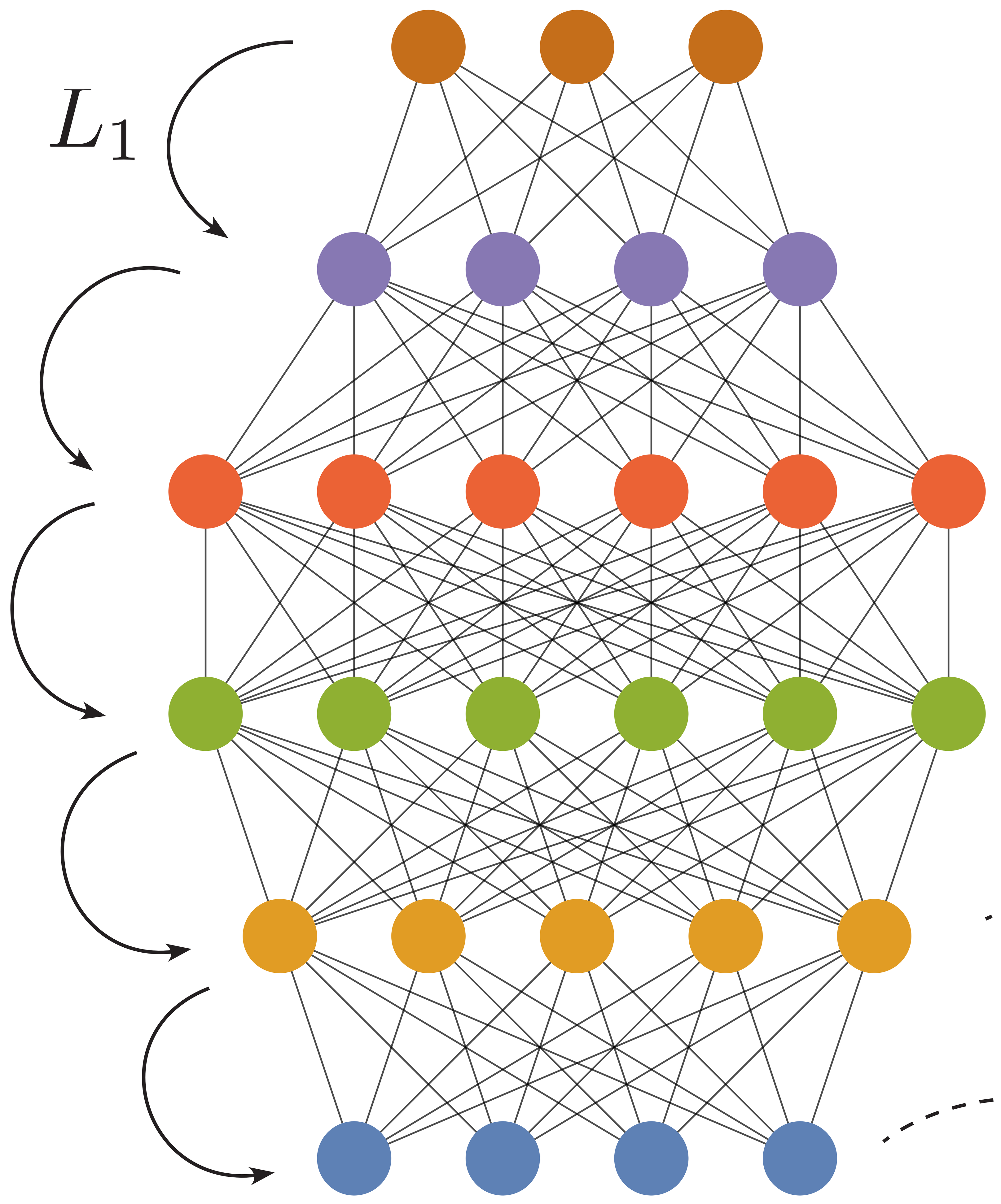
Transformer neural networks

Transformer neural networks

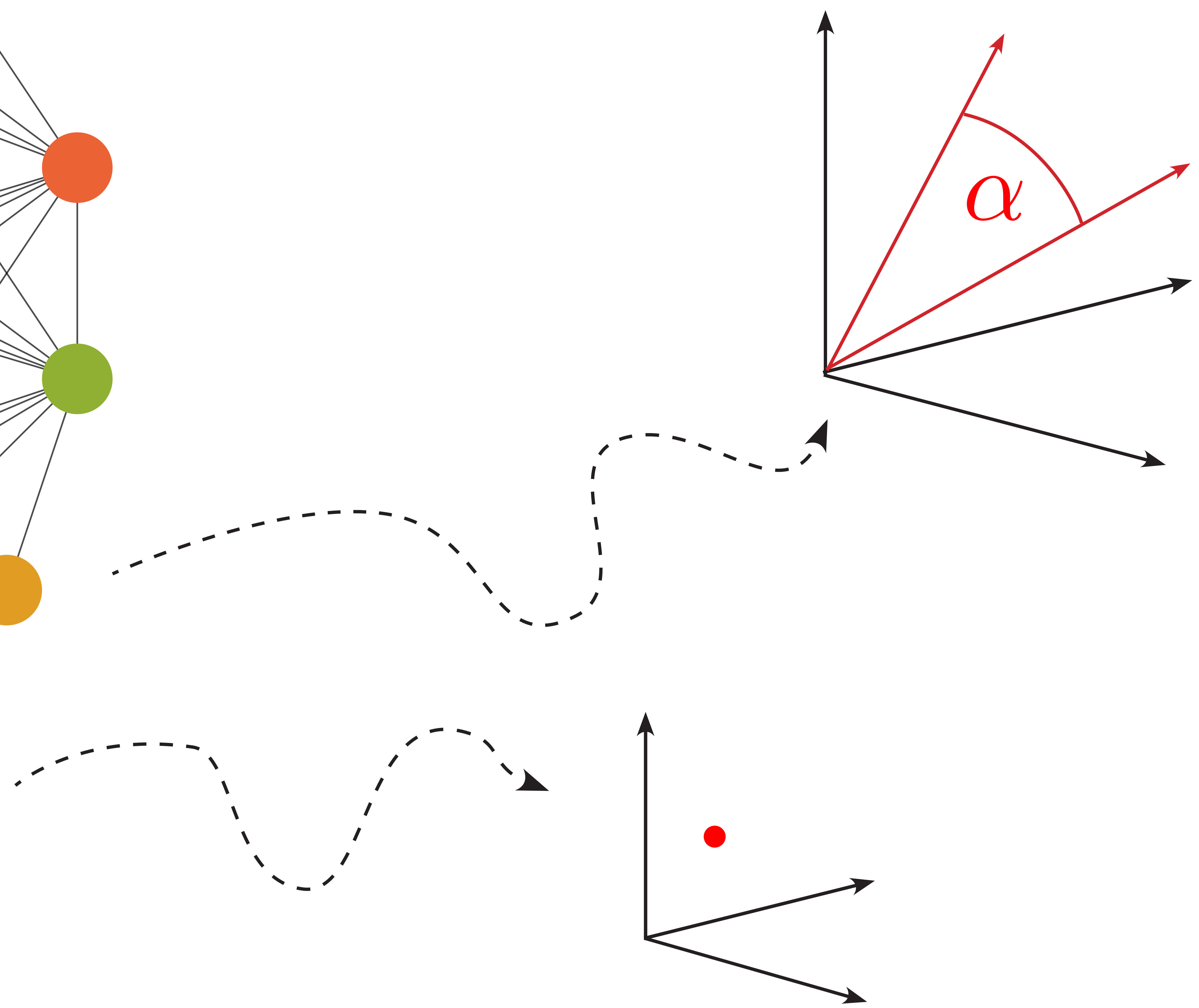


correlation between inputs in feature space: attention

Transformer neural networks



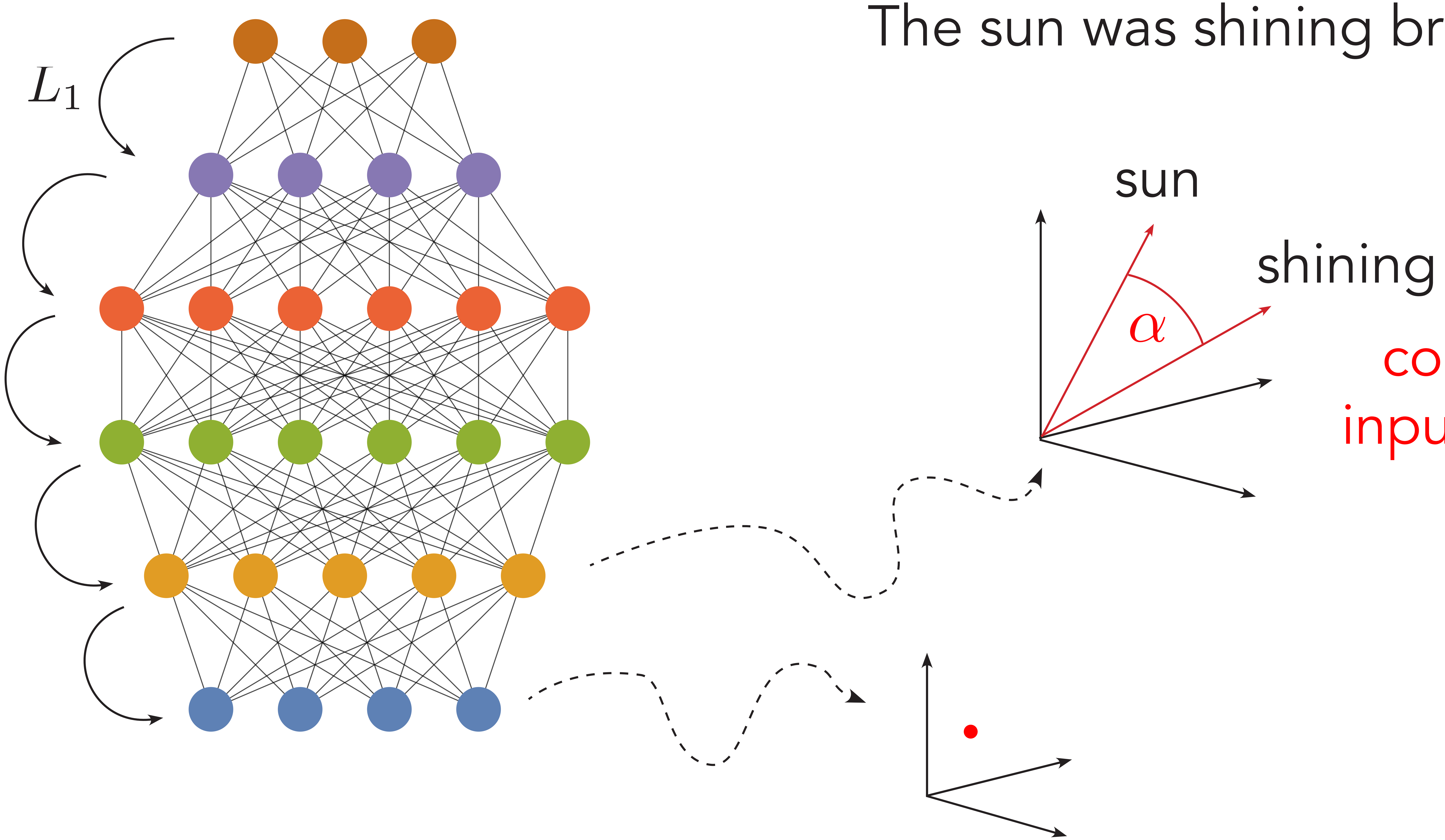
The sun was shining bright.



correlation between
inputs in feature space:
attention

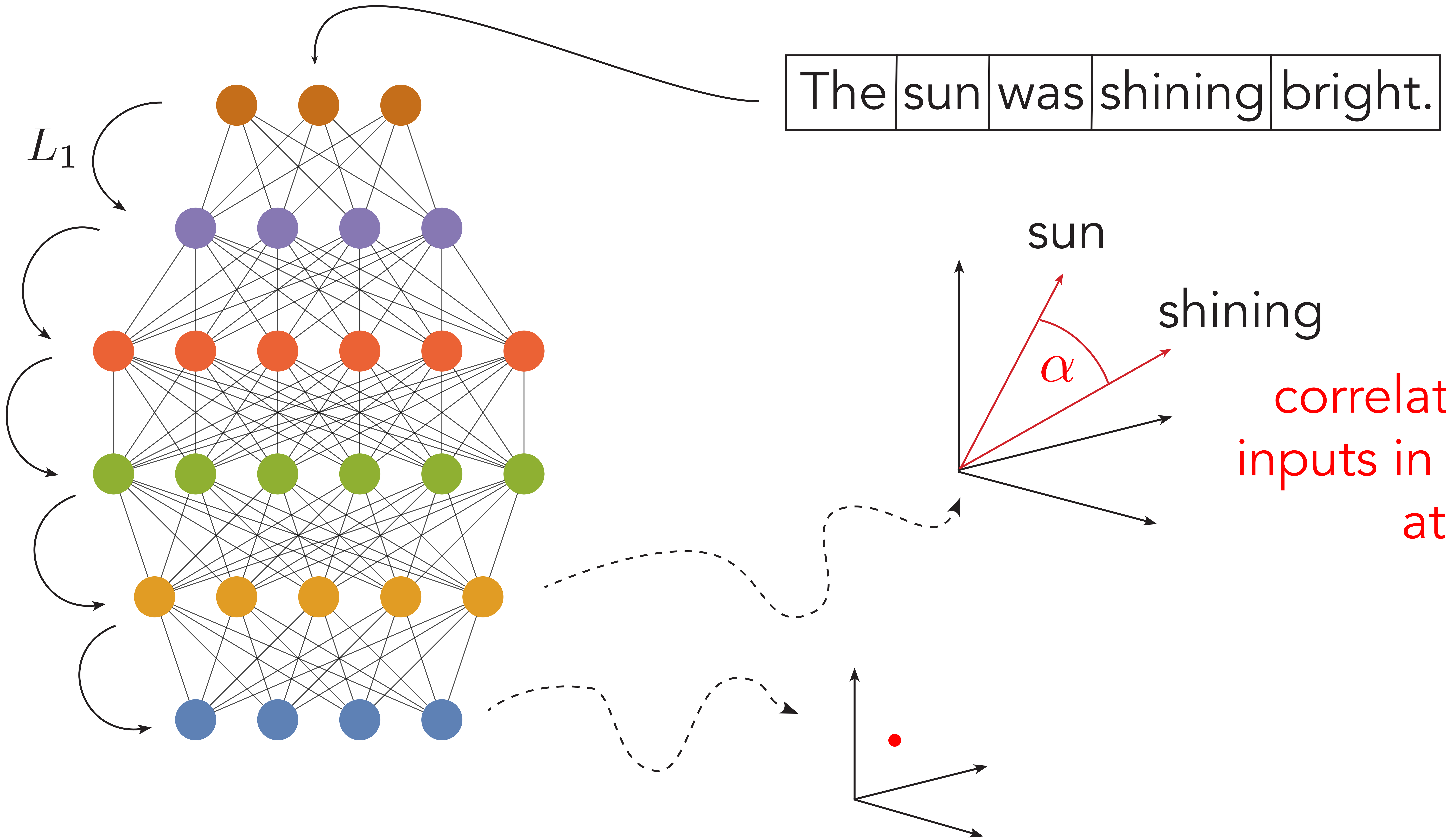
Transformer neural networks

The sun was shining bright.



correlation between inputs in feature space: attention

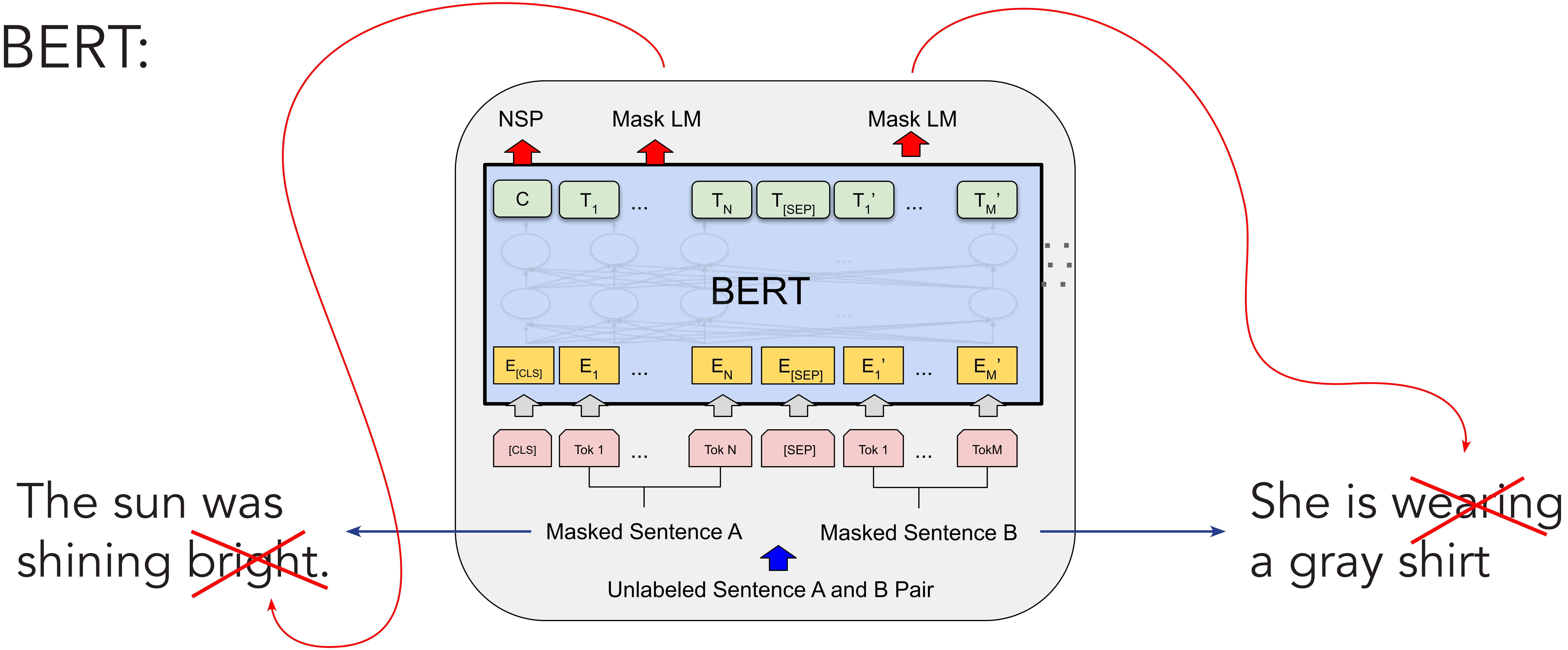
Transformer neural networks



correlation between inputs in feature space: attention

Transformer neural networks

- BERT:



Transformer neural networks

- BERT:

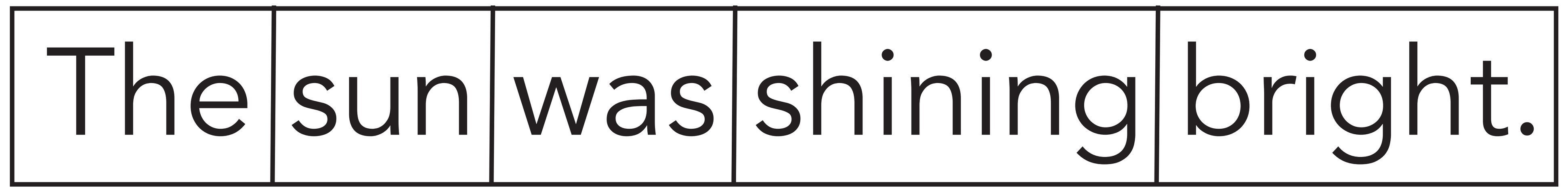
"The law will never be perfect, but its application should be just, this is what we are missing, in my opinion."

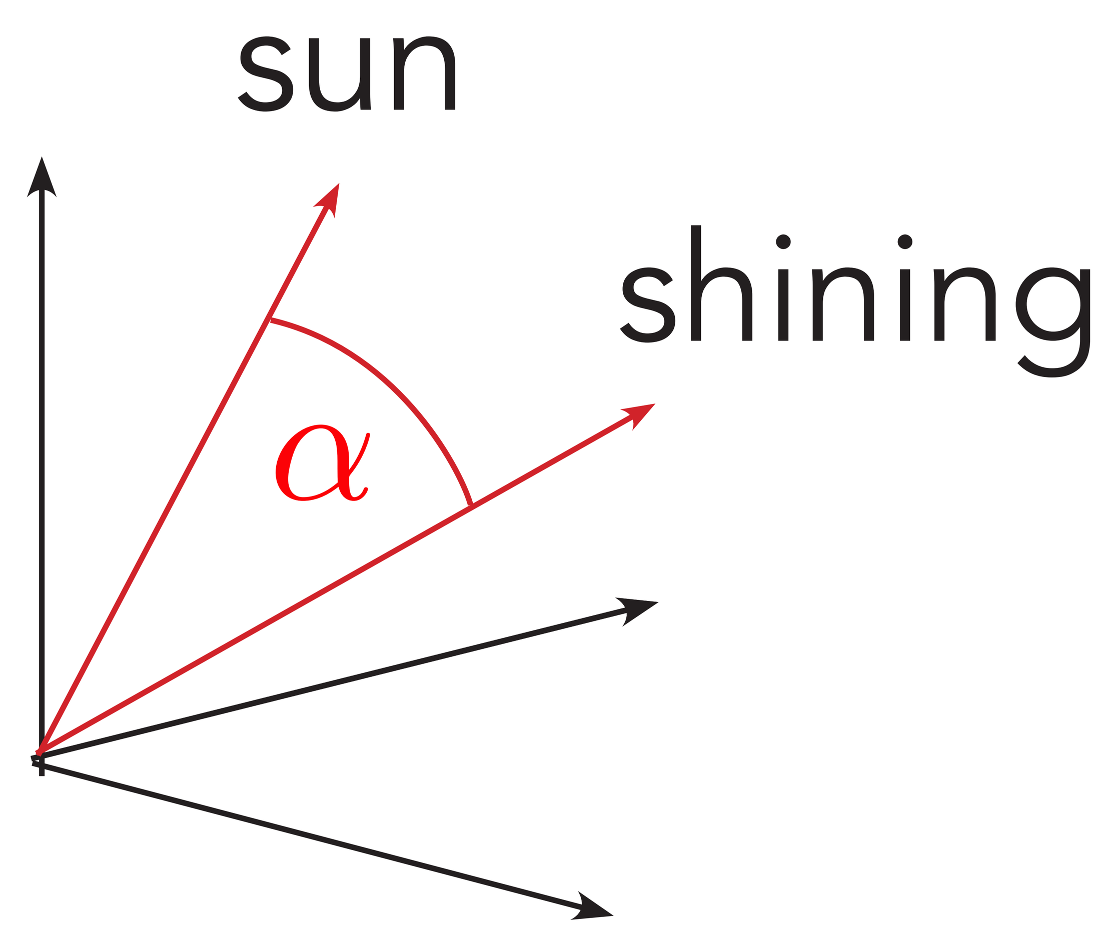
Transformer neural networks

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Transformer neural networks

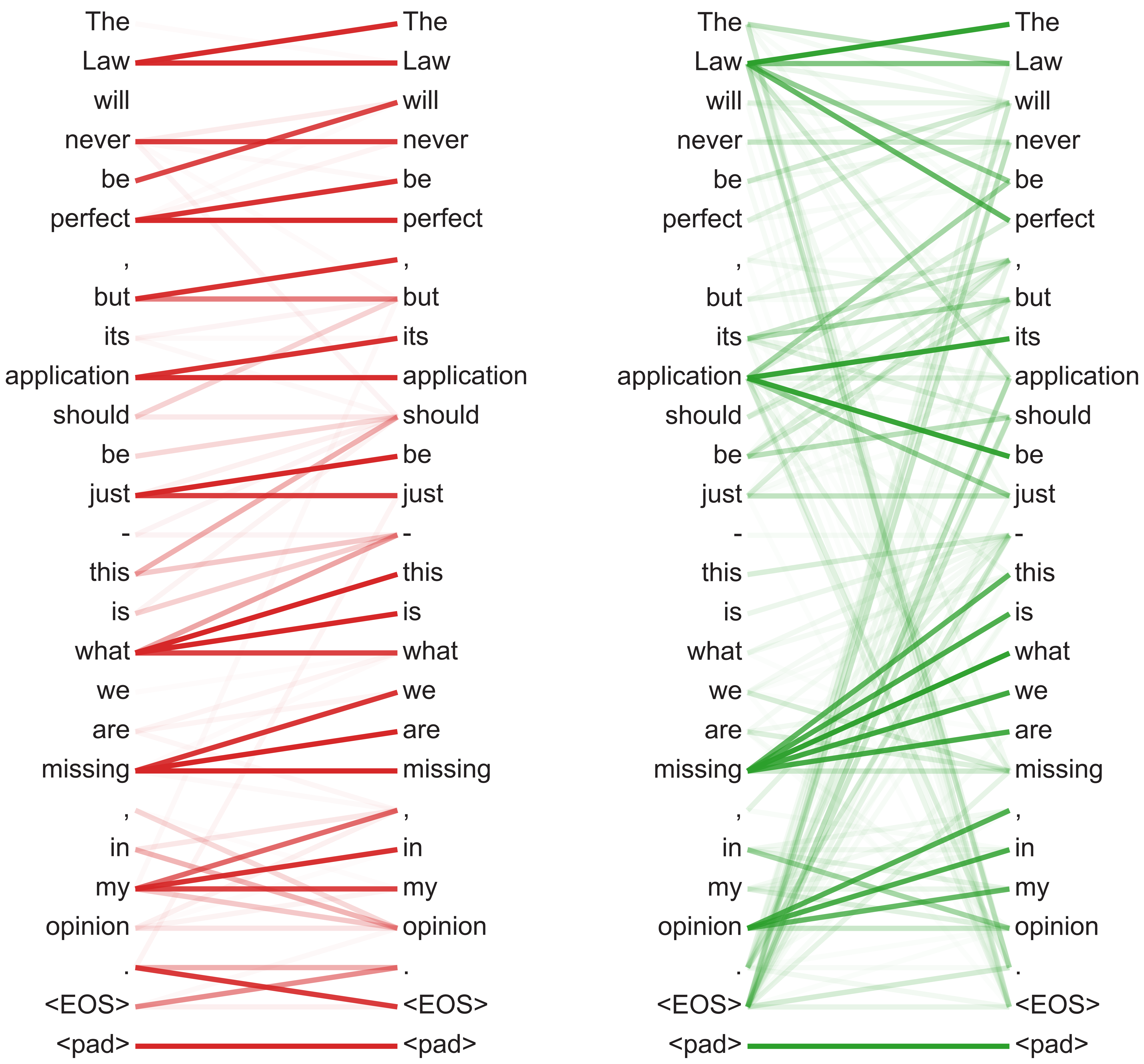
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.





Transformer neural networks

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.



Transformer neural networks

- DINO (computer vision):

M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. CoRR, abs/2104.14294, 2021.

Transformer neural networks

- DINO (computer vision):

M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. CoRR, abs/2104.14294, 2021.

Transformer neural networks

- Scale well to highly parallel training and billions of parameters, especially for sequential data
- Attention mechanism allows to model complex dependencies in data
  - Learned inner product in feature space with “legs” in input domain
  - Direct interpretation of learned representations through attention maps

How can we adapt these ideas to the Earth sciences?

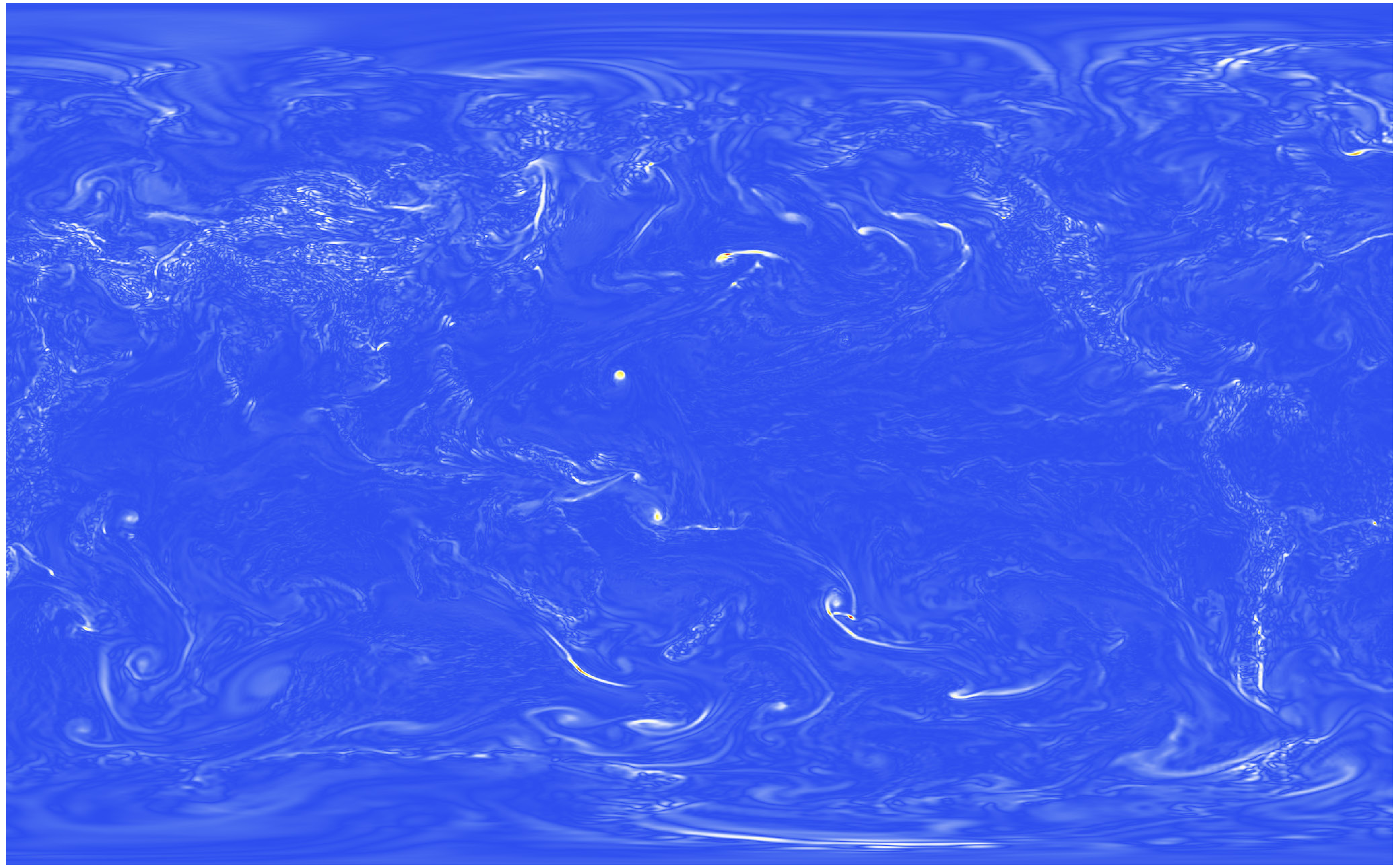
The facts

- Large amounts of data and growing fast:
  - ERA5, OCEAN5, MetOp-SG, ...
  - Unlabeled data
  - (Quasi-) observational data describing effects from the whole system
- Very limited amounts of labeled data for many appli-cations

AtmoDist

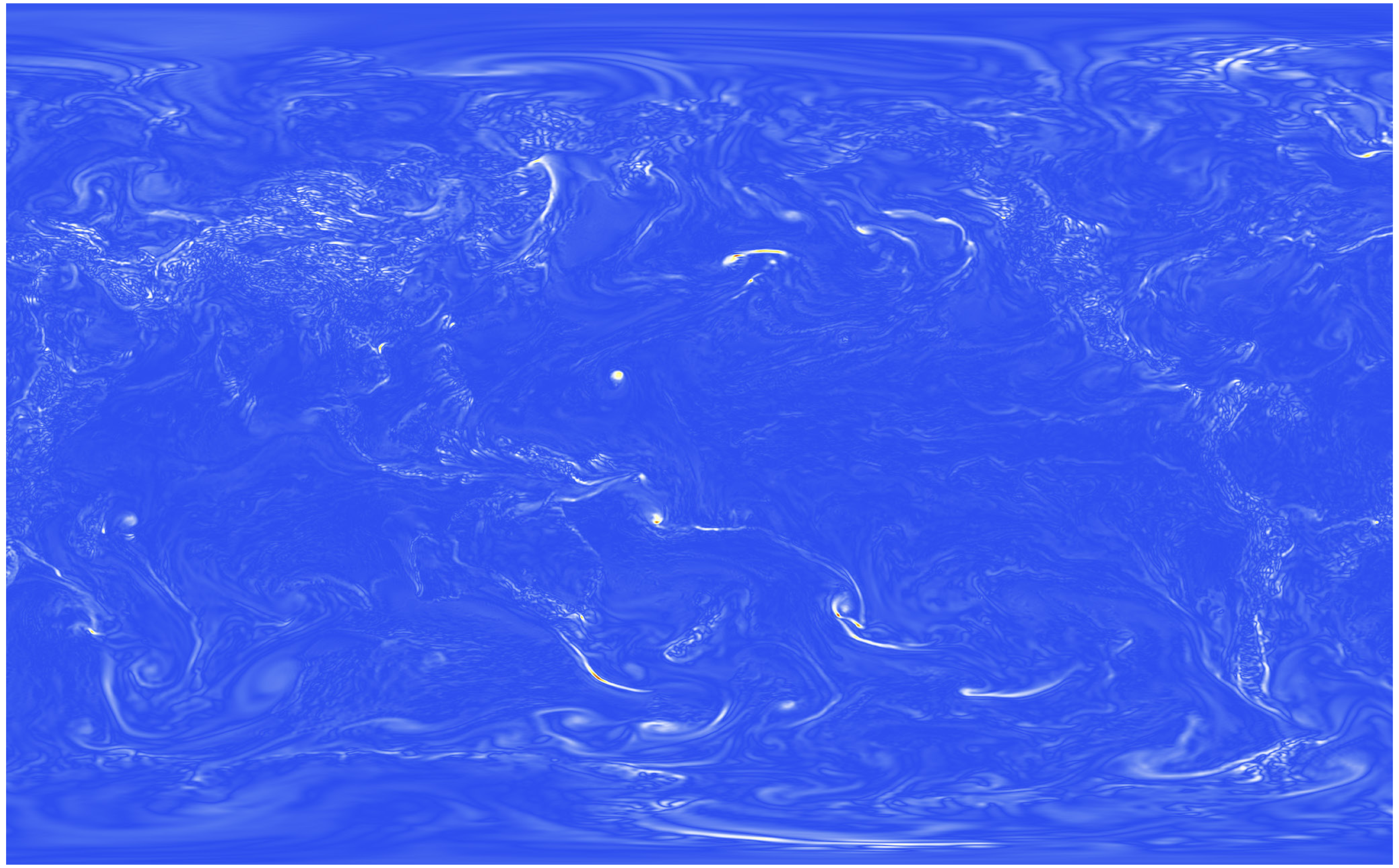
- Starting point:
  - Good distance function is critical for many machine learning applications
  - Standard ones from mathematics are often of limited utility

Motivation

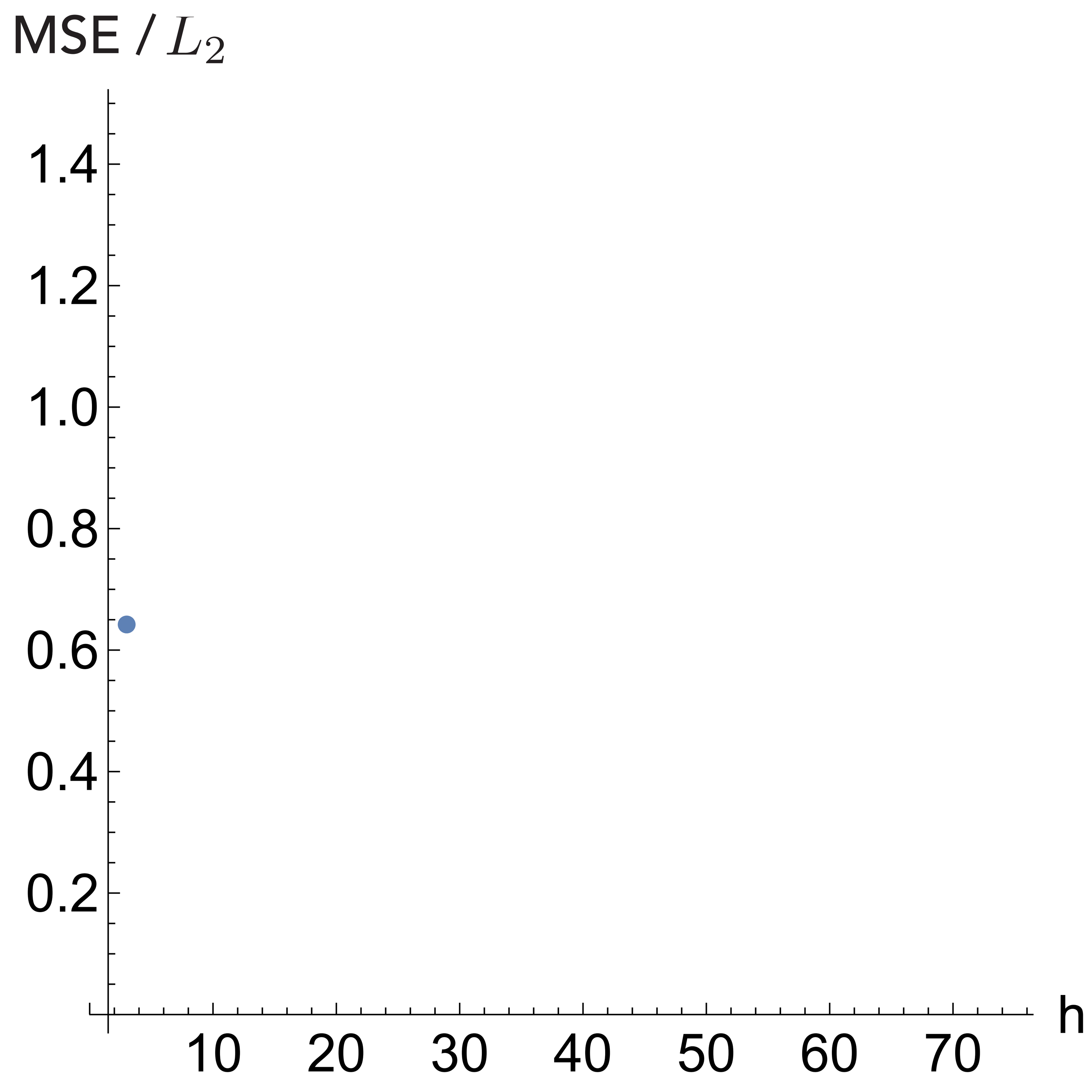


vorticity

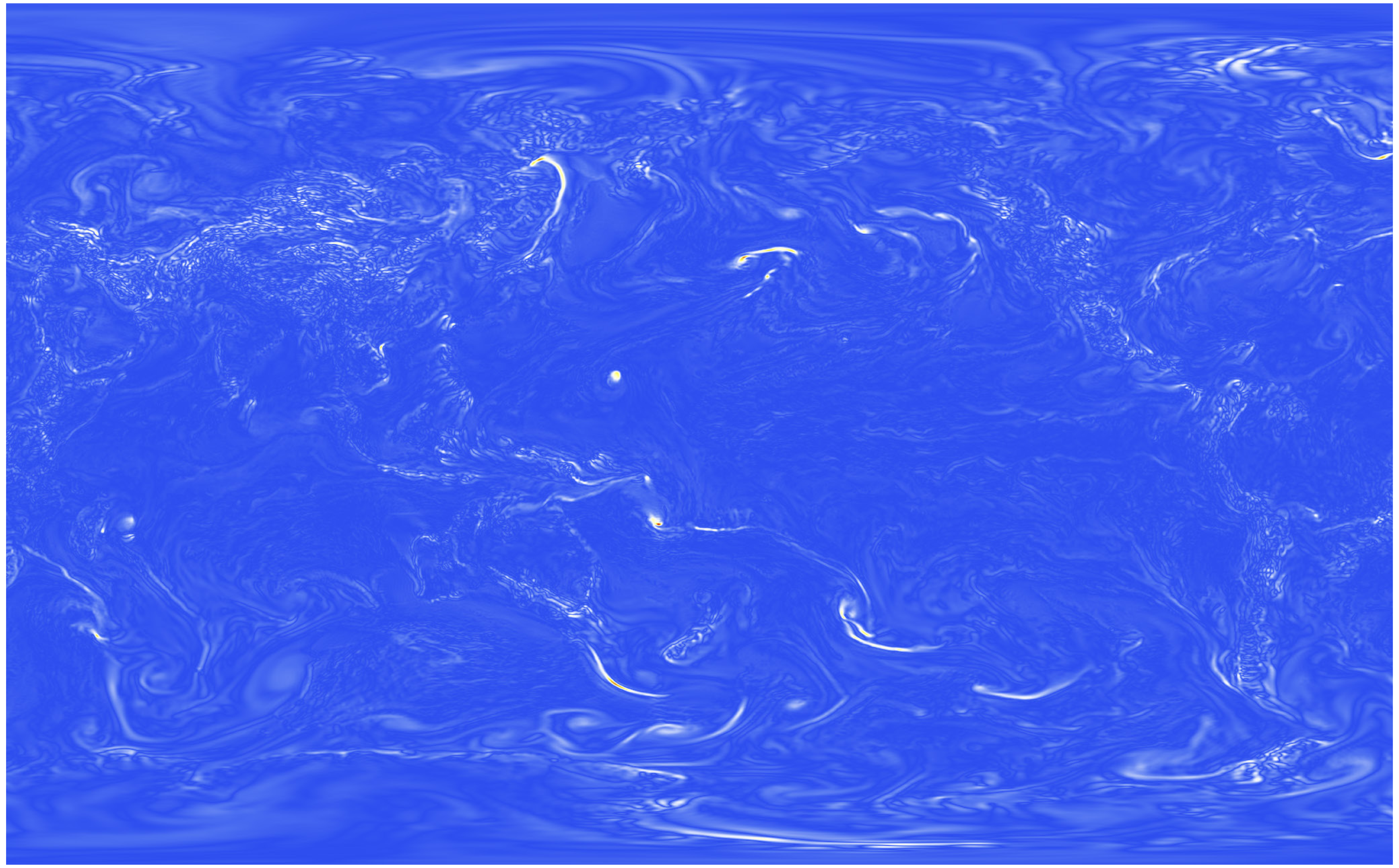
Motivation



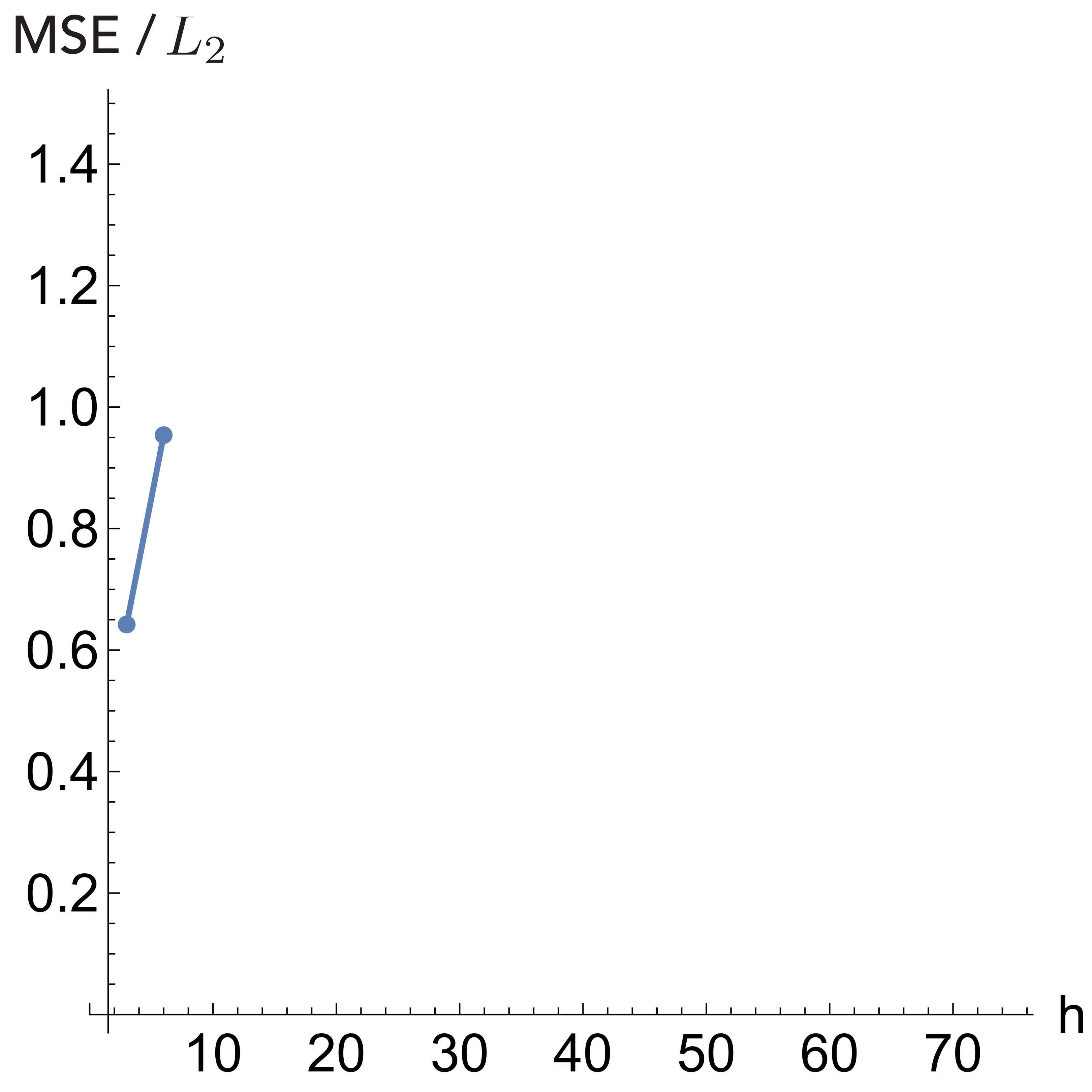
vorticity



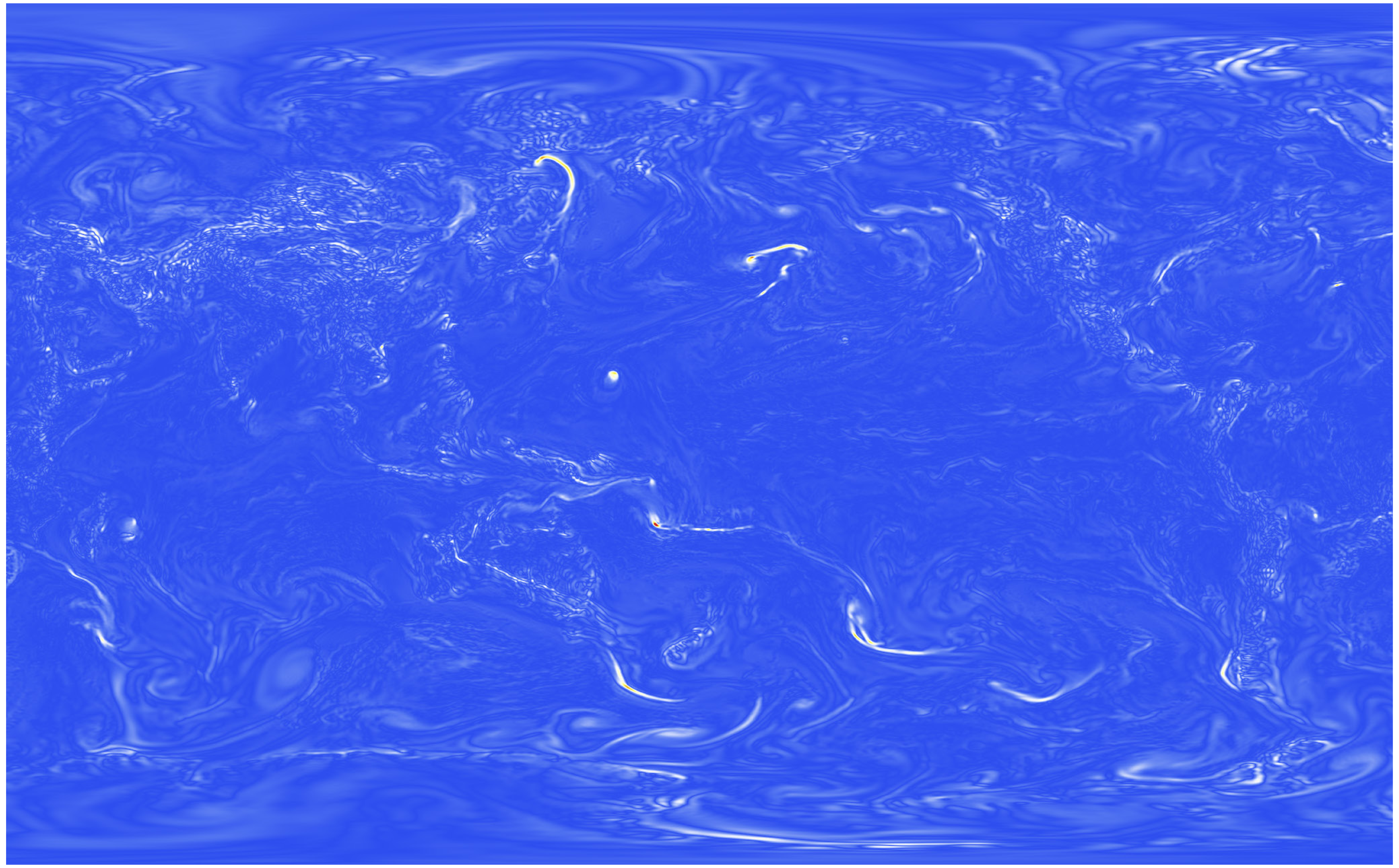
Motivation



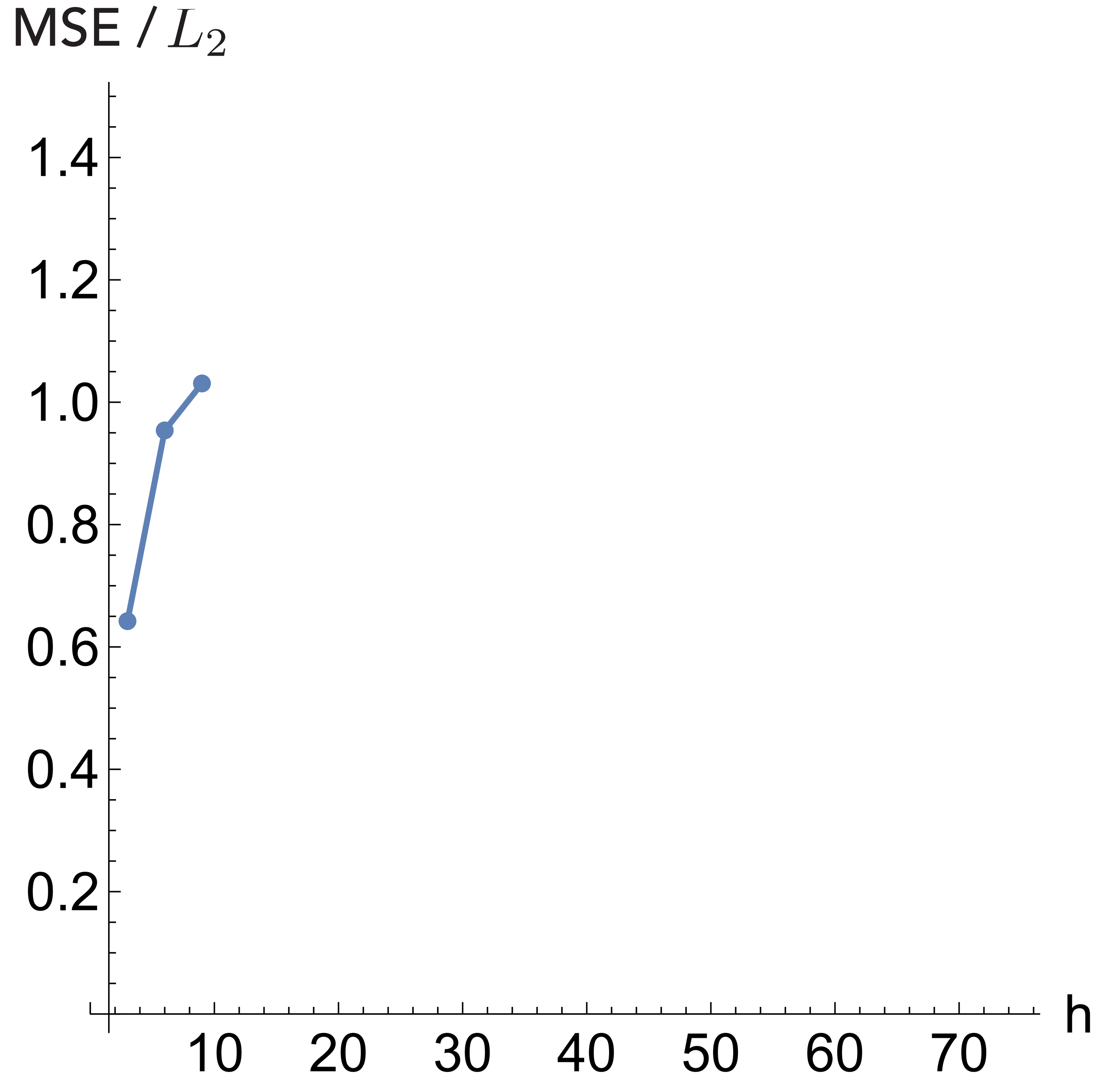
vorticity



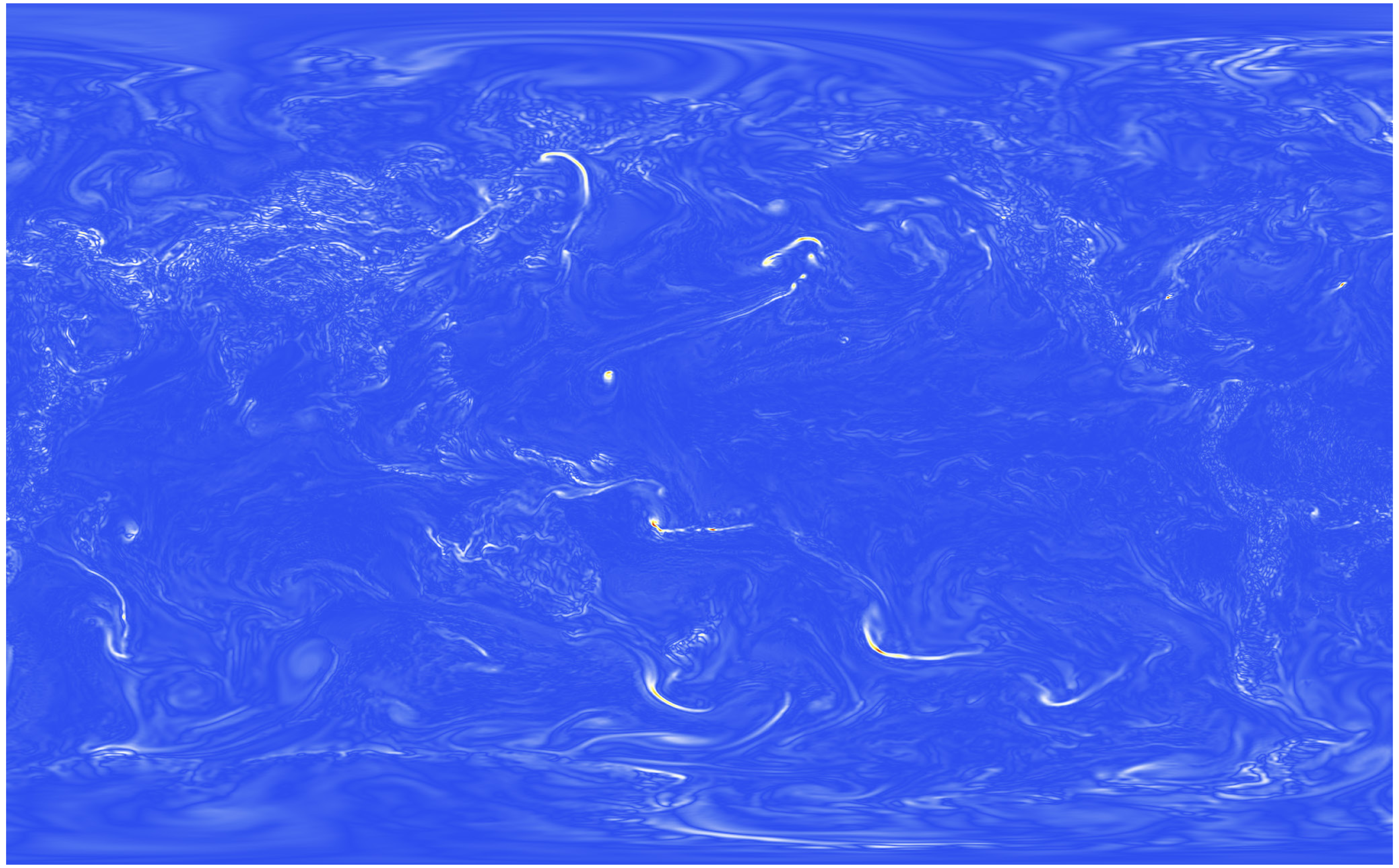
Motivation



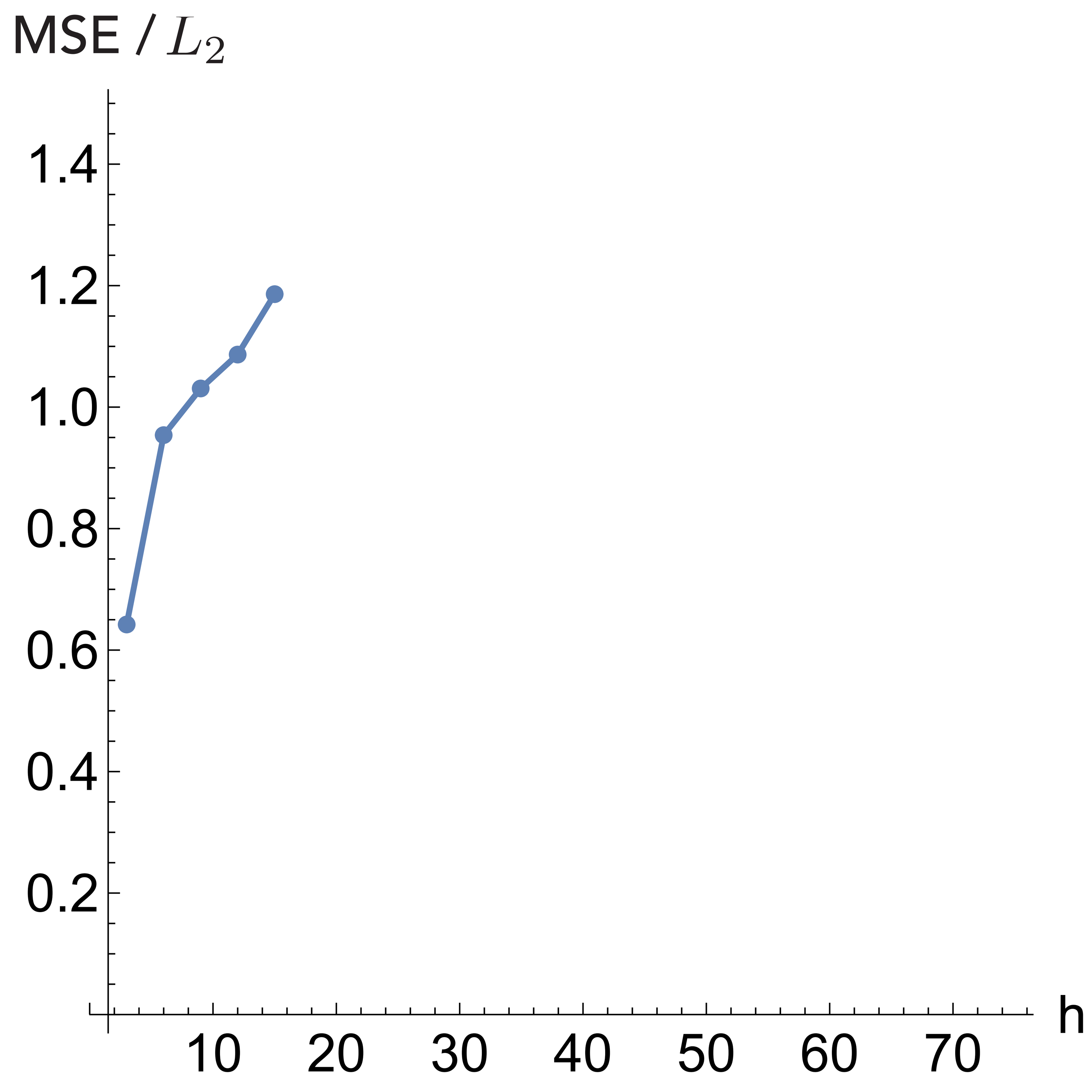
vorticity



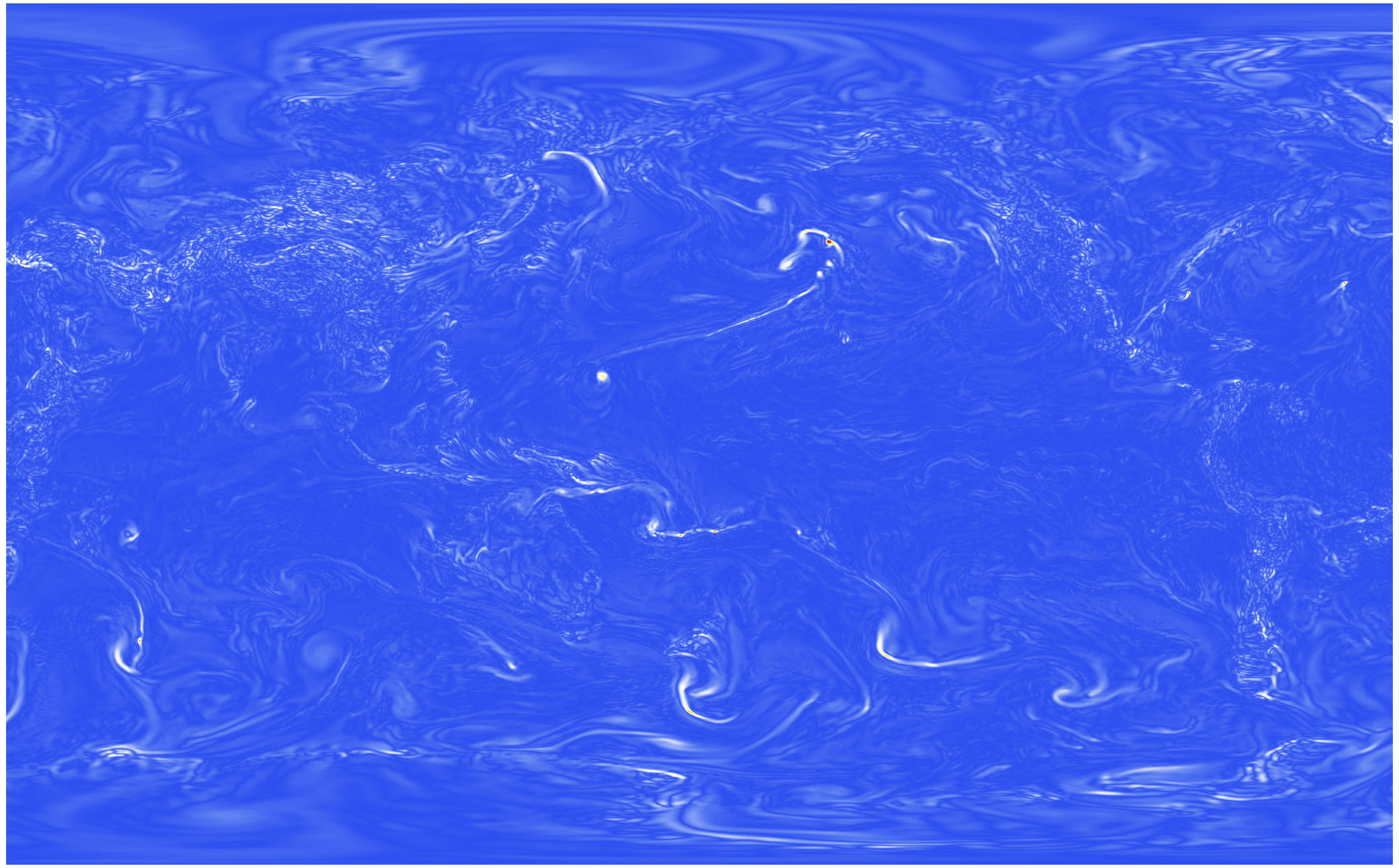
Motivation



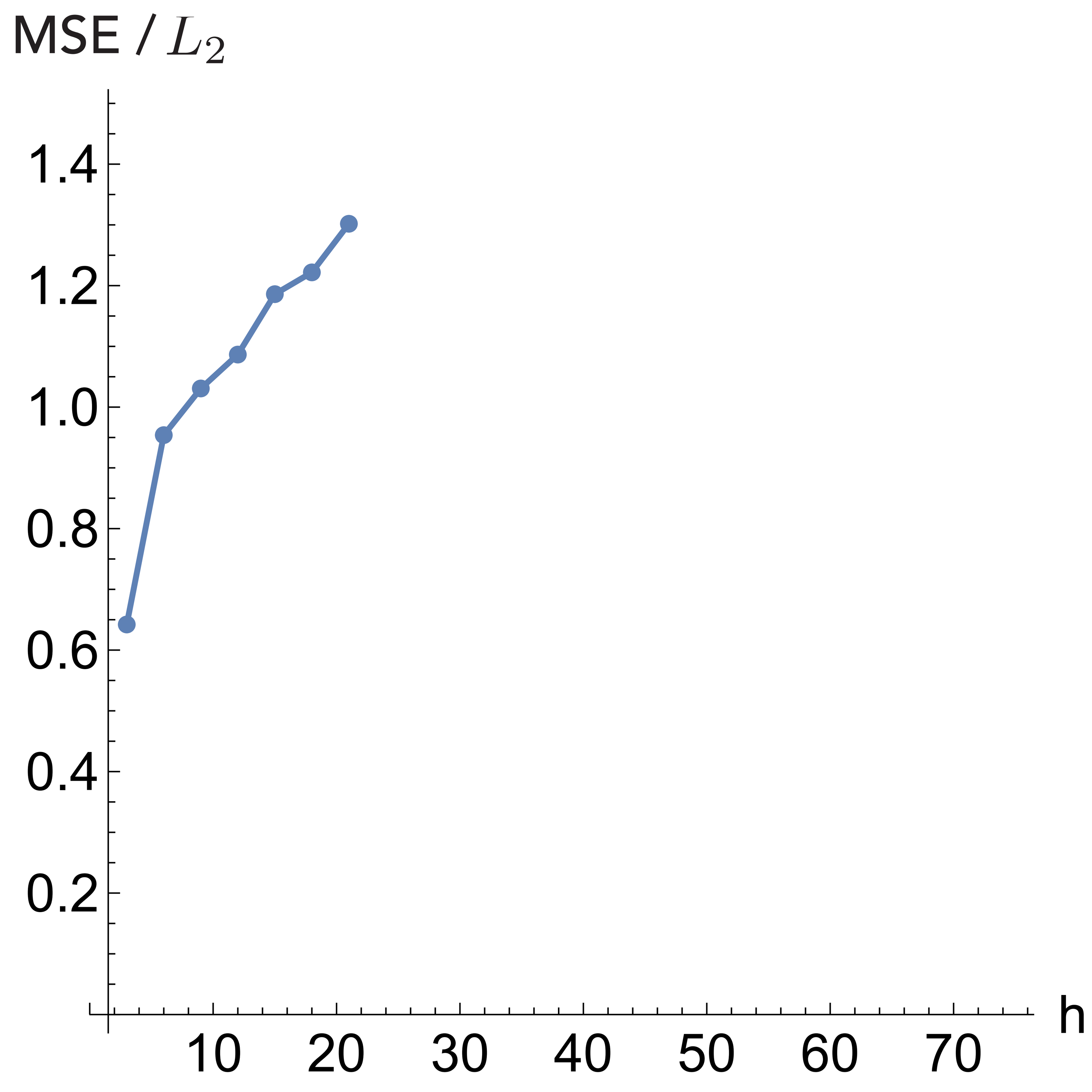
vorticity



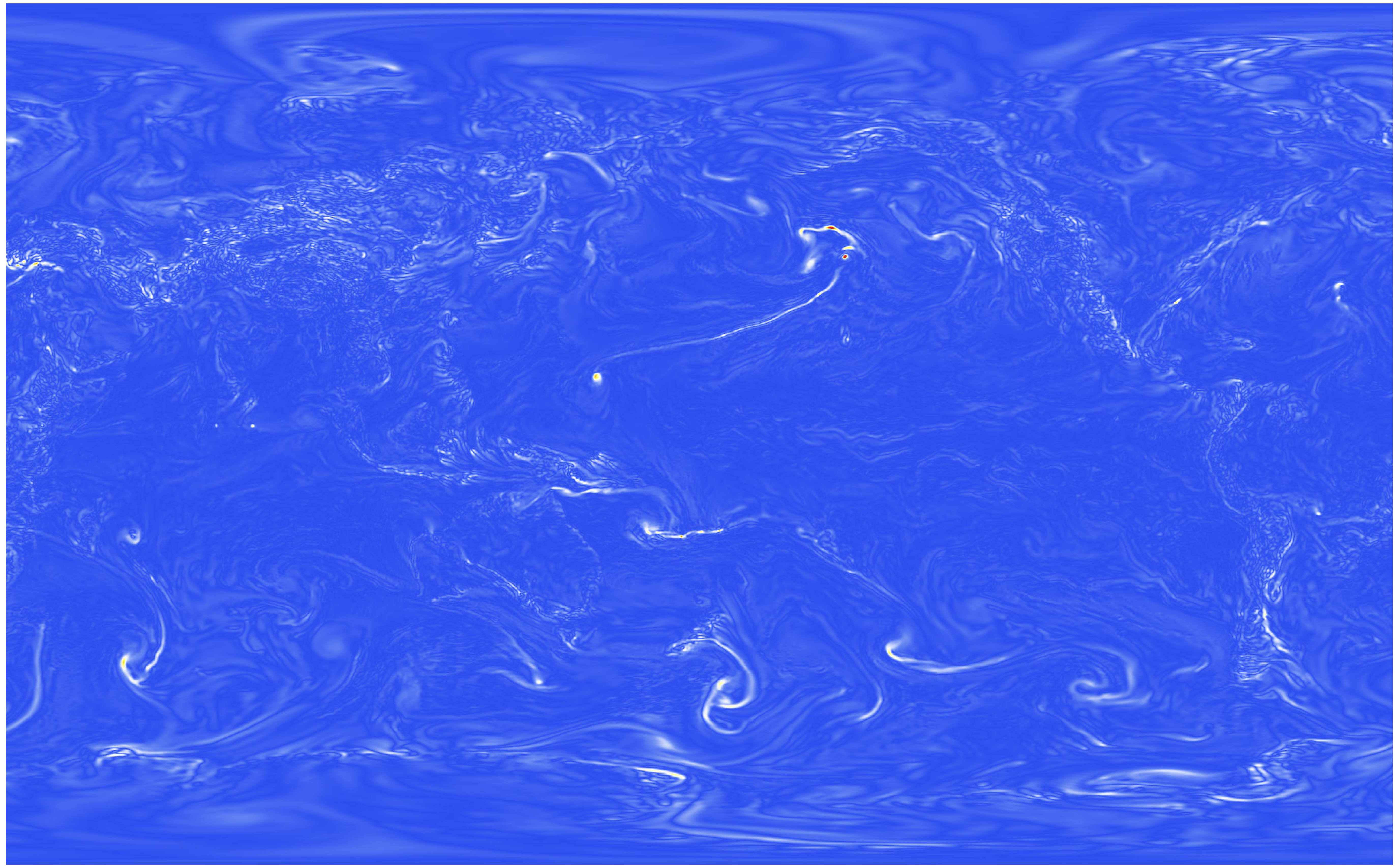
Motivation



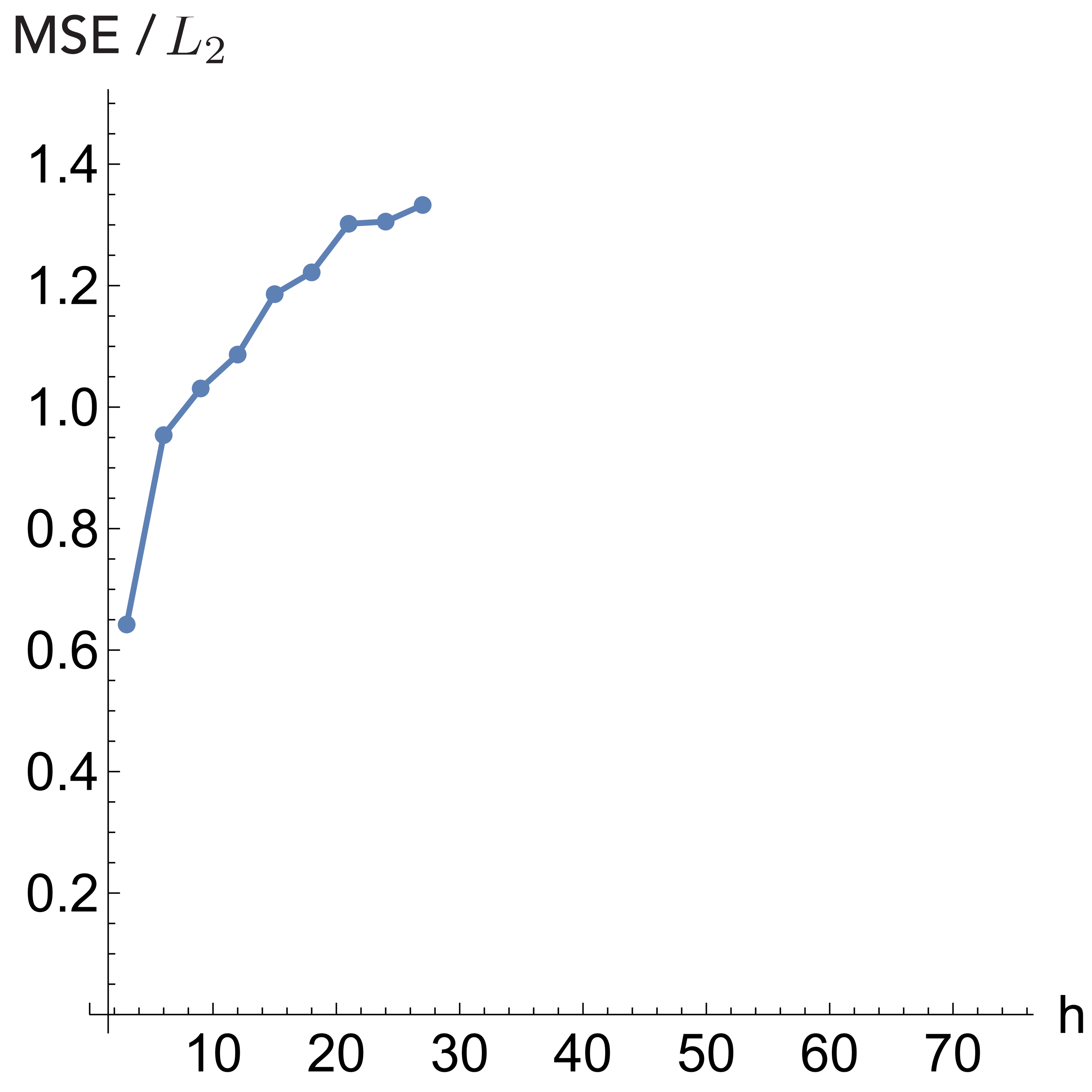
vorticity



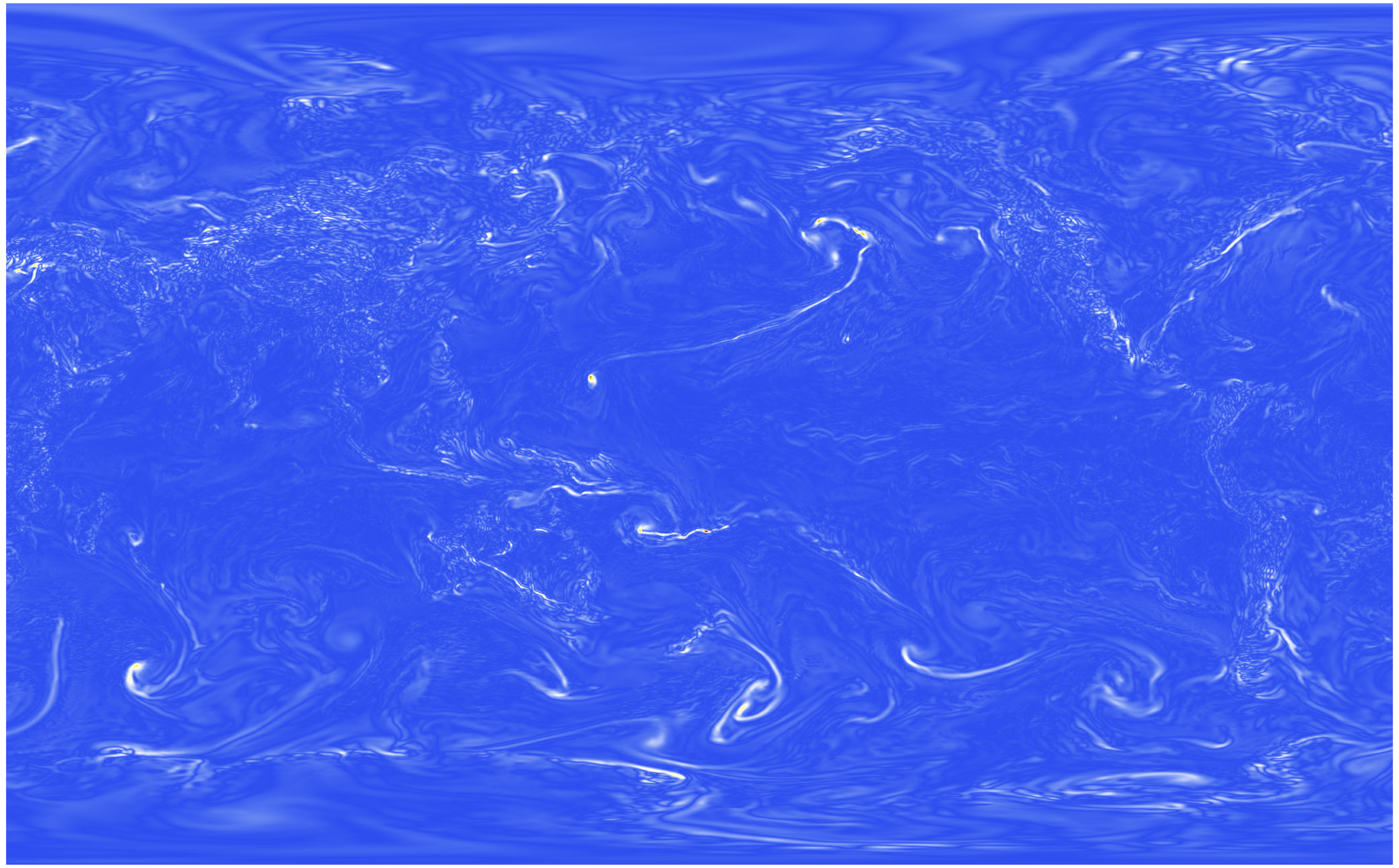
Motivation



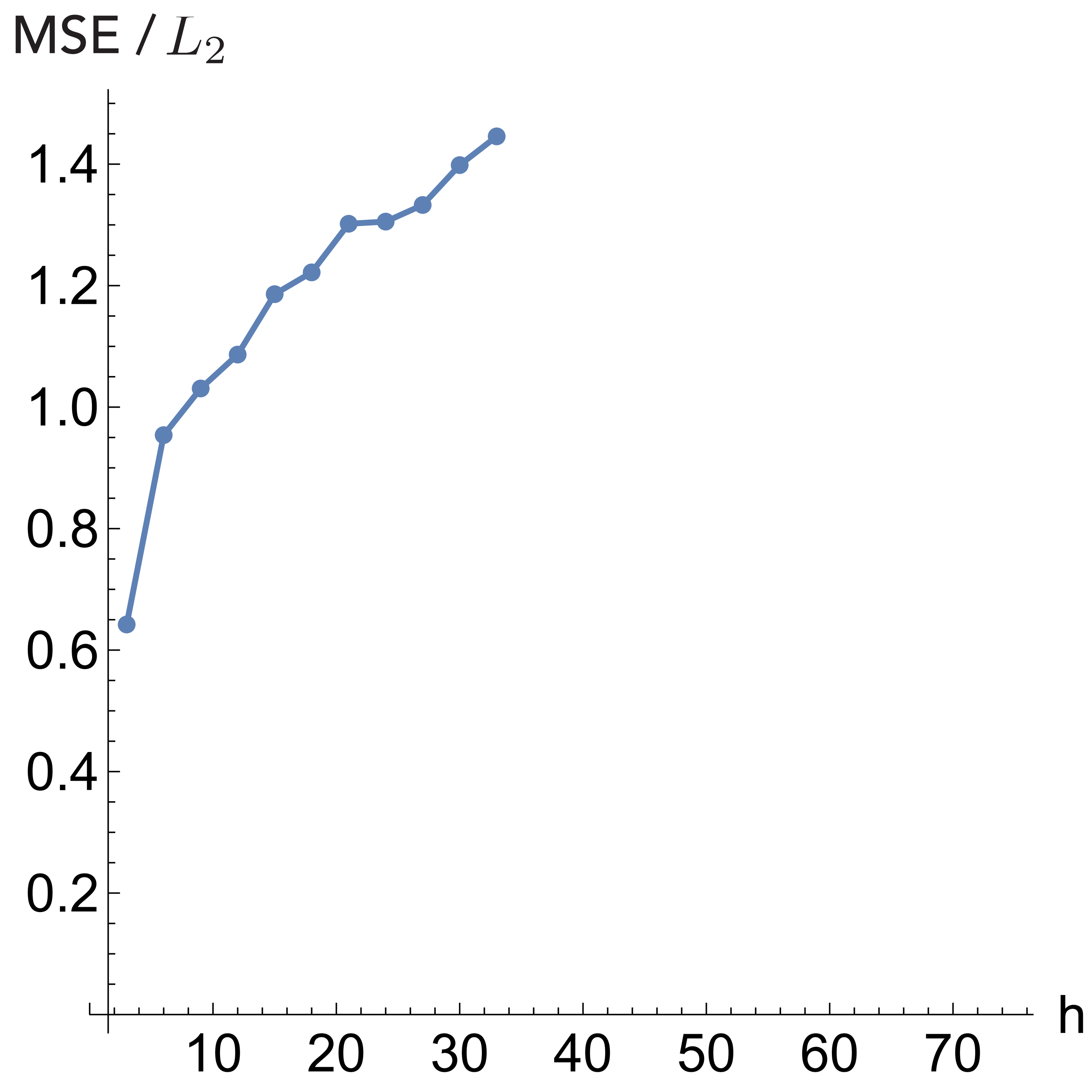
vorticity



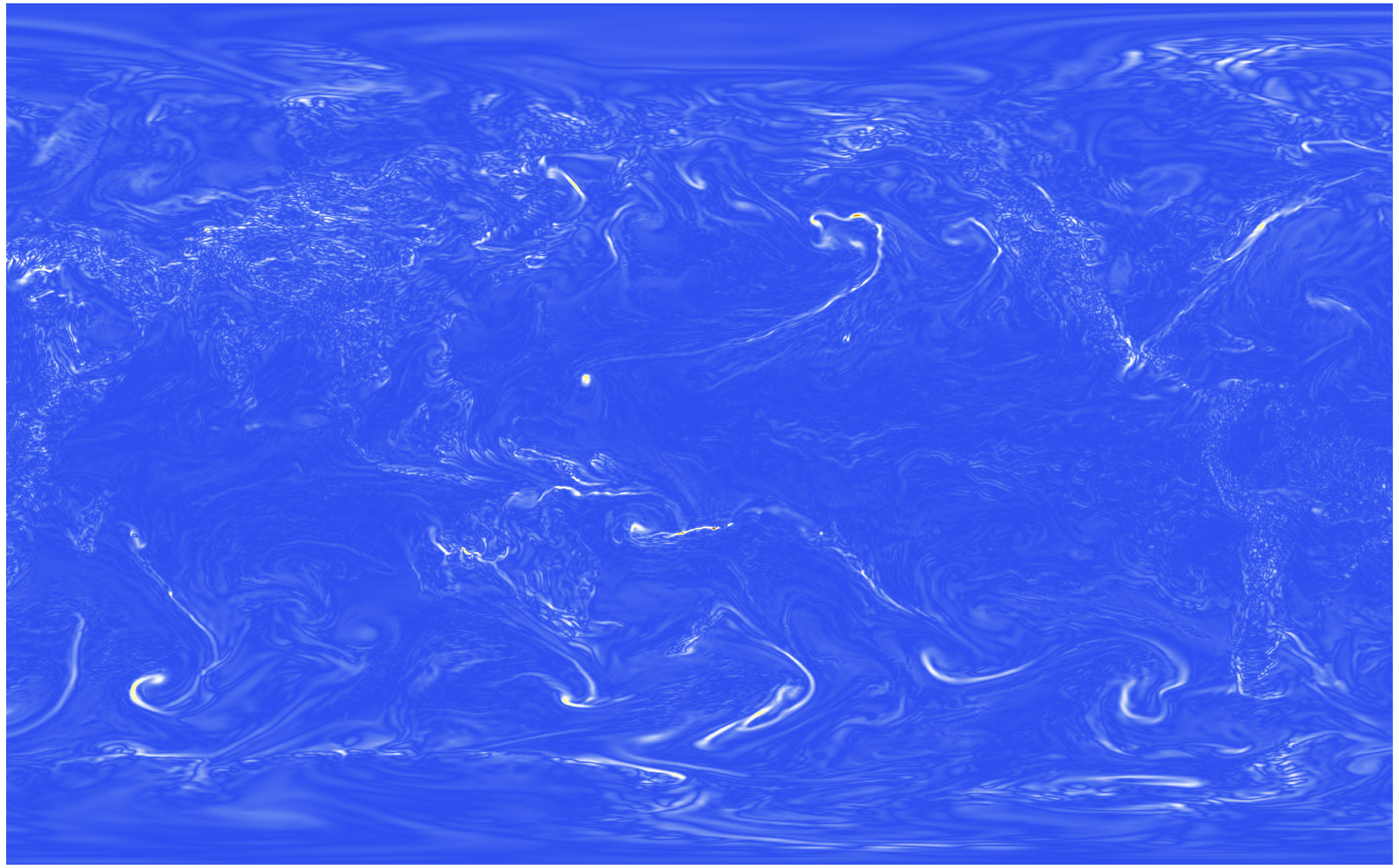
Motivation



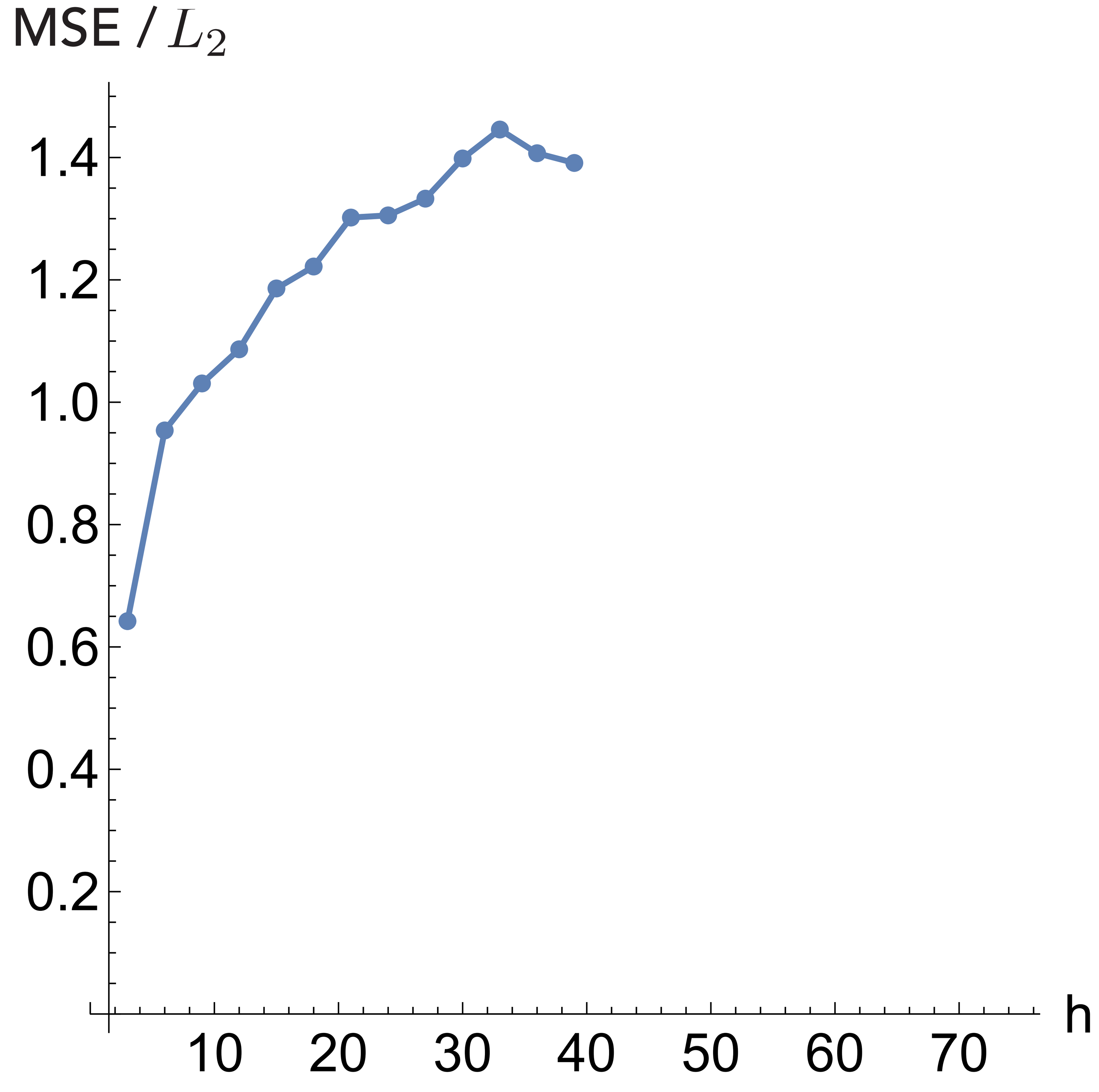
vorticity



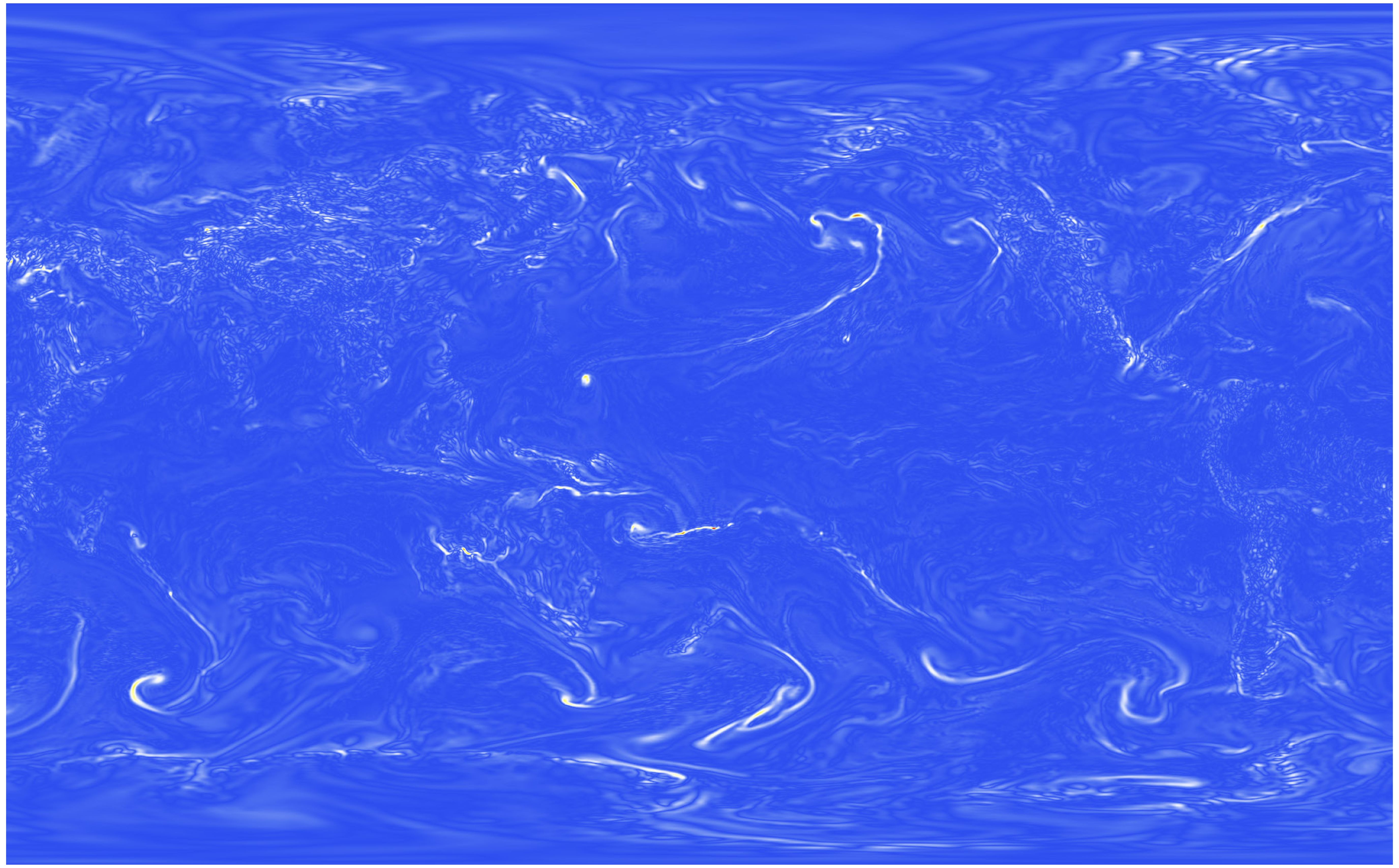
Motivation



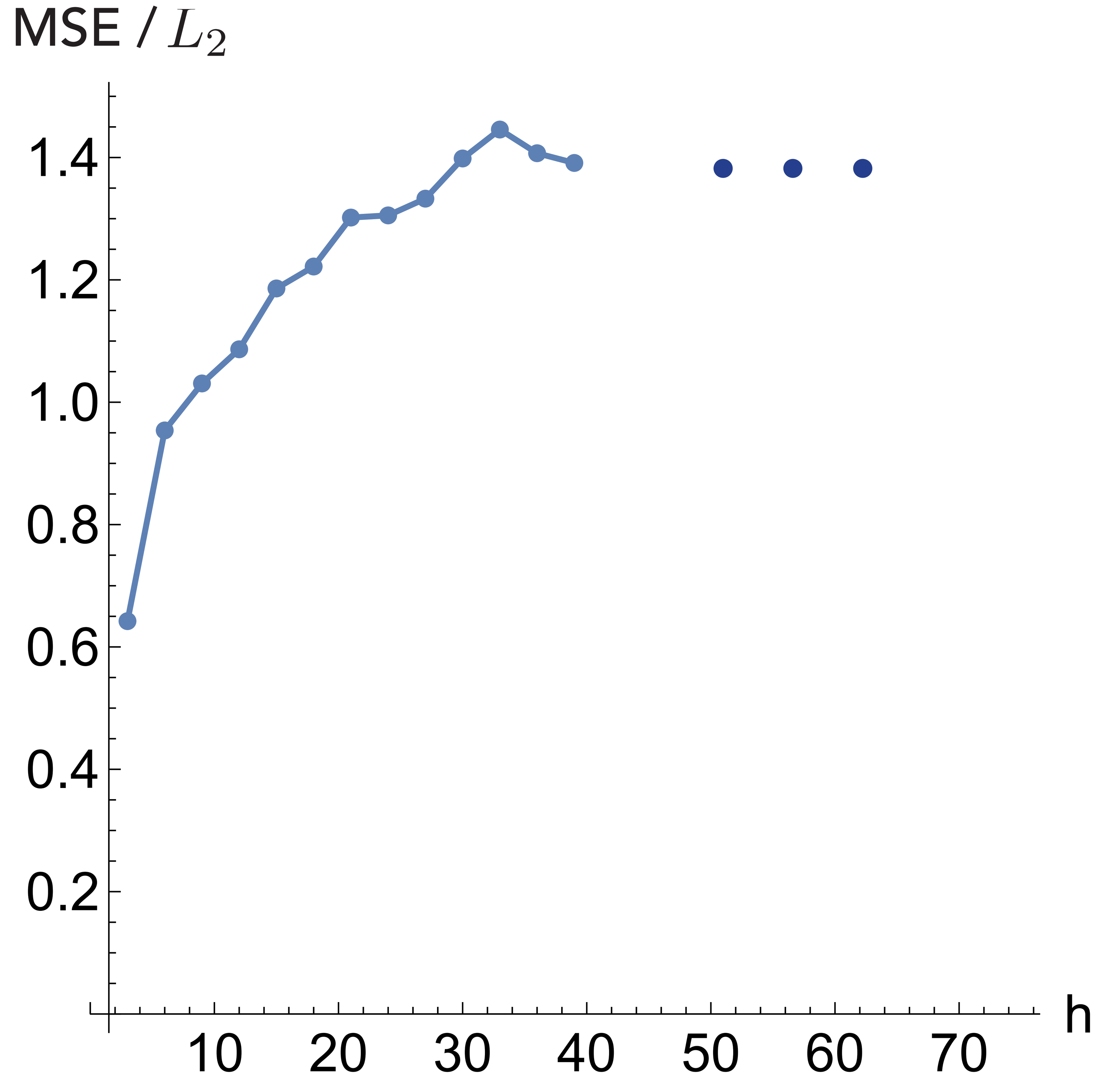
vorticity



Motivation



vorticity

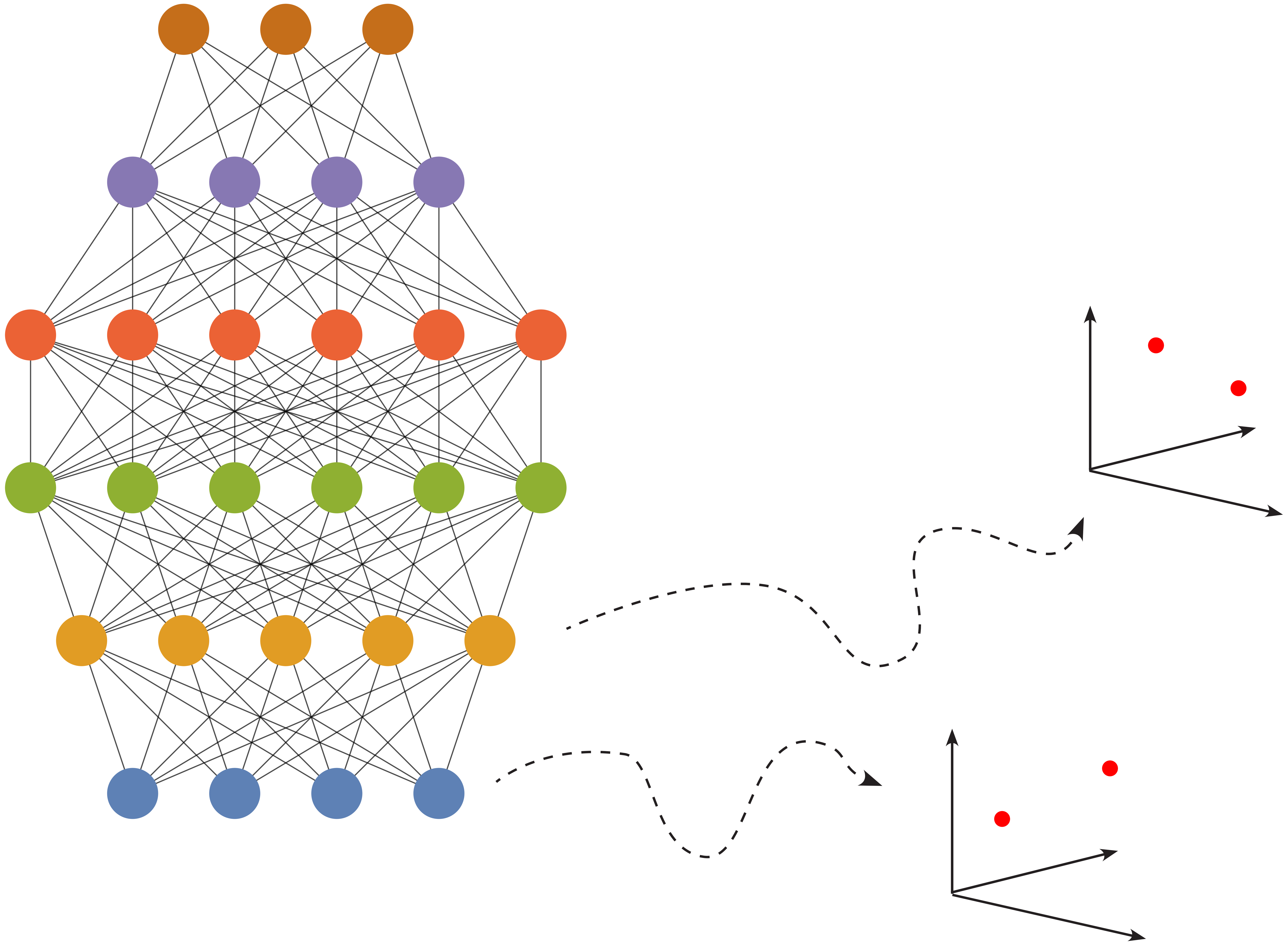


AtmoDist[1]

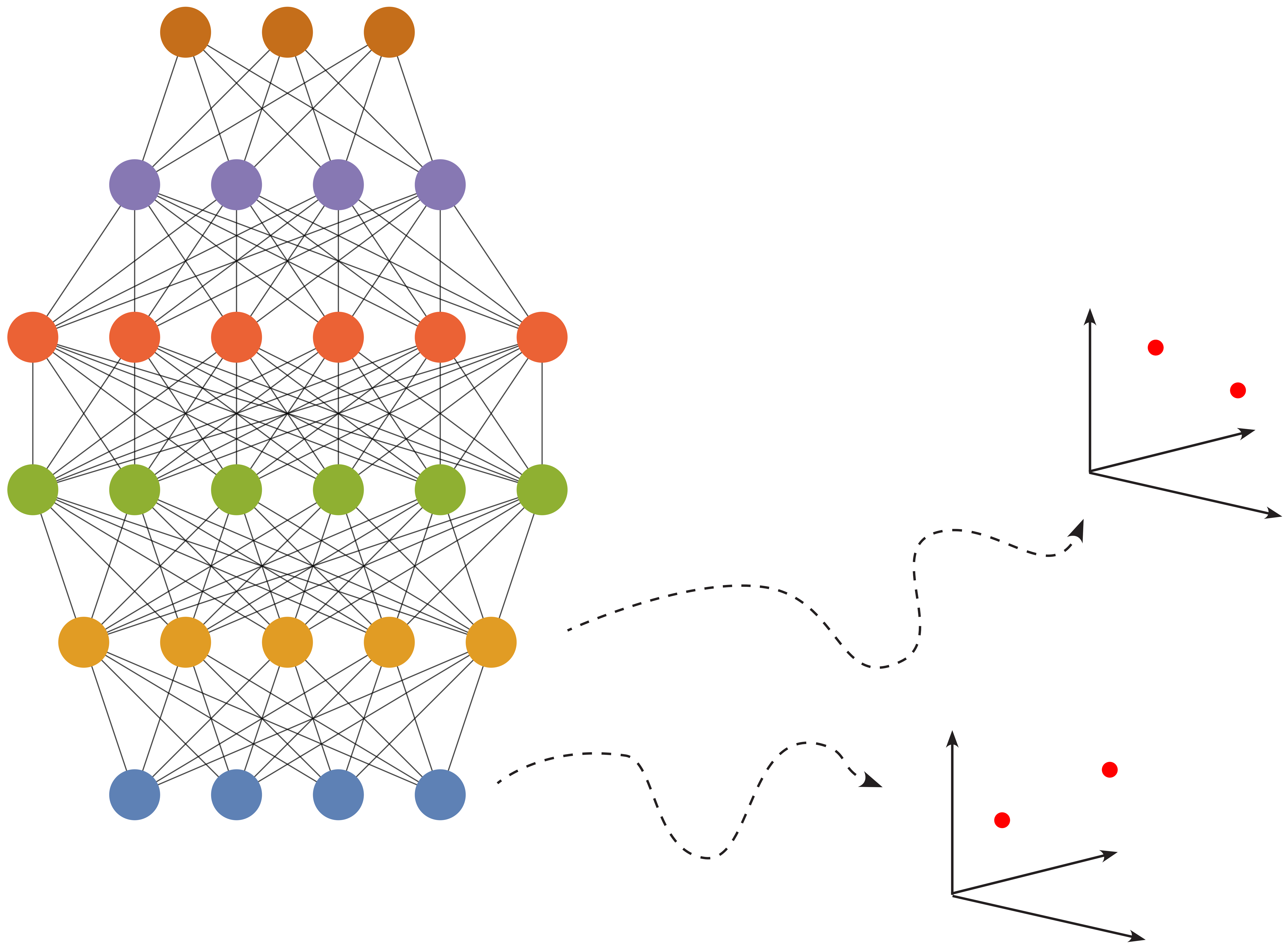
- Custom distance metric for vorticity + divergence (wind velocity vector field)
- Application: GAN-based super-resolution/downscaling
  - Recent work by Stengel et al.[2] as baseline

[1] S. Hoffmann and C. Lessig. Towards representation learning for atmospheric data. In NEURIPS 2021 Workshop on Climate Change (poster), 2021.
[2] K. Stengel, A. Glaws, D. Hettinger, and R. N. King. Adversarial super-resolution of climatological wind and solar data. Proceedings of the National Academy of Sciences, 117(29):16805–16815, 2020.

AtmoDist

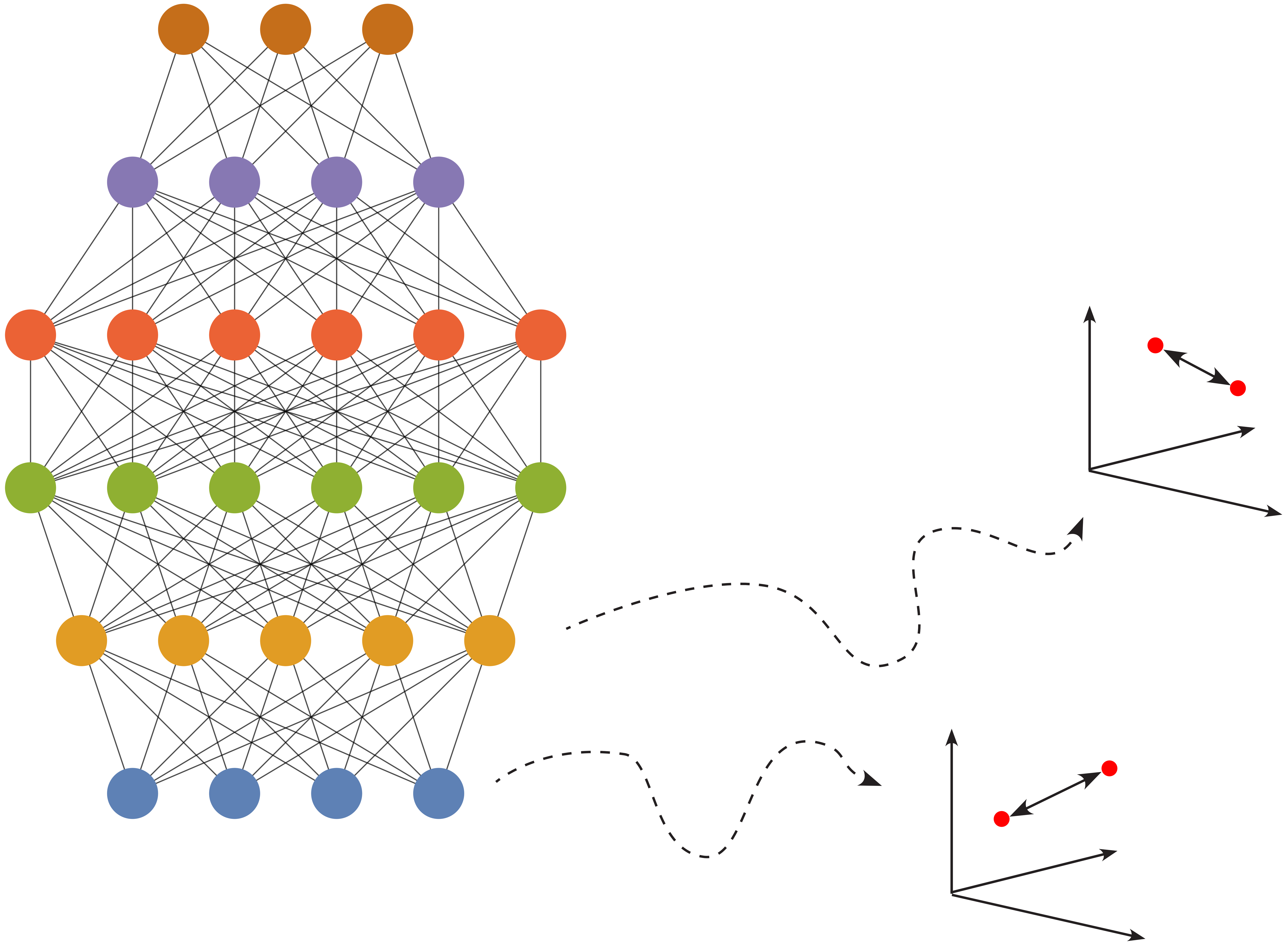


AtmoDist



Feature space is task / domain specific

AtmoDist



Feature space is task / domain specific

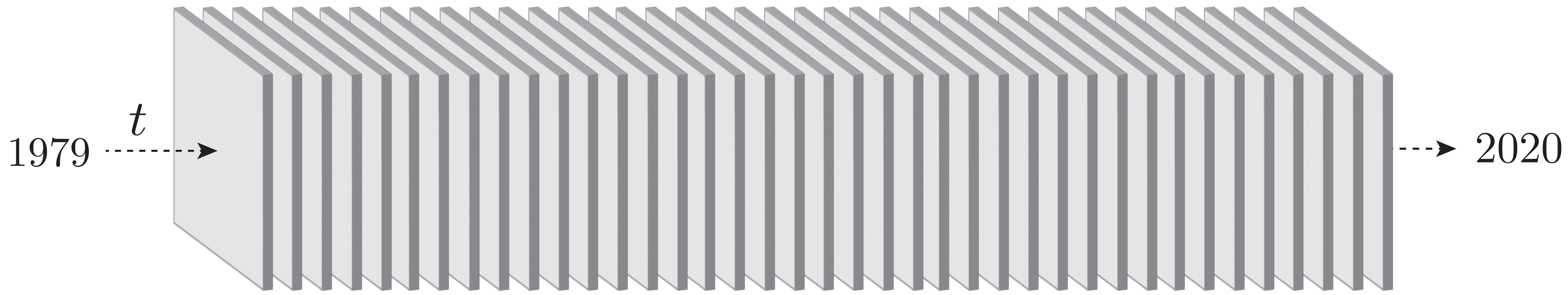
Compute distance there!

AtmoDist

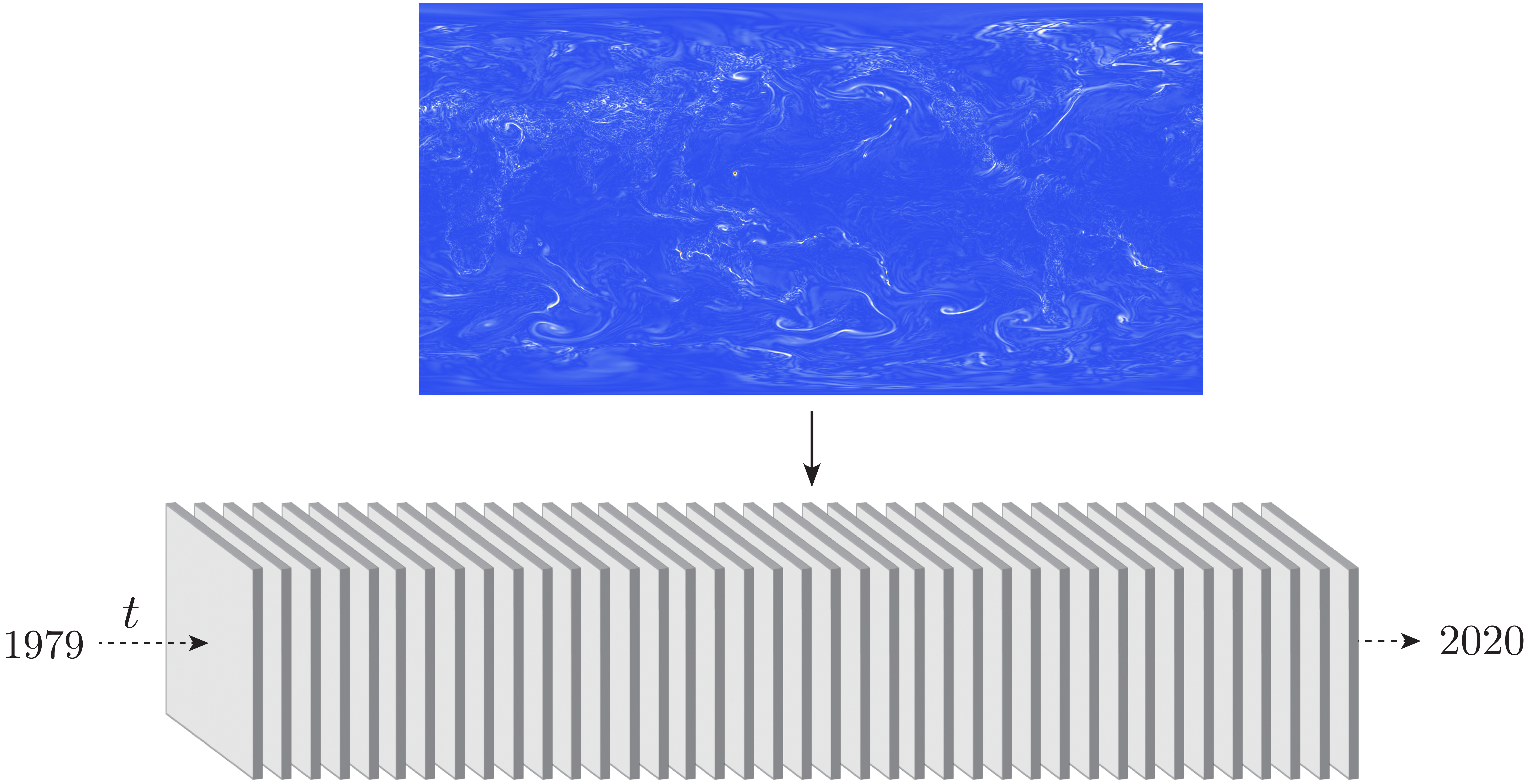
- ERA5[1] provides well curated data set of (processed) atmospheric observations for training
  - Vorticity and divergence
  - One vertical layer (≈ 883 hPa)
  - But unlabelled

[1] Hersbach et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 2020.

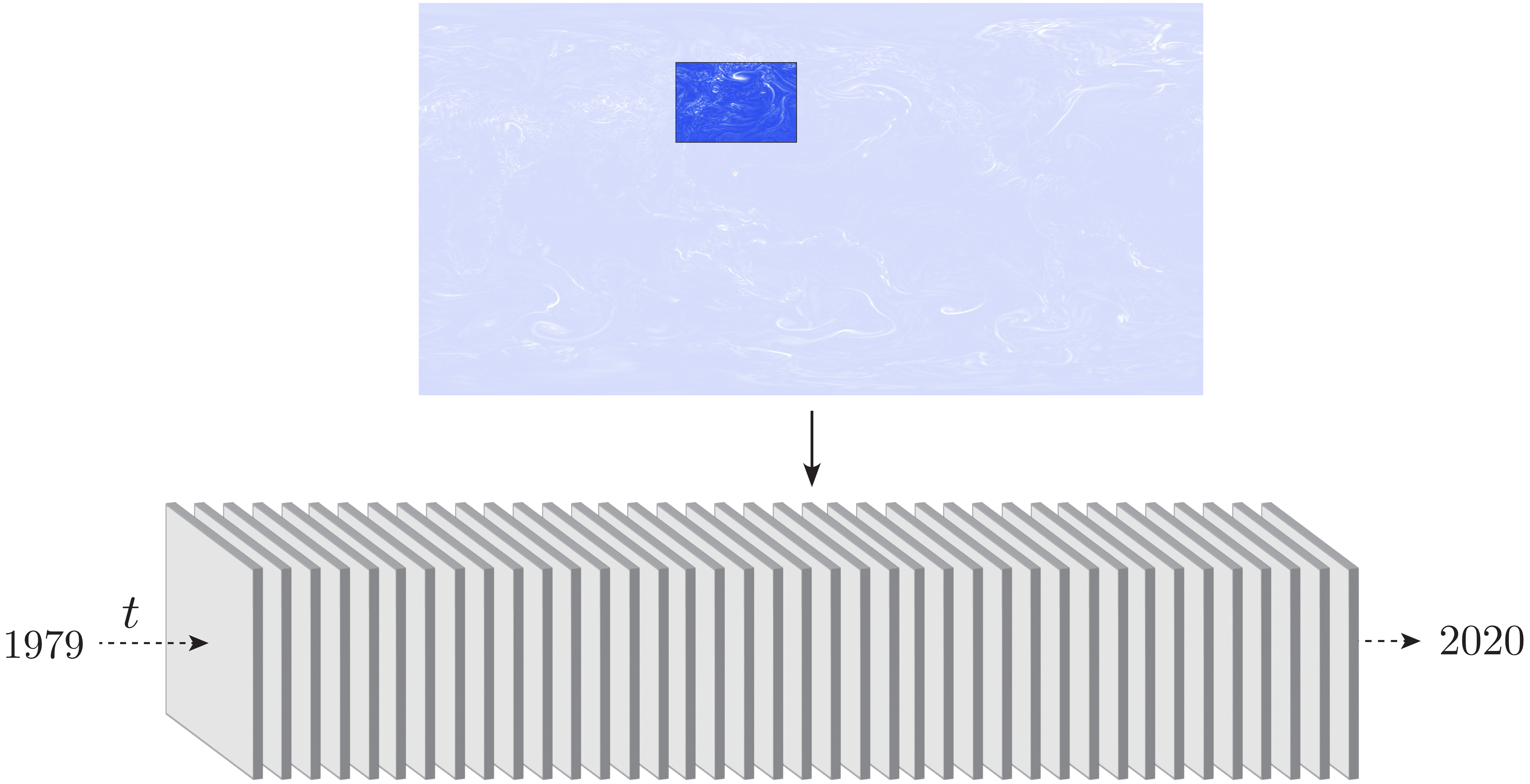
AtmoDist



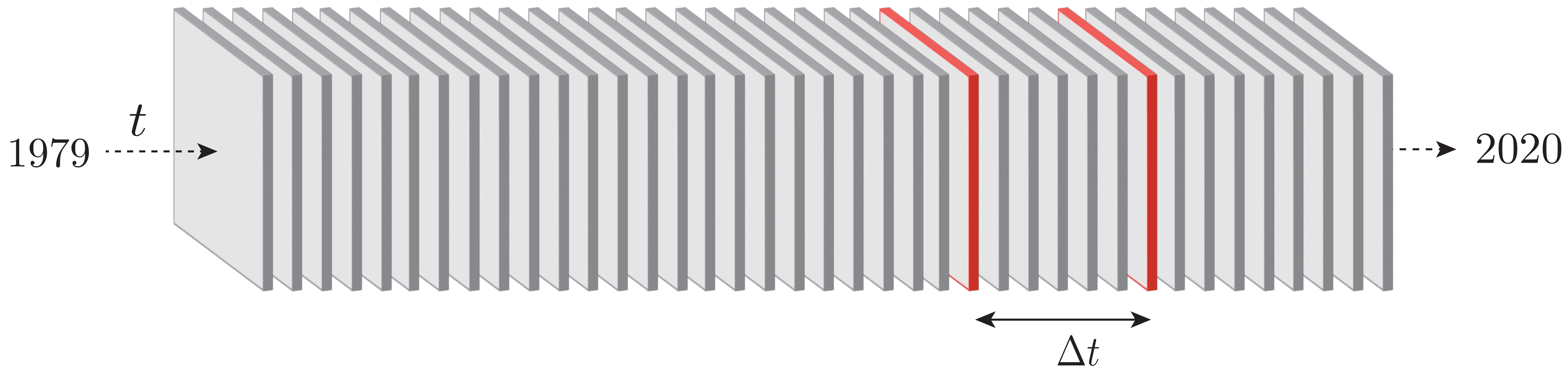
AtmoDist



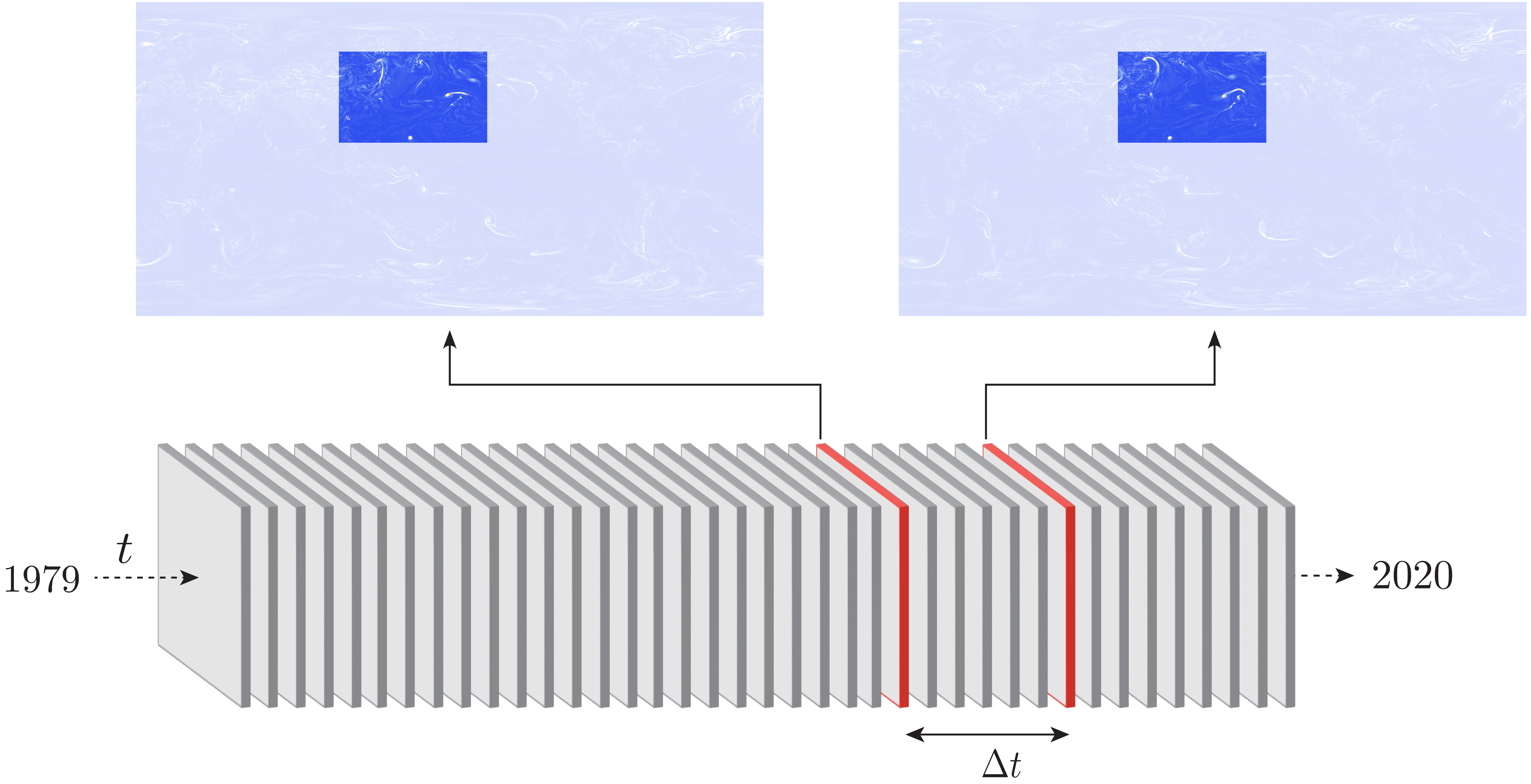
AtmoDist



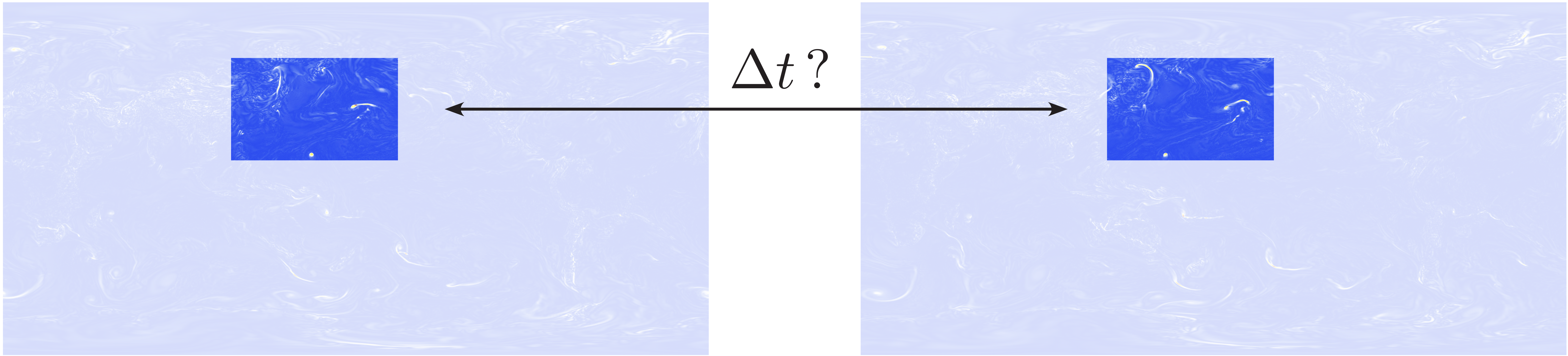
AtmoDist

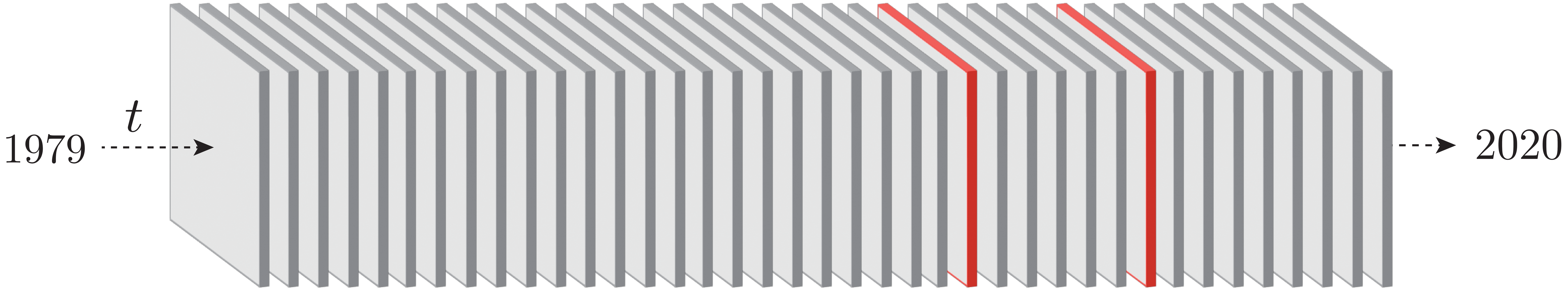


AtmoDist

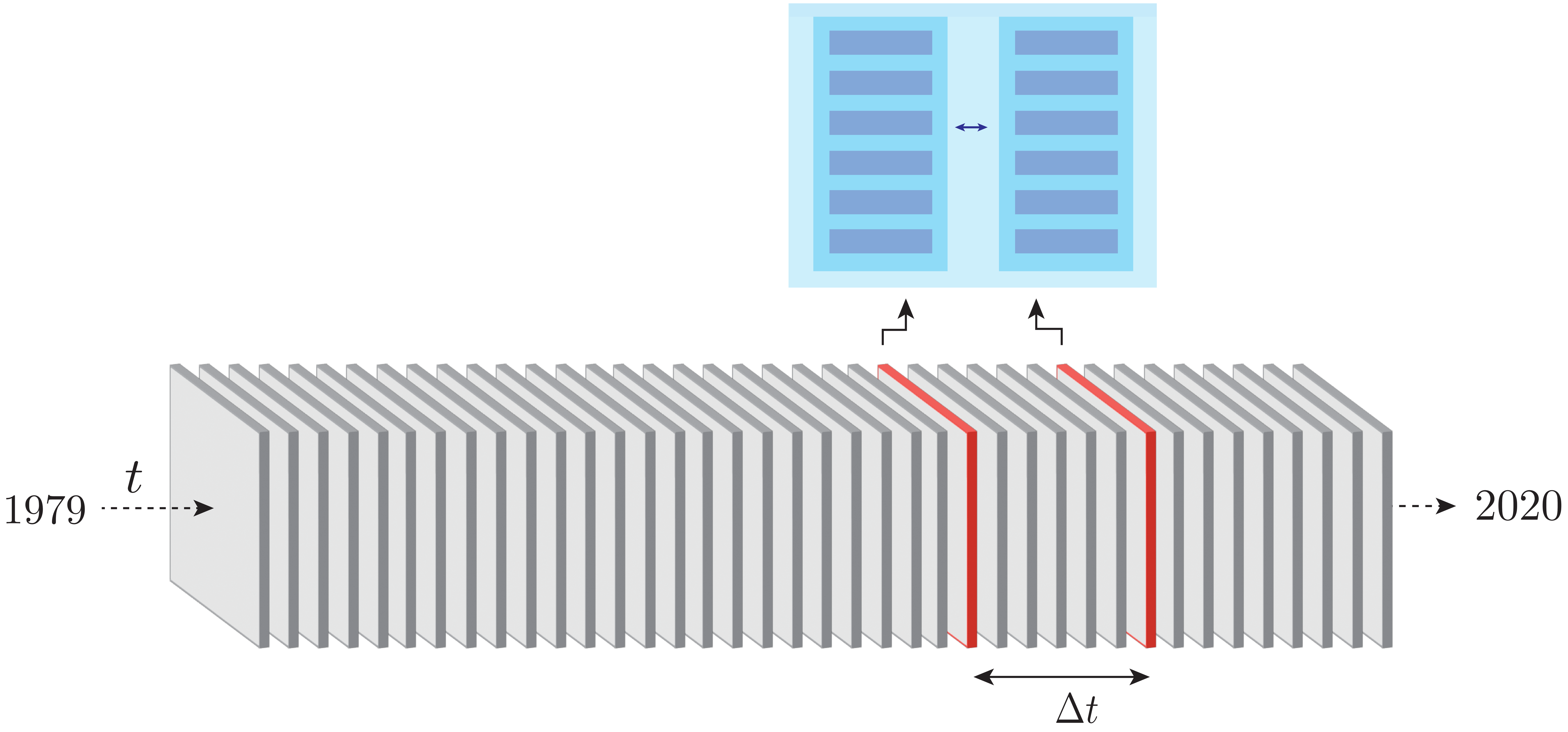


AtmoDist

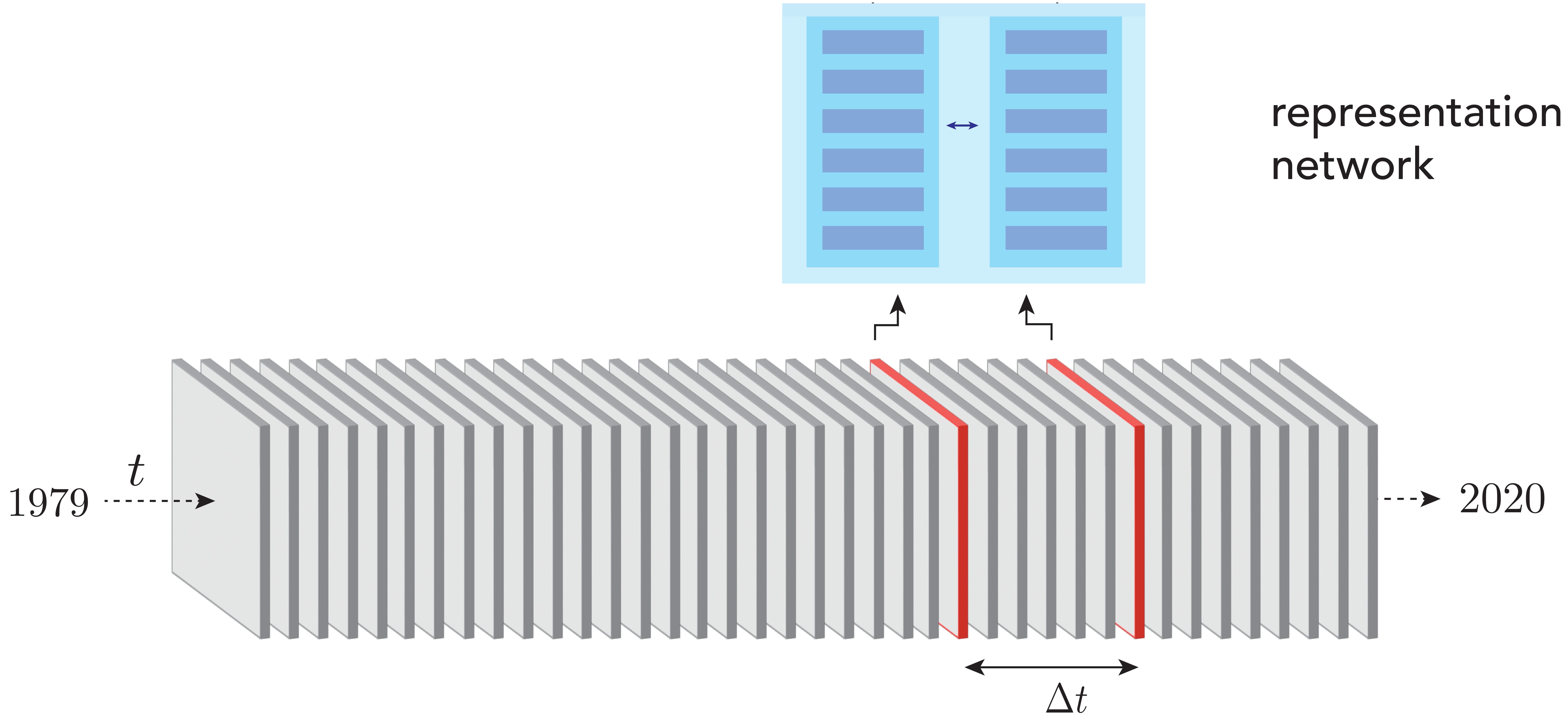




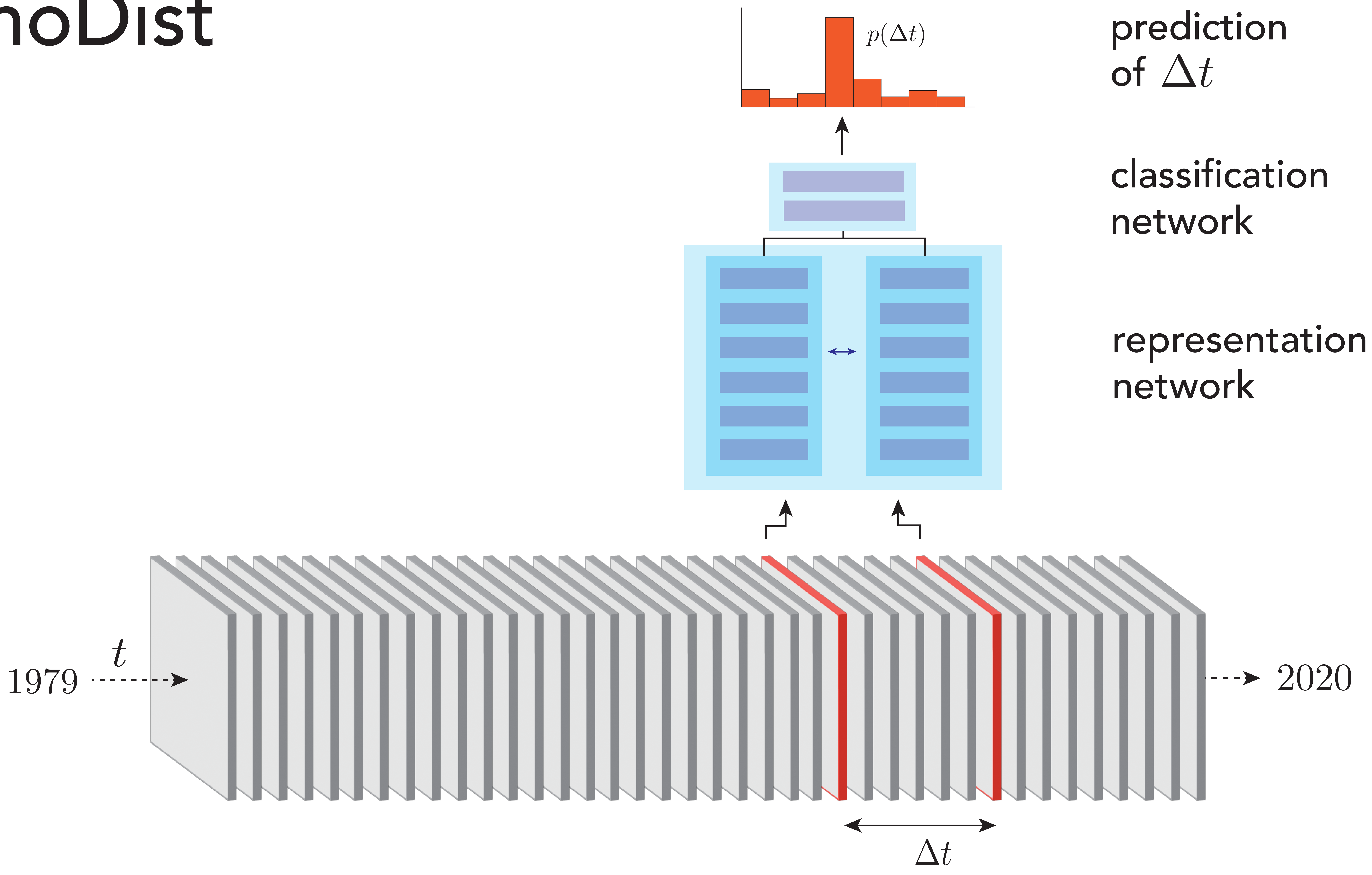
AtmoDist



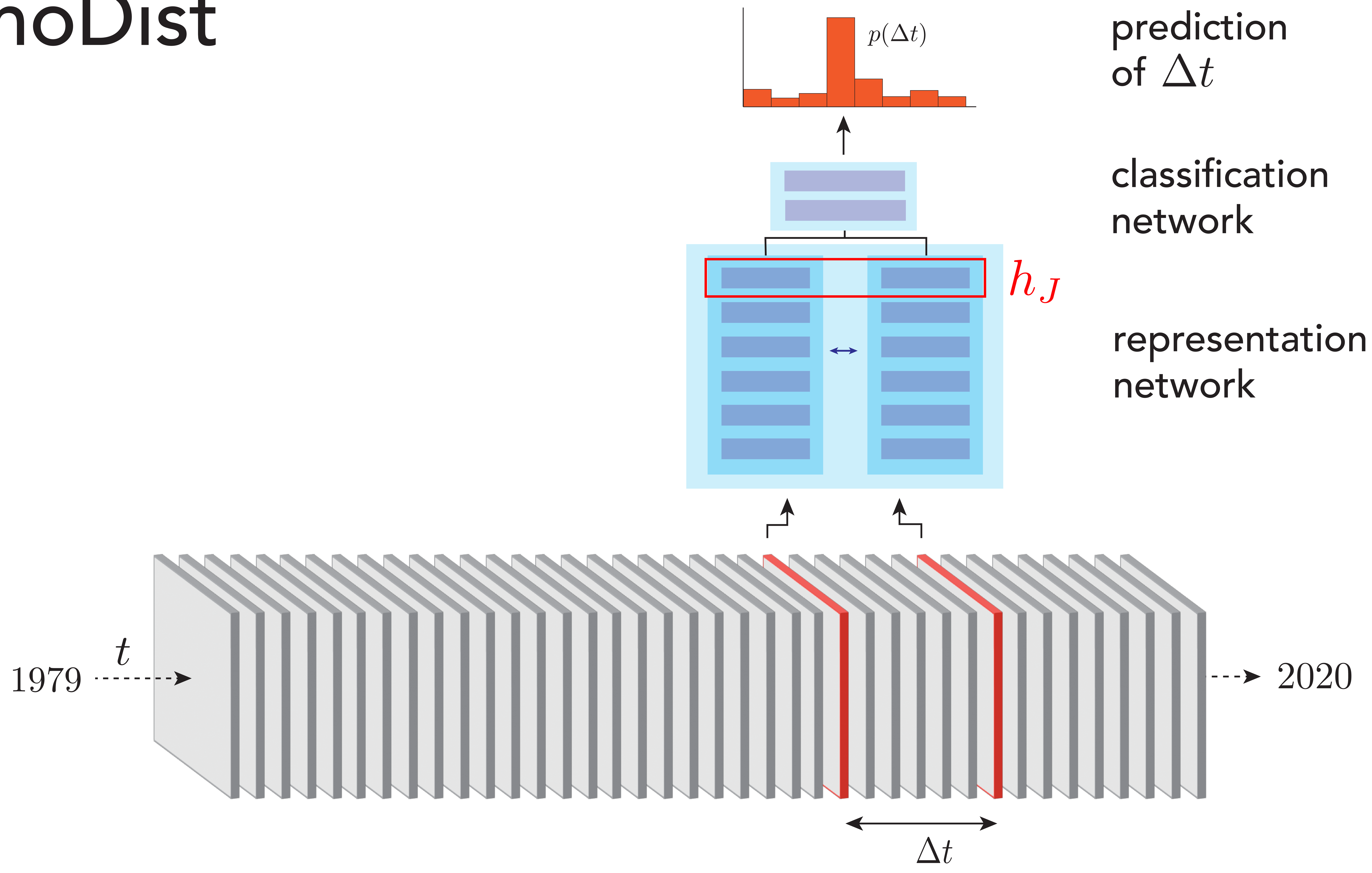
AtmoDist



AtmoDist

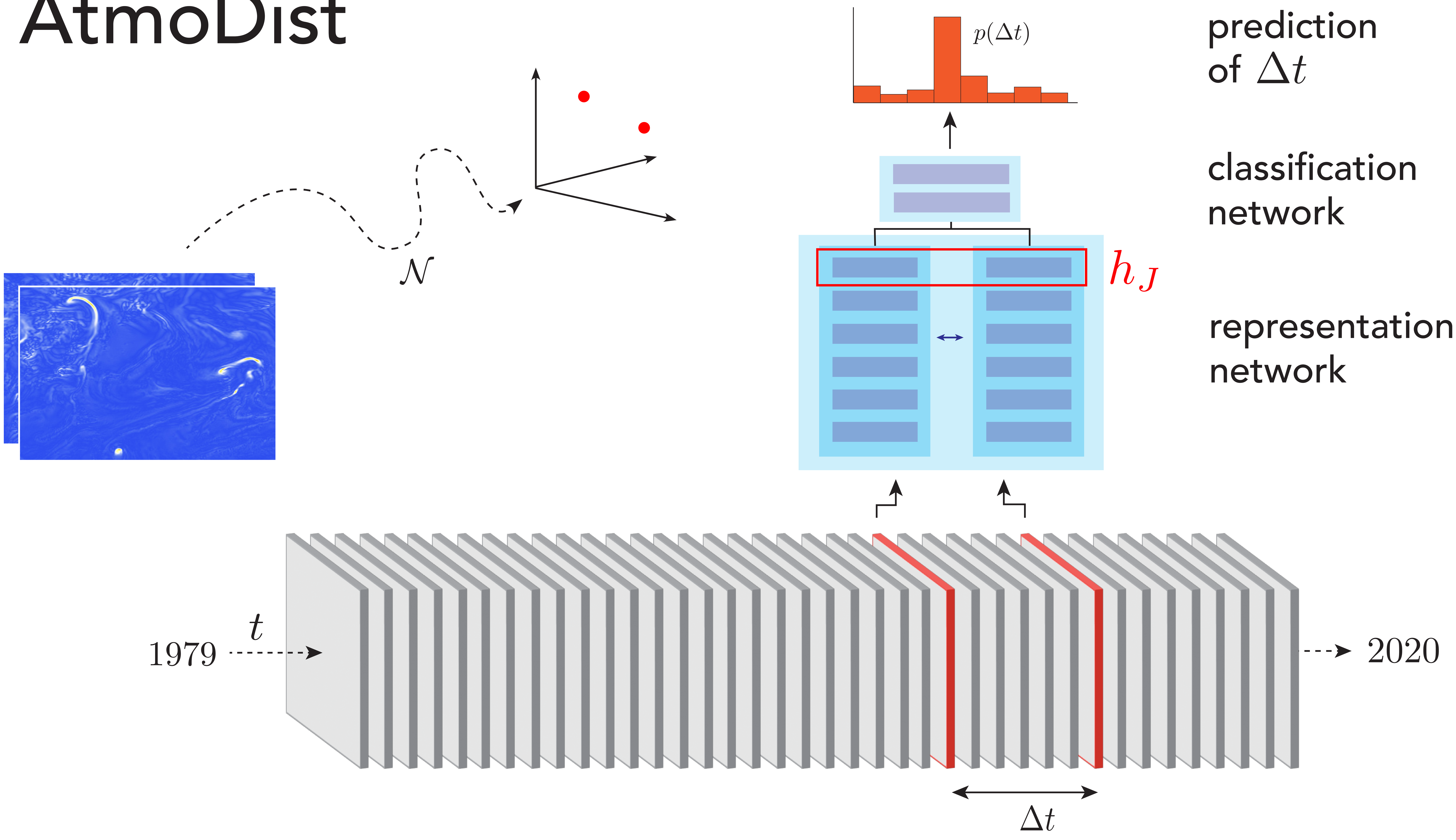

$p(\Delta t)$
prediction of Δt
classification network
representation network
1979
t
2020
Δt

AtmoDist



$p(\Delta t)$
prediction of Δt
classification network
h_J
representation network
1979
t
2020
Δt

AtmoDist

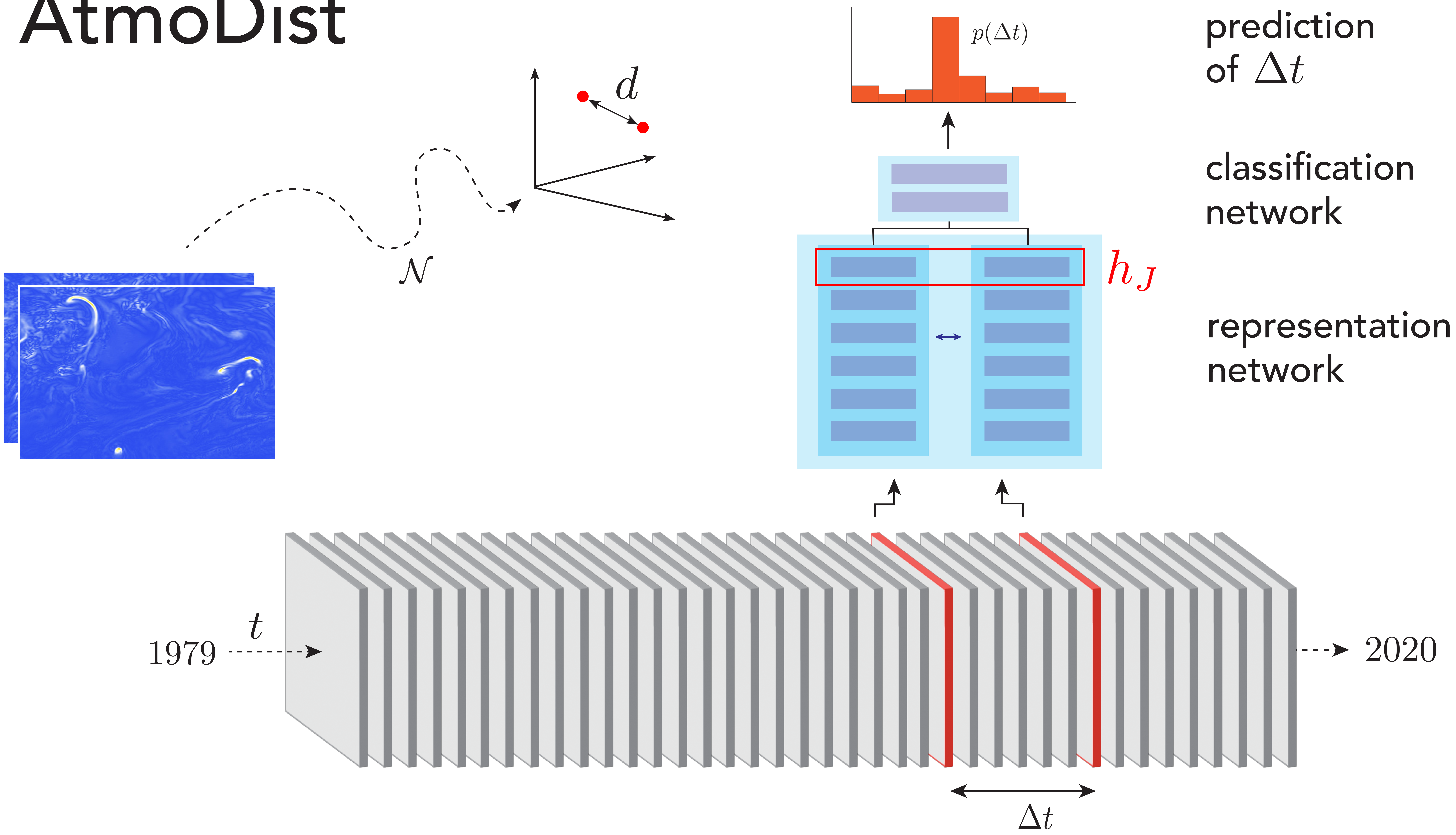


prediction of Δt

classification network

representation network

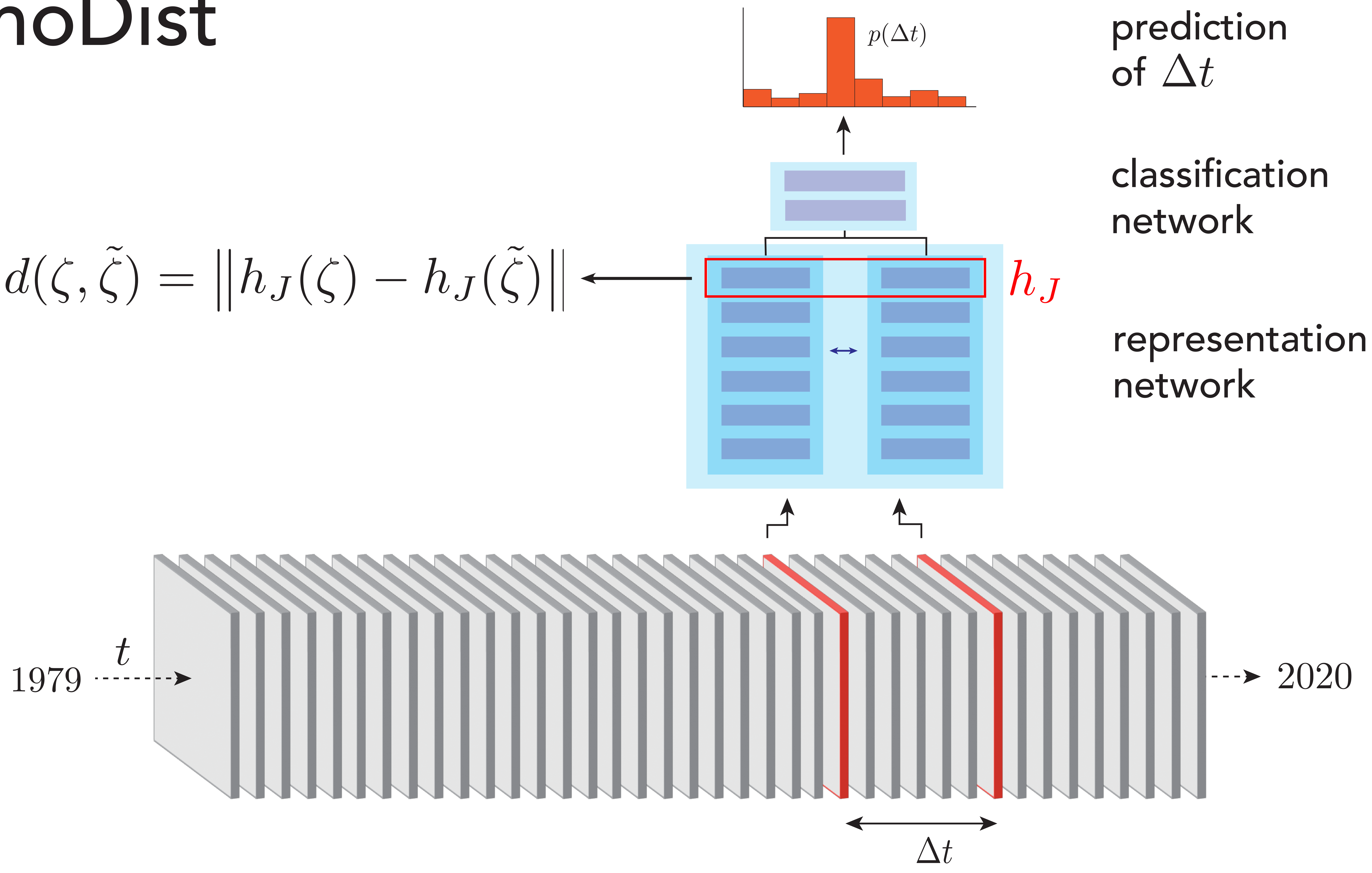
AtmoDist



d
$\mathcal{N}$
$p(\Delta t)$
prediction of Δt
classification network
h_J
representation network
1979
t
2020
Δt

AtmoDist

$$d(\zeta, \tilde{\zeta}) = \left\| h_J(\zeta) - h_J(\tilde{\zeta}) \right\|$$



Super-resolution using AtmoDist

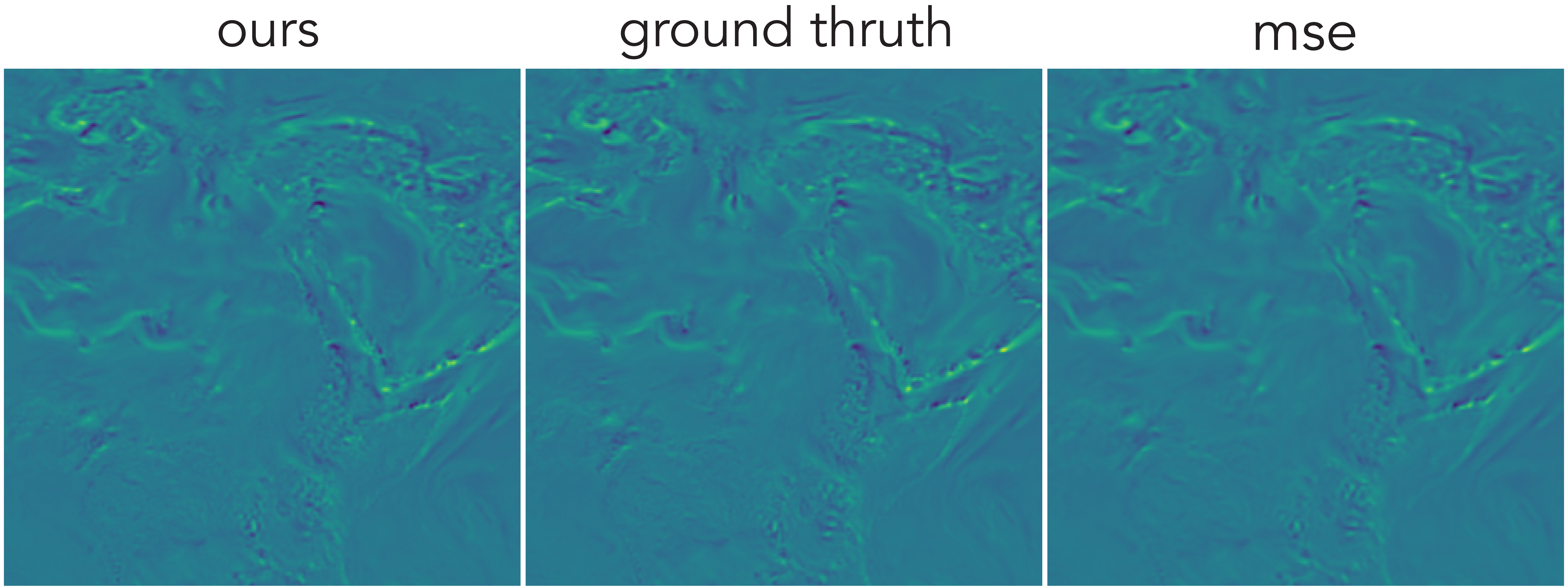
- Objective: down-scaling / upsampling of coarse fields

Super-resolution using AtmoDist

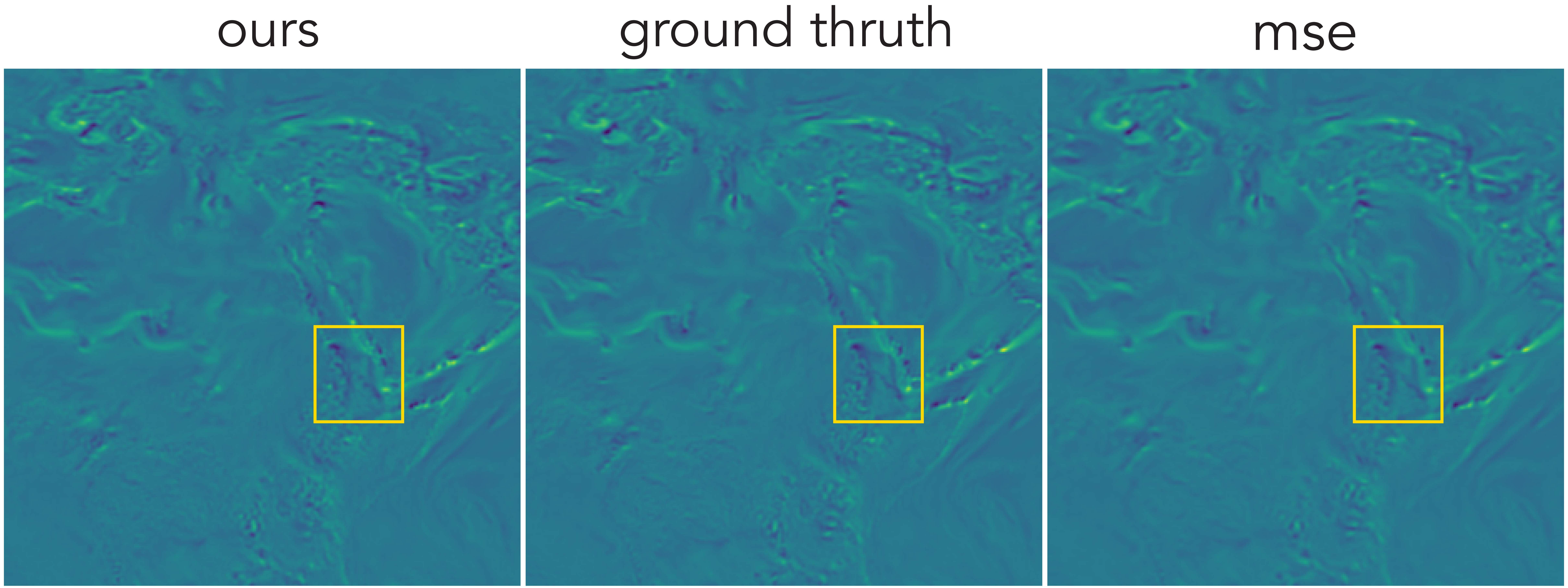
- Objective: down-scaling / upsampling of coarse fields
- Comparison using and with the GAN of Stengel et al.[1]
  - Our content loss replaces mean squared error

[1] K. Stengel, A. Glaws, D. Hettinger, and R. N. King. Adversarial super-resolution of climatological wind and solar data. Proceedings of the National Academy of Sciences, 117(29):16805–16815, 2020.

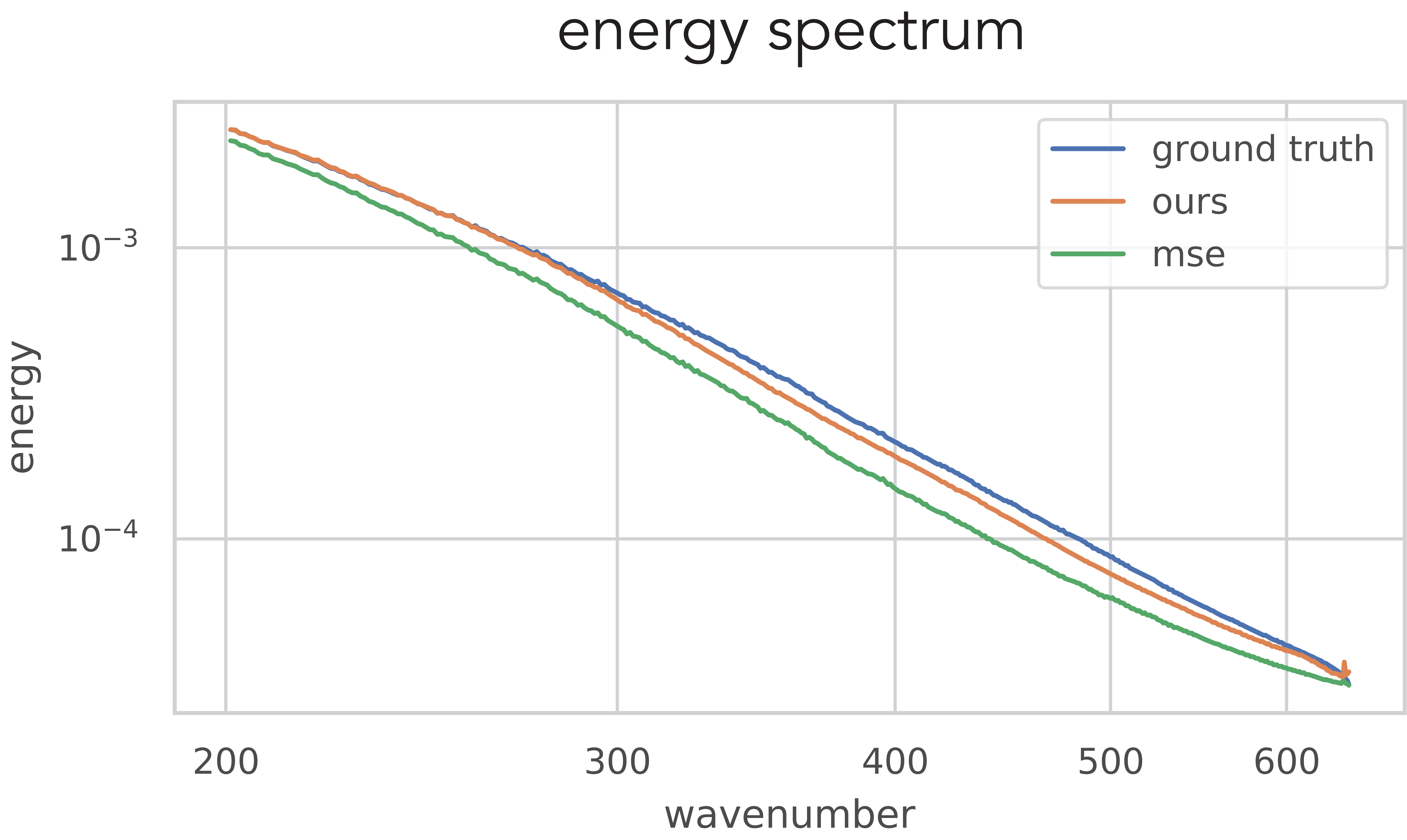
Super-resolution using AtmoDist



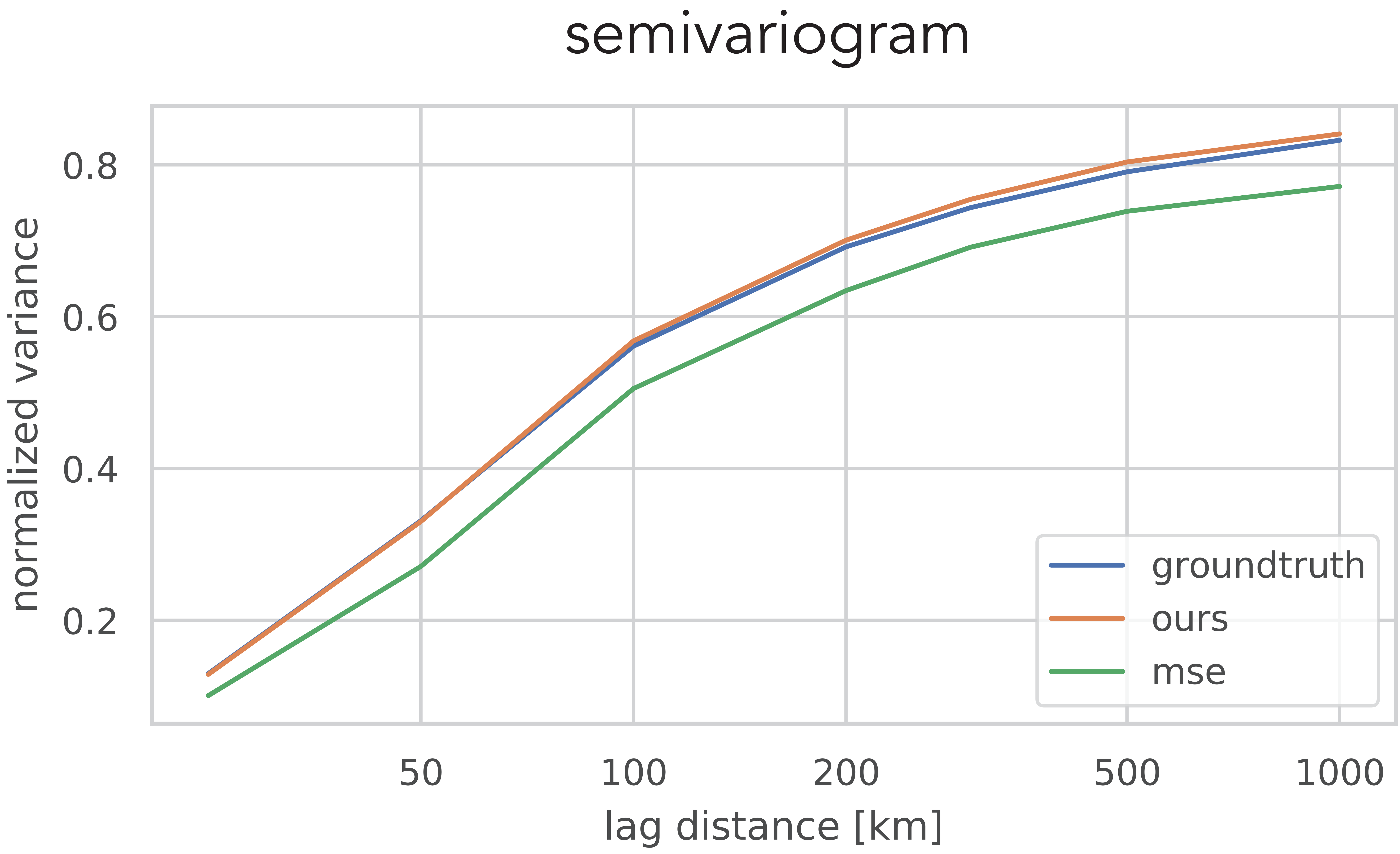
Super-resolution using AtmoDist



Super-resolution using AtmoDist

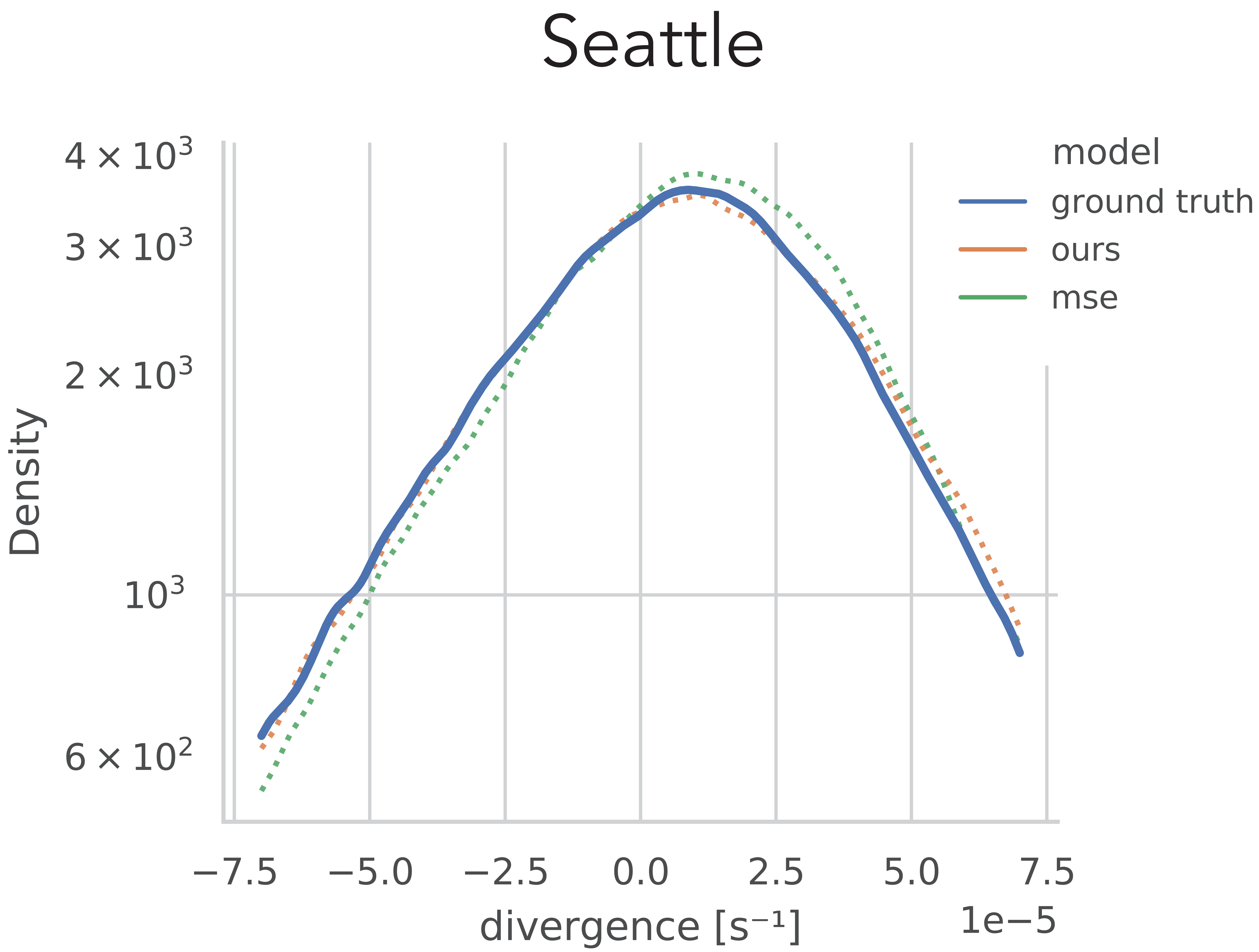


Super-resolution using AtmoDist

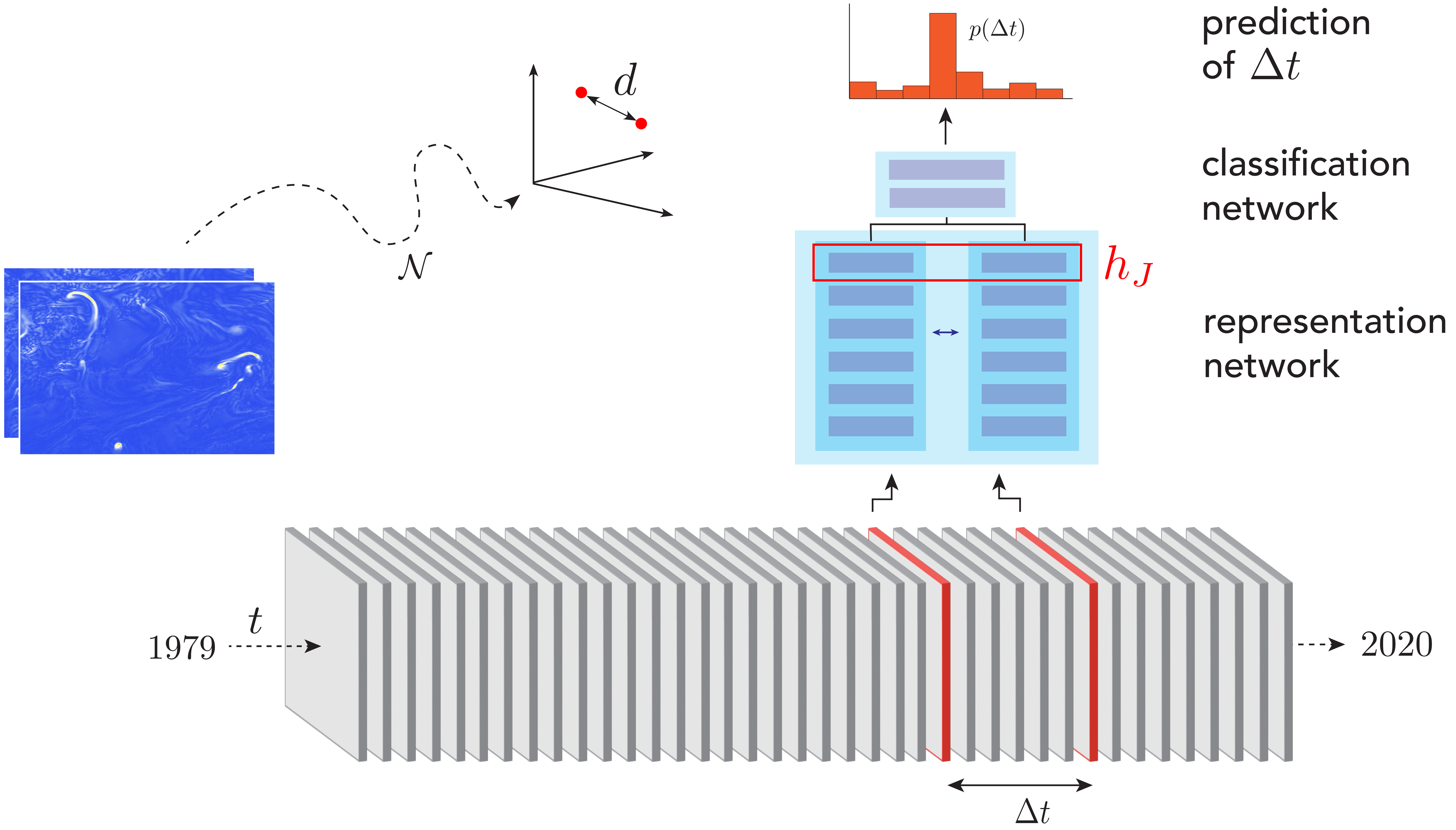


Super-resolution using AtmoDist

- Local statistics by averaging over super-resolution predictions for entire reanalysis data set
- 150 big cities as locations



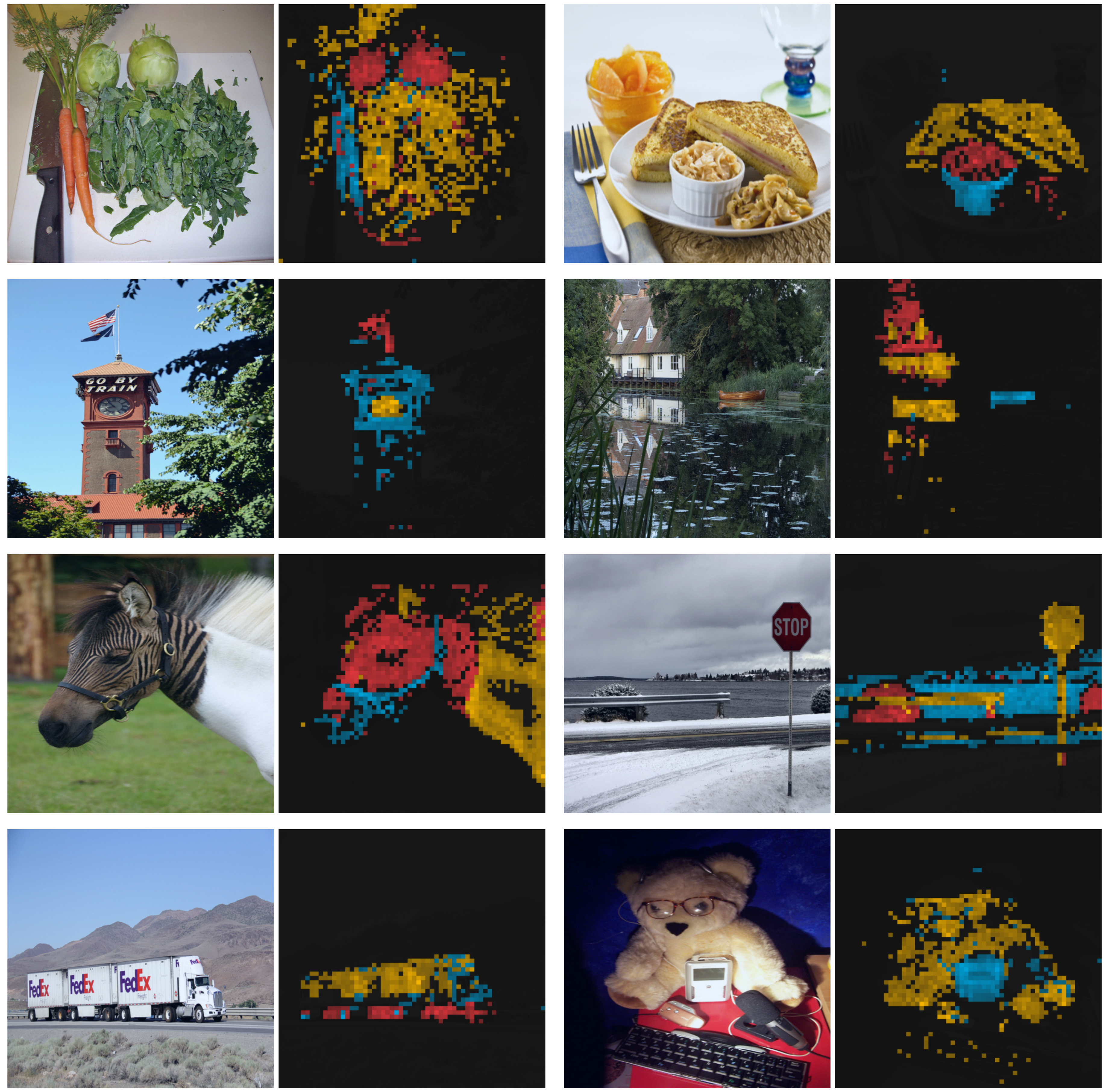
AtmoDist



$p(\Delta t)$
prediction of Δt
classification network
h_J
representation network
d
$\mathcal{N}$
t
1979
2020
Δt

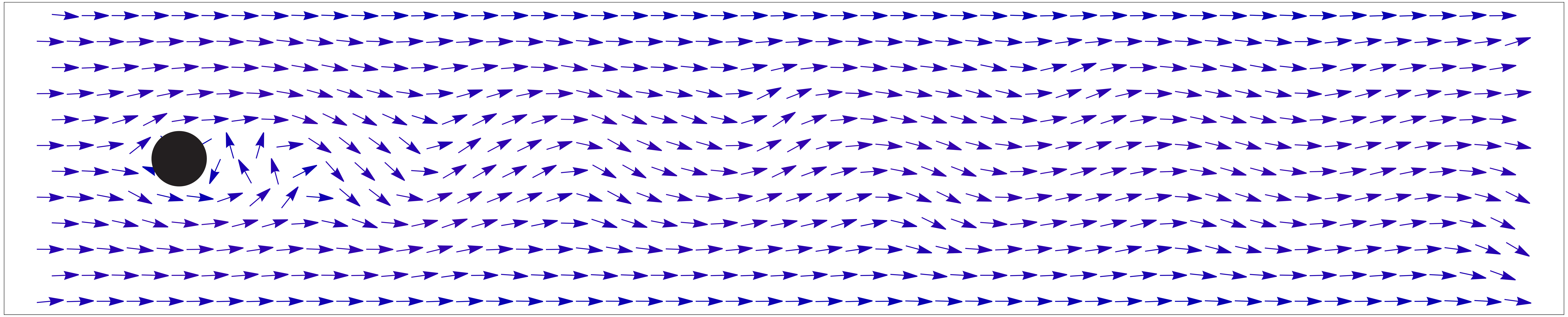
Transformers and attention

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.



M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. CoRR, abs/2104.14294, 2021.

Transformers and attention

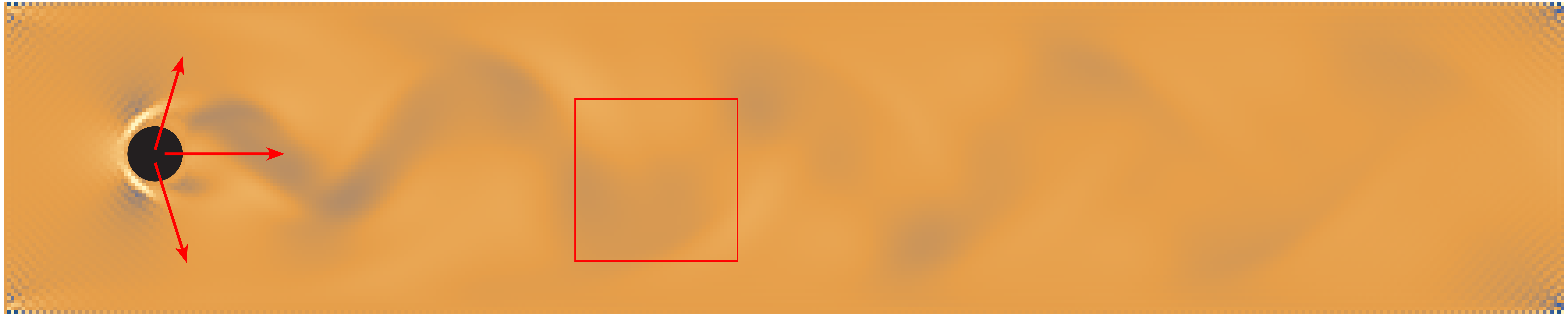


Fluid flow (vorticity)

Fluid flow (vorticity)

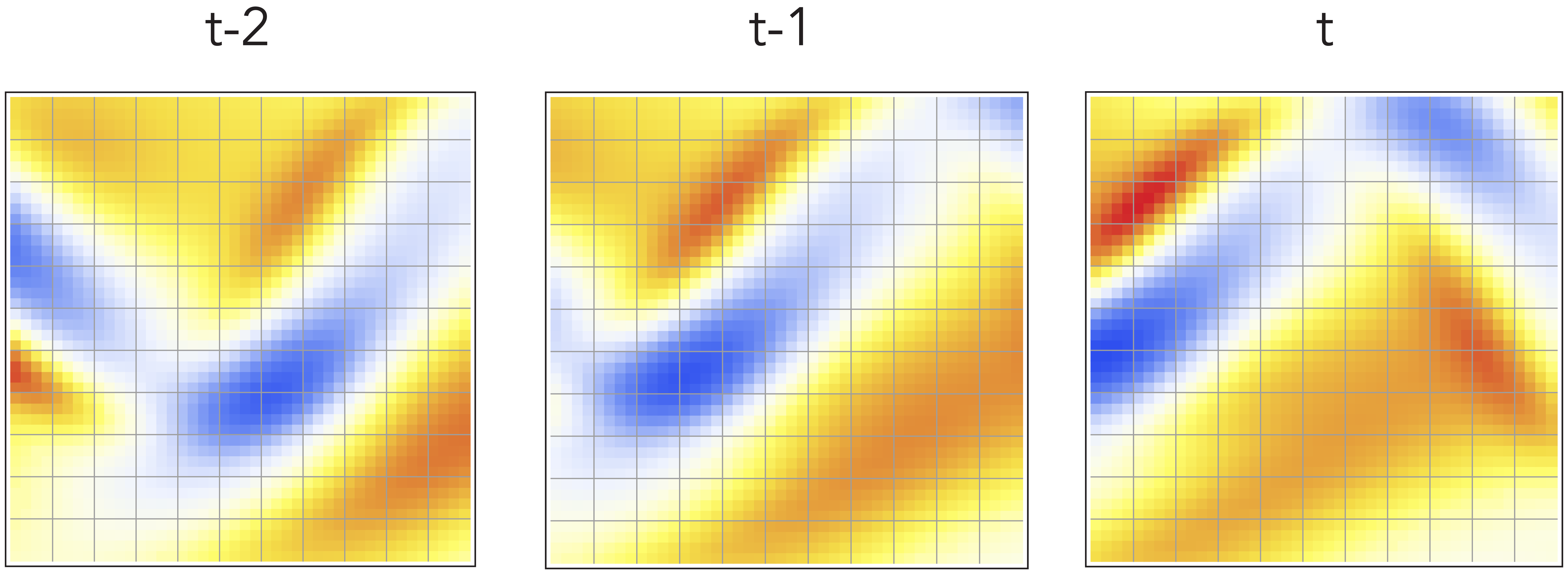
training with varying position
and spherical eccentrity

Fluid flow (vorticity)



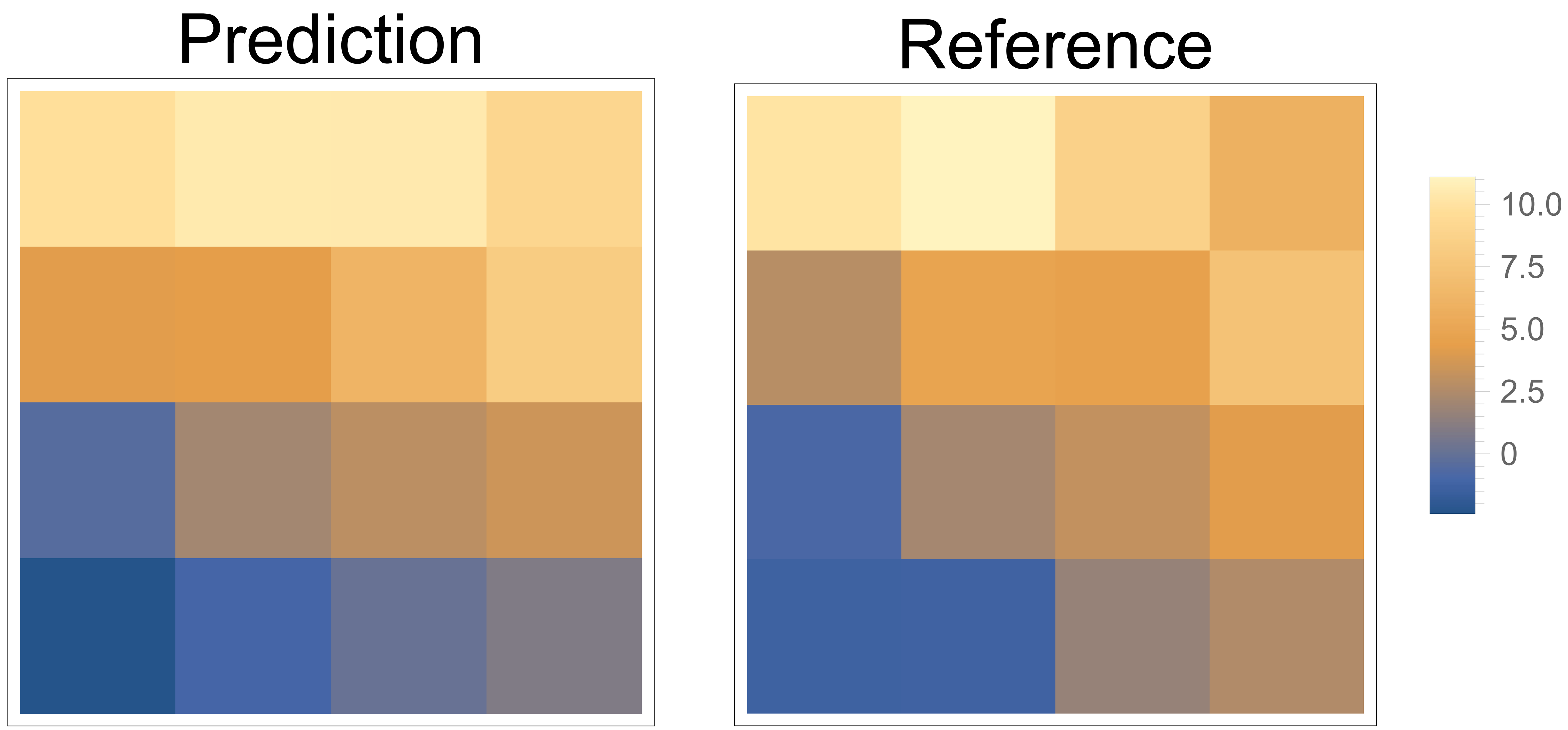
training with varying position
and spherical eccentrity

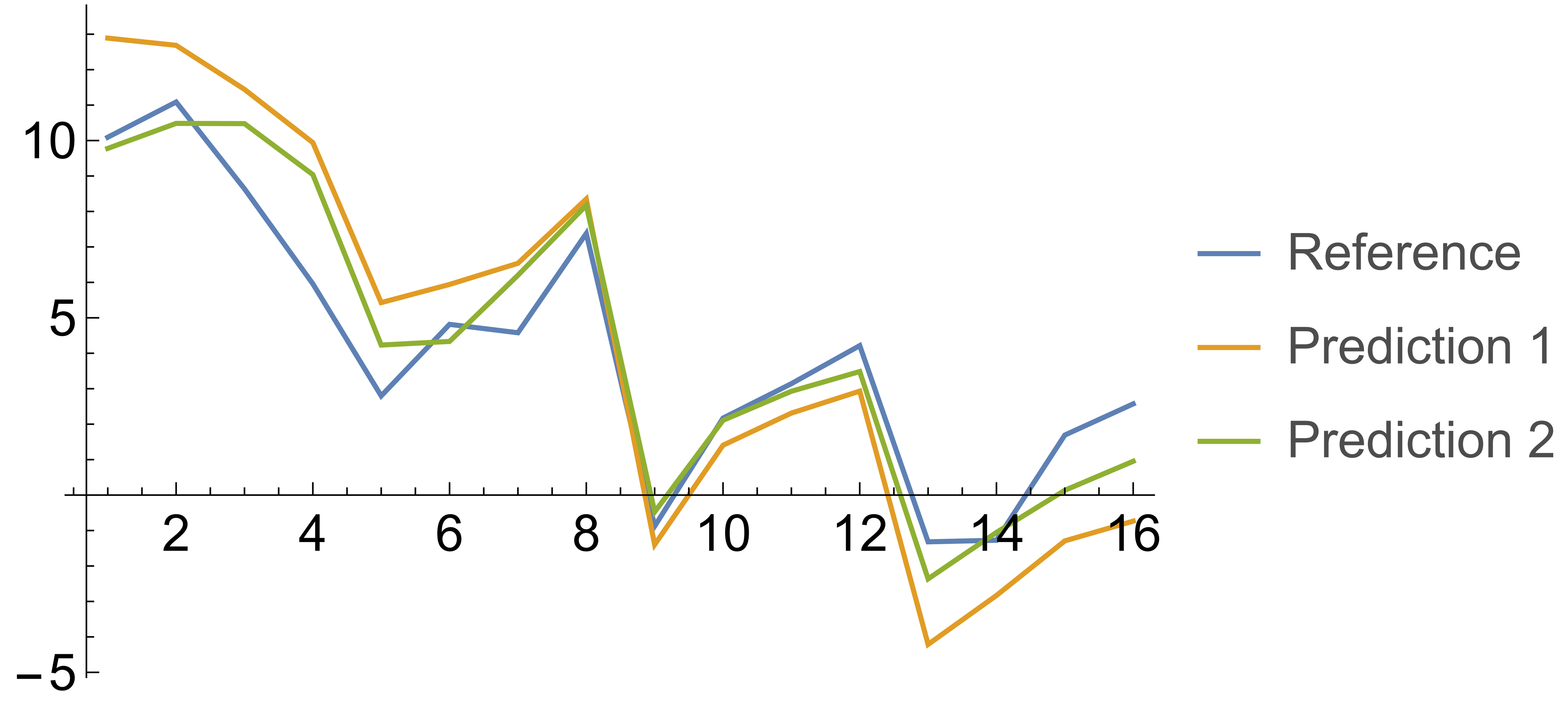
Fluid flow (vorticity)



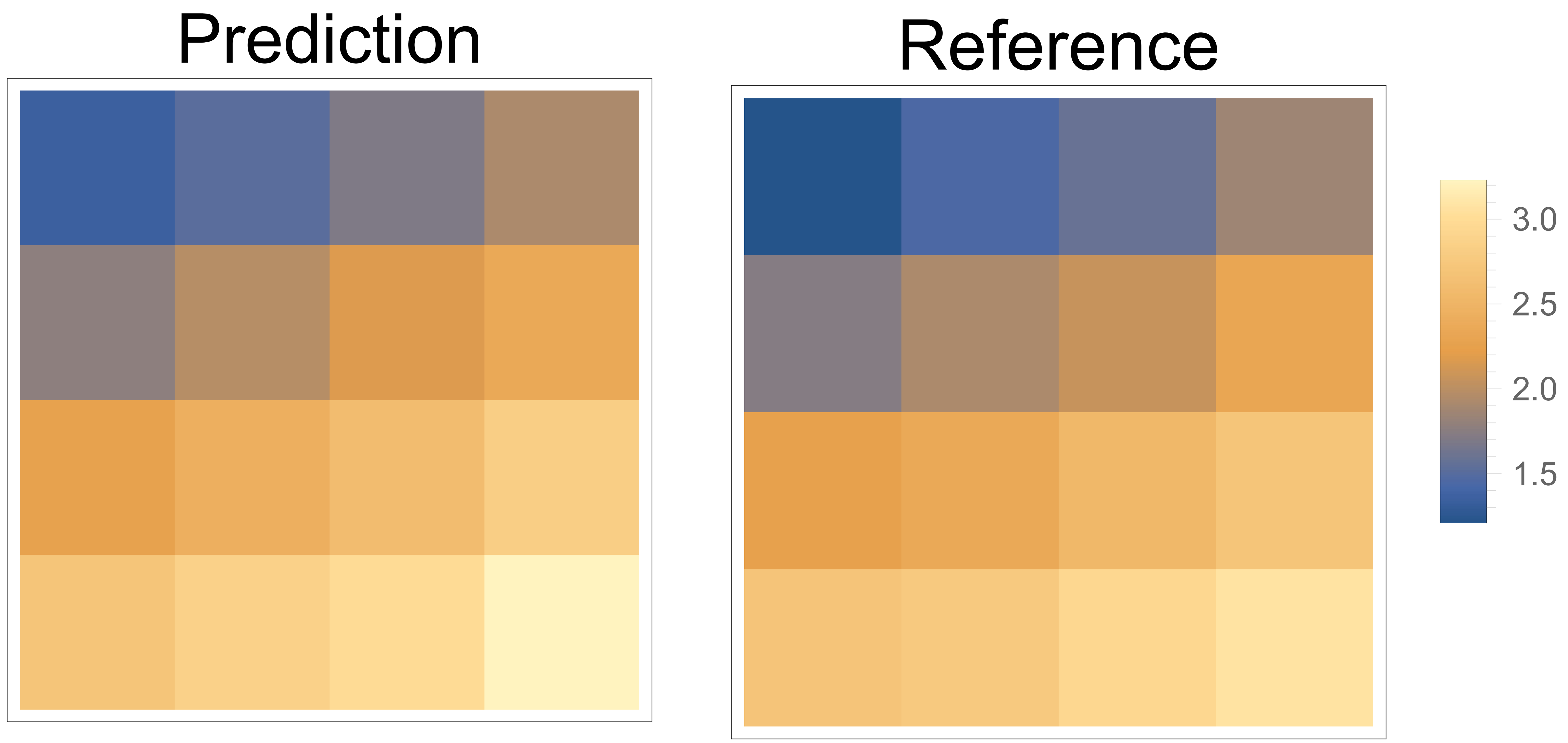
Fluid flow (vorticity)

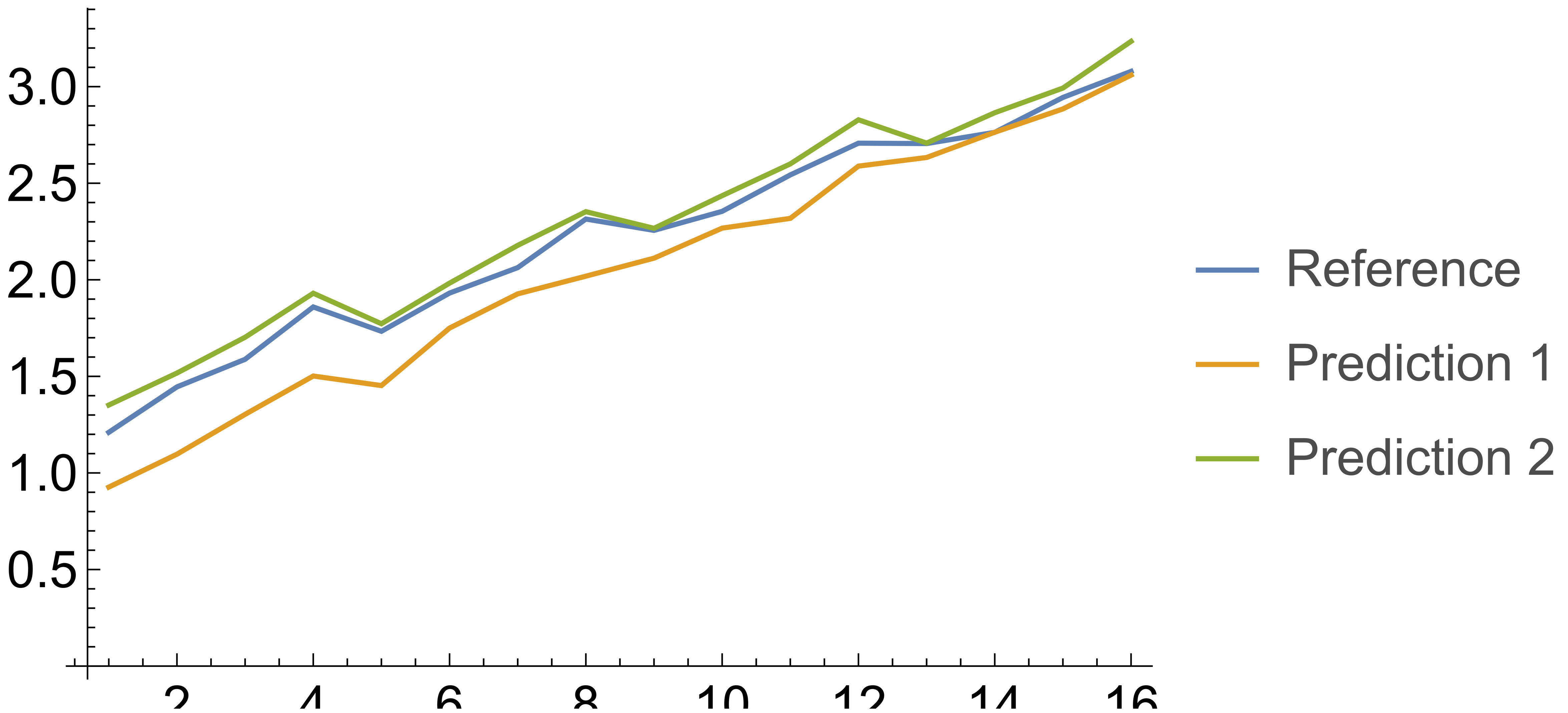
Fluid flow


Prediction
Reference
10.0
7.5
5.0
2.5
0

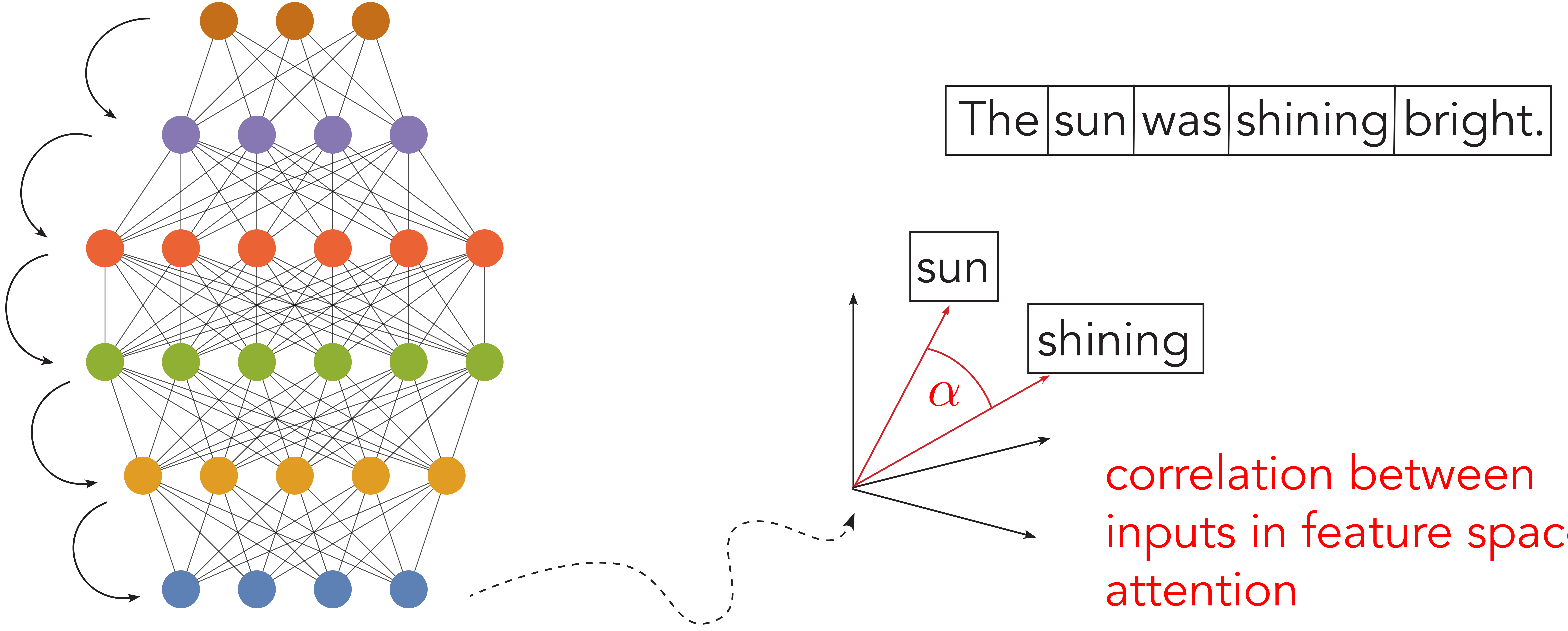

10
5
−5
2
4
6
8
10
12
14
16
Reference
Prediction 1
Prediction 2

Fluid flow


Prediction
Reference
3.0
2.5
2.0
1.5


3.0
2.5
2.0
1.5
1.0
0.5
Reference
Prediction 1
Prediction 2

Fluid flow

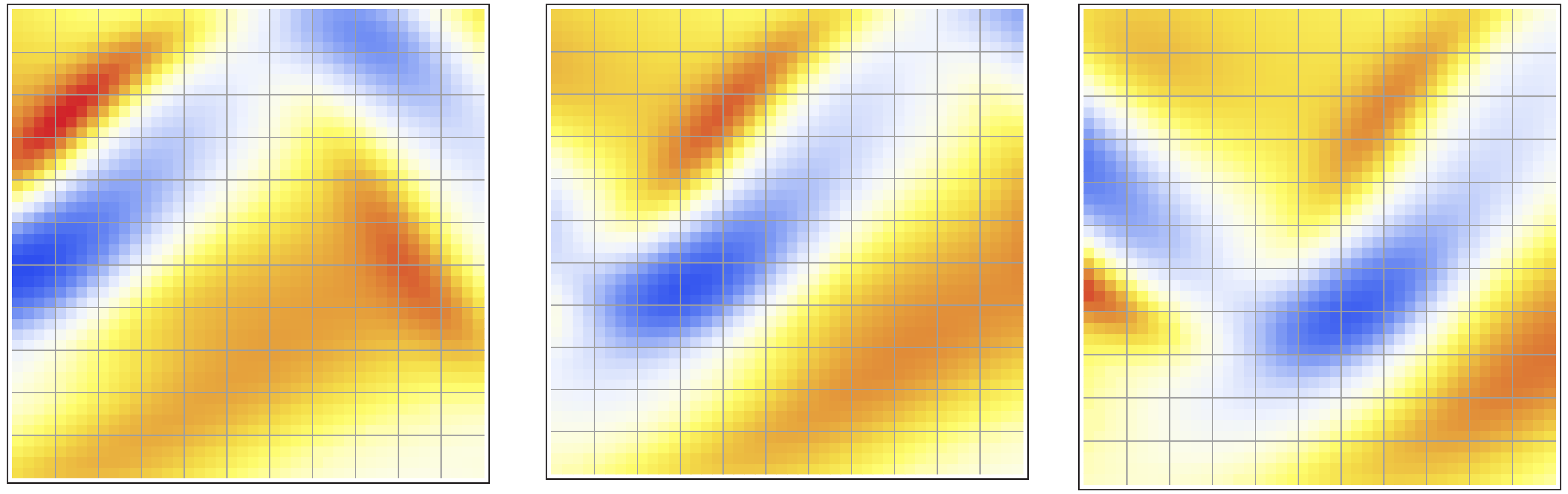


correlation between inputs in feature space: attention

Fluid flow

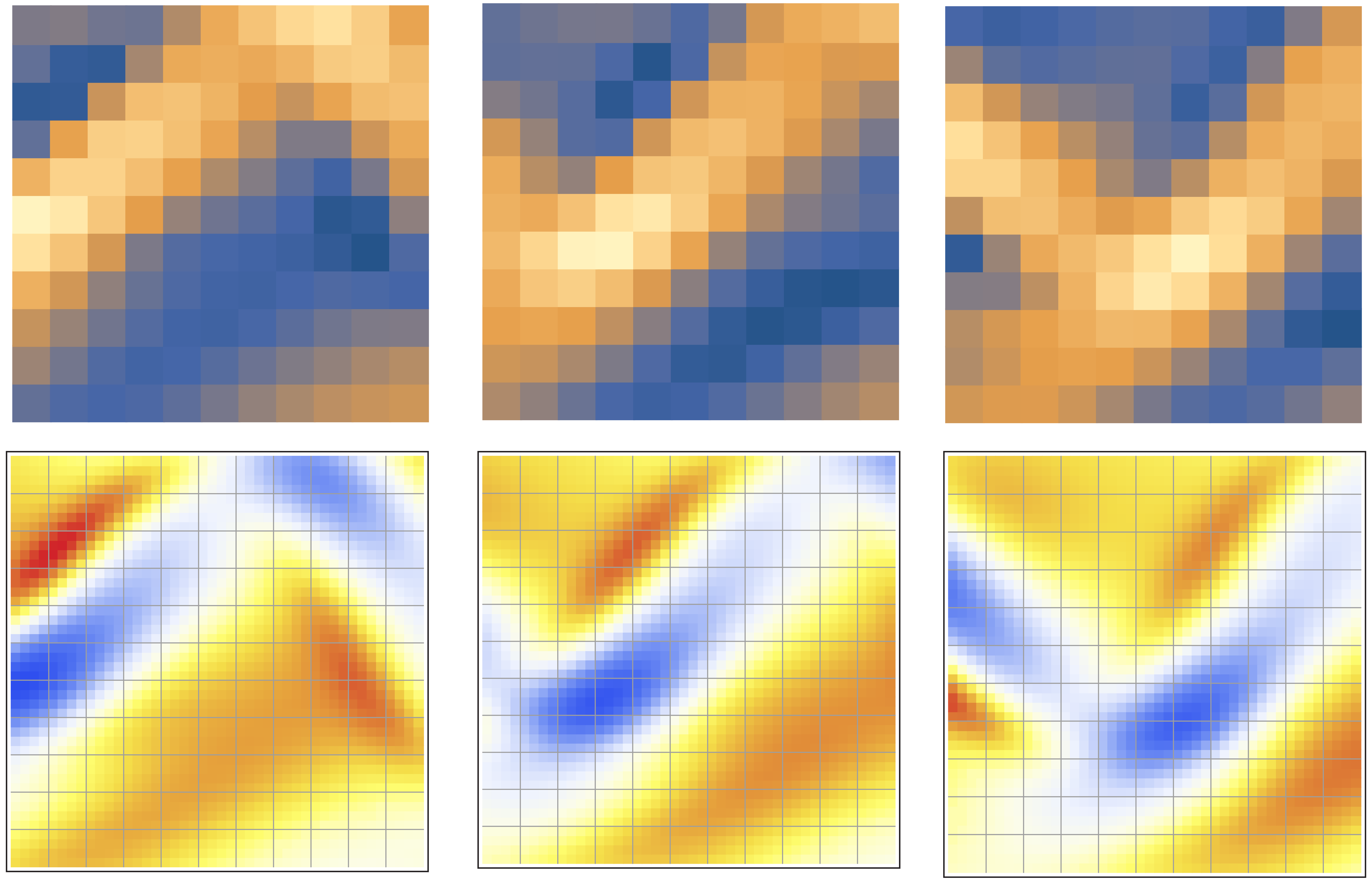
correlation between inputs in feature space: attention

Fluid flow



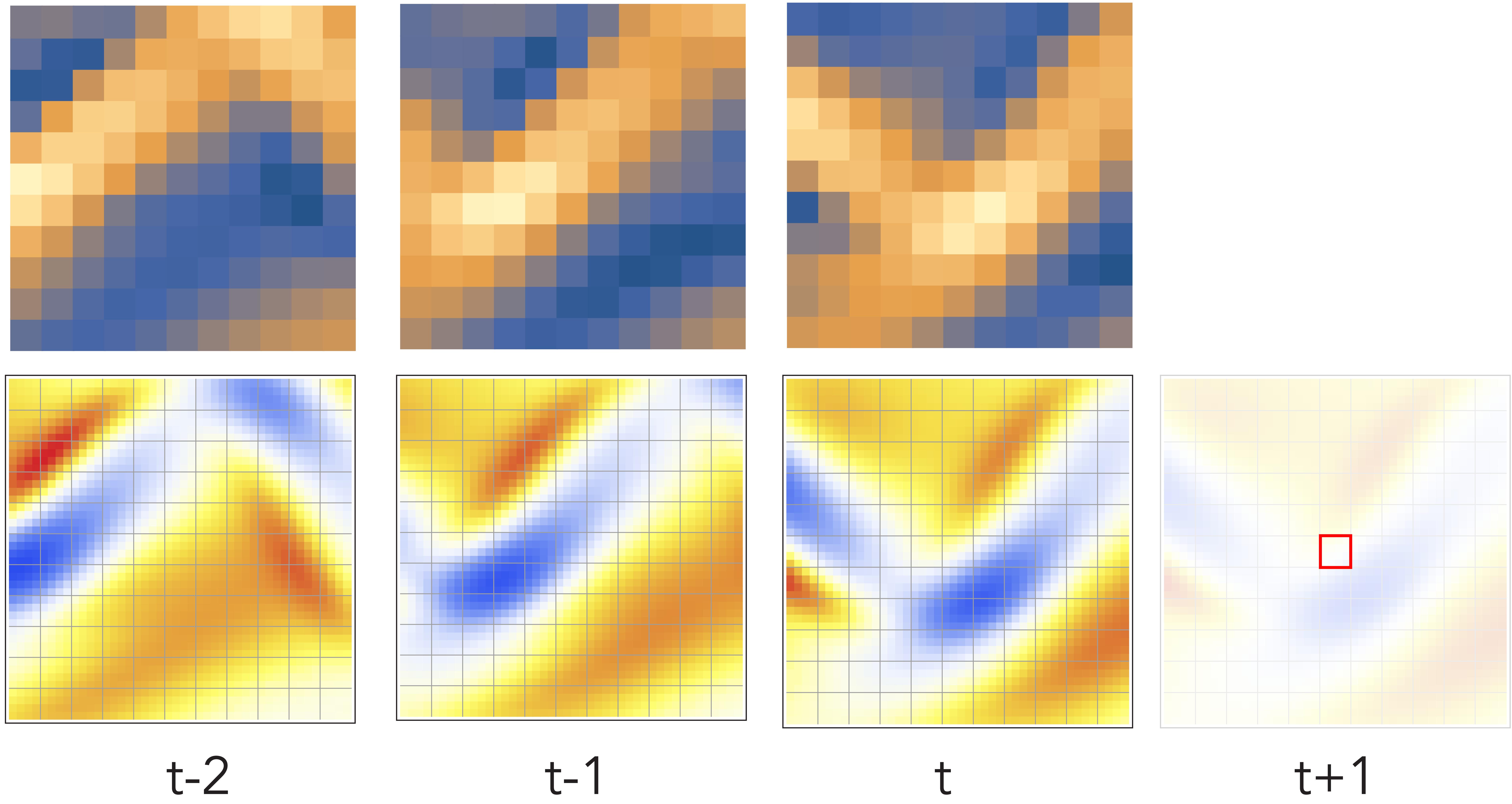
t-2 t-1 t

Fluid flow

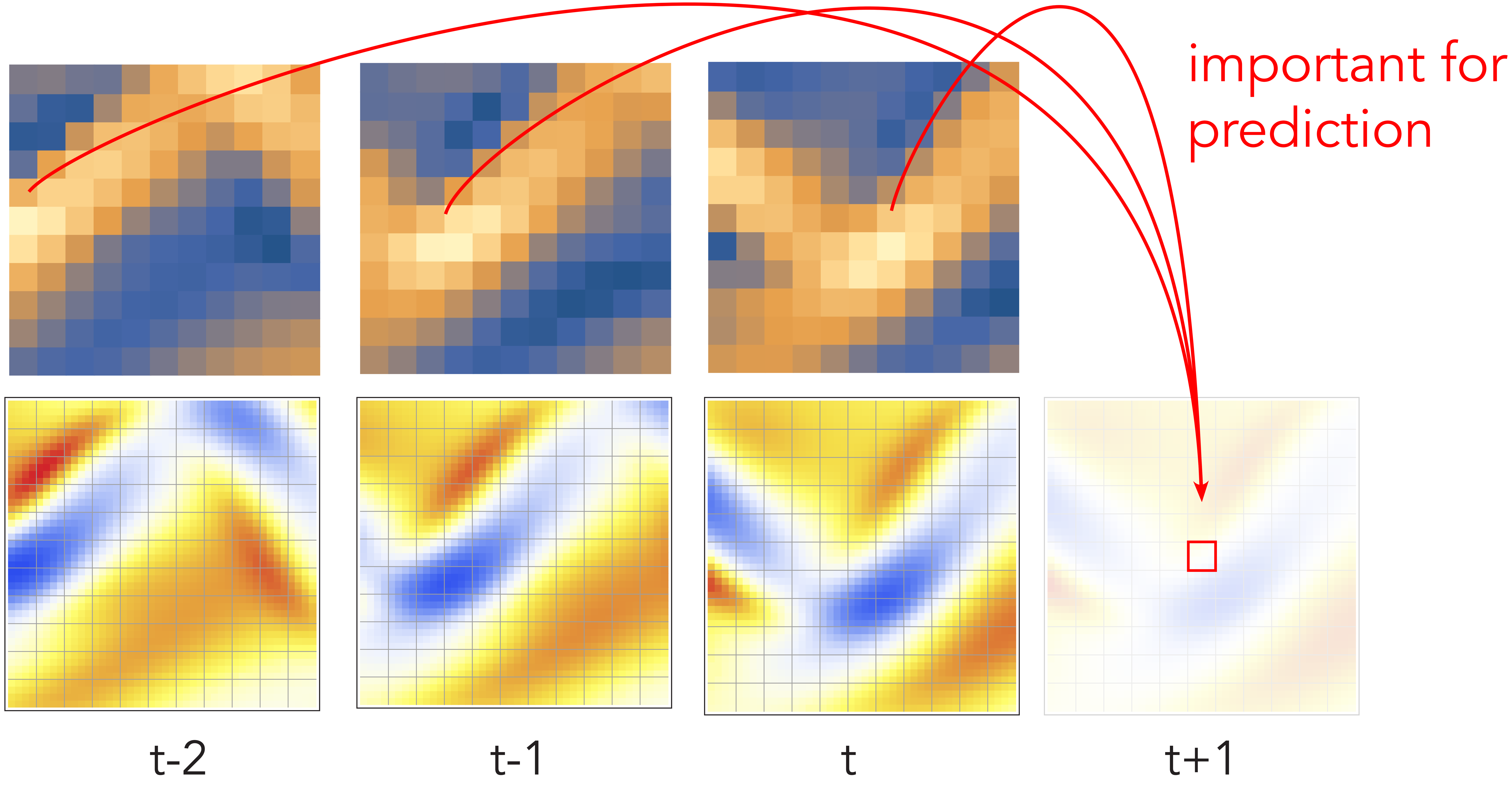


t-2 t-1 t

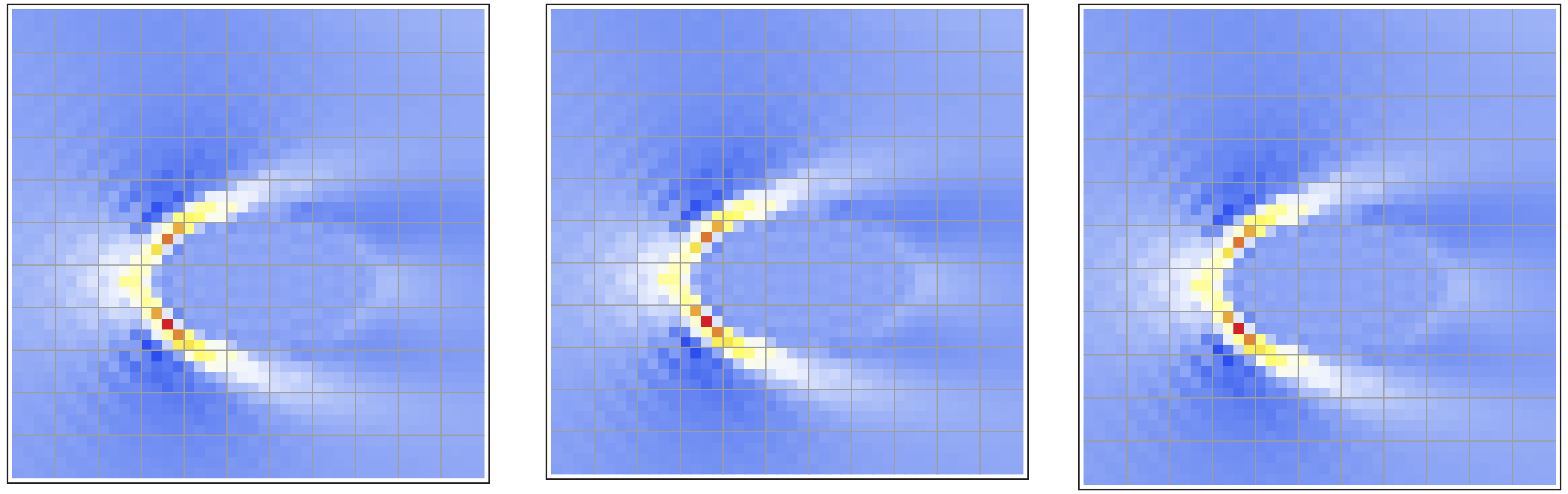
Fluid flow



Fluid flow

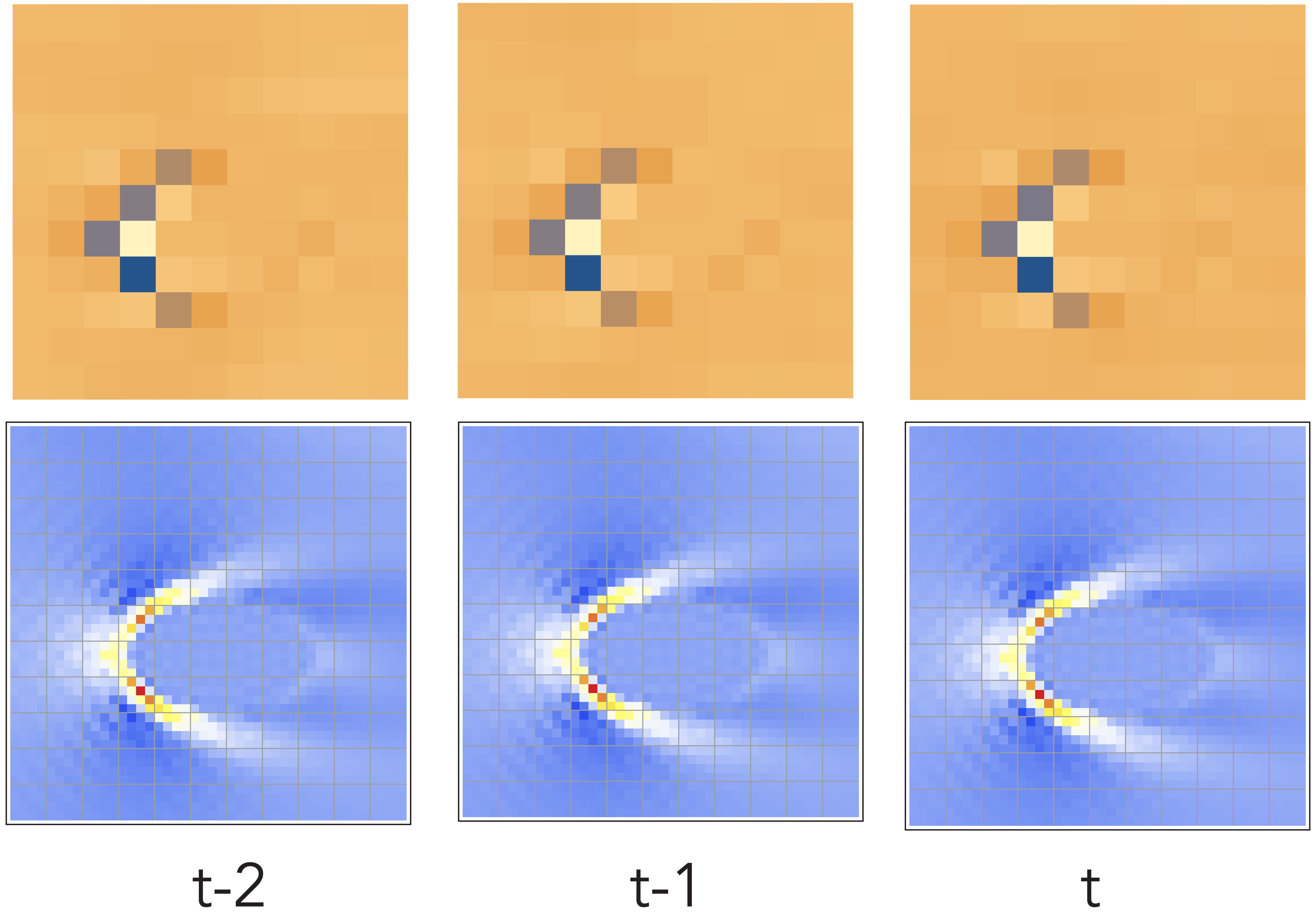


Fluid flow

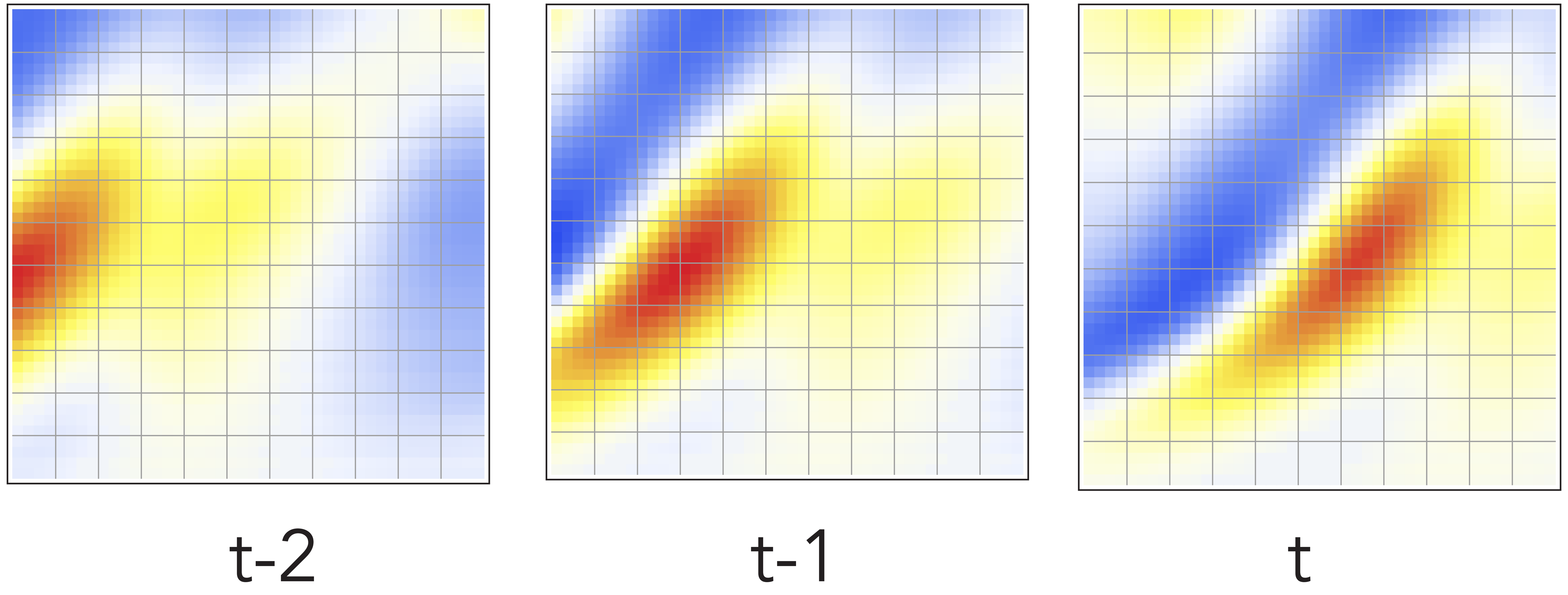


t-2 t-1 t

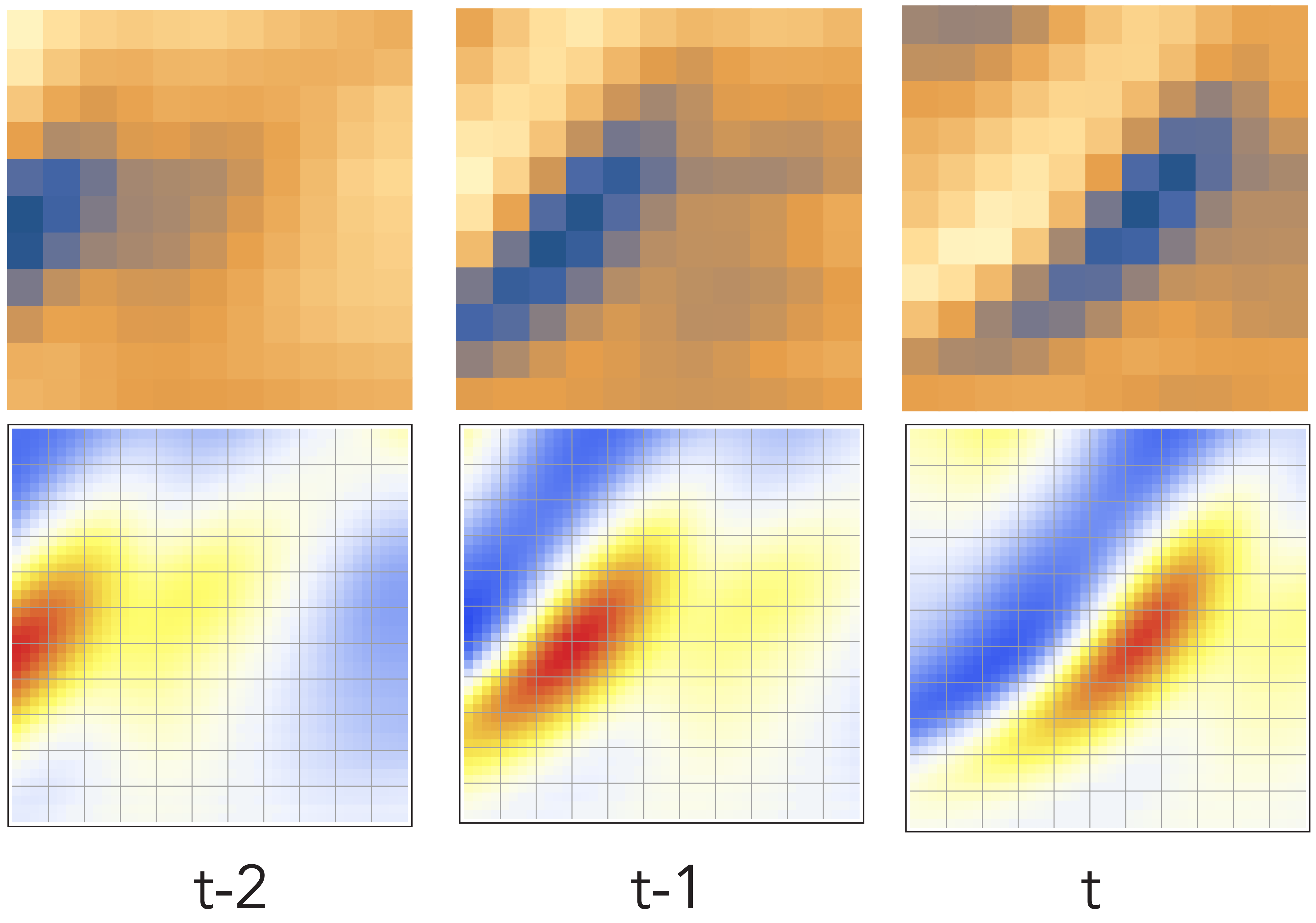
Fluid flow



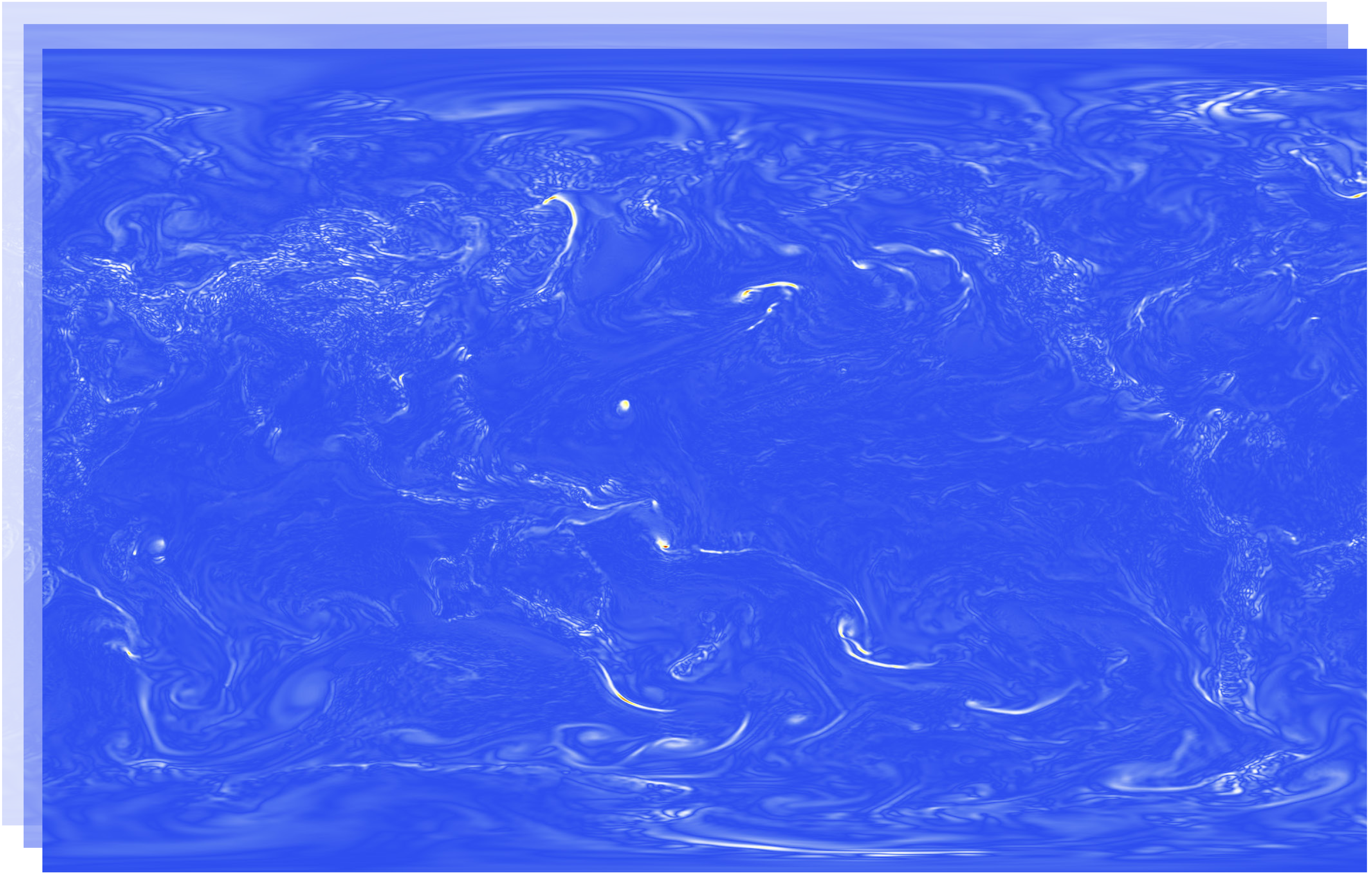
Fluid flow



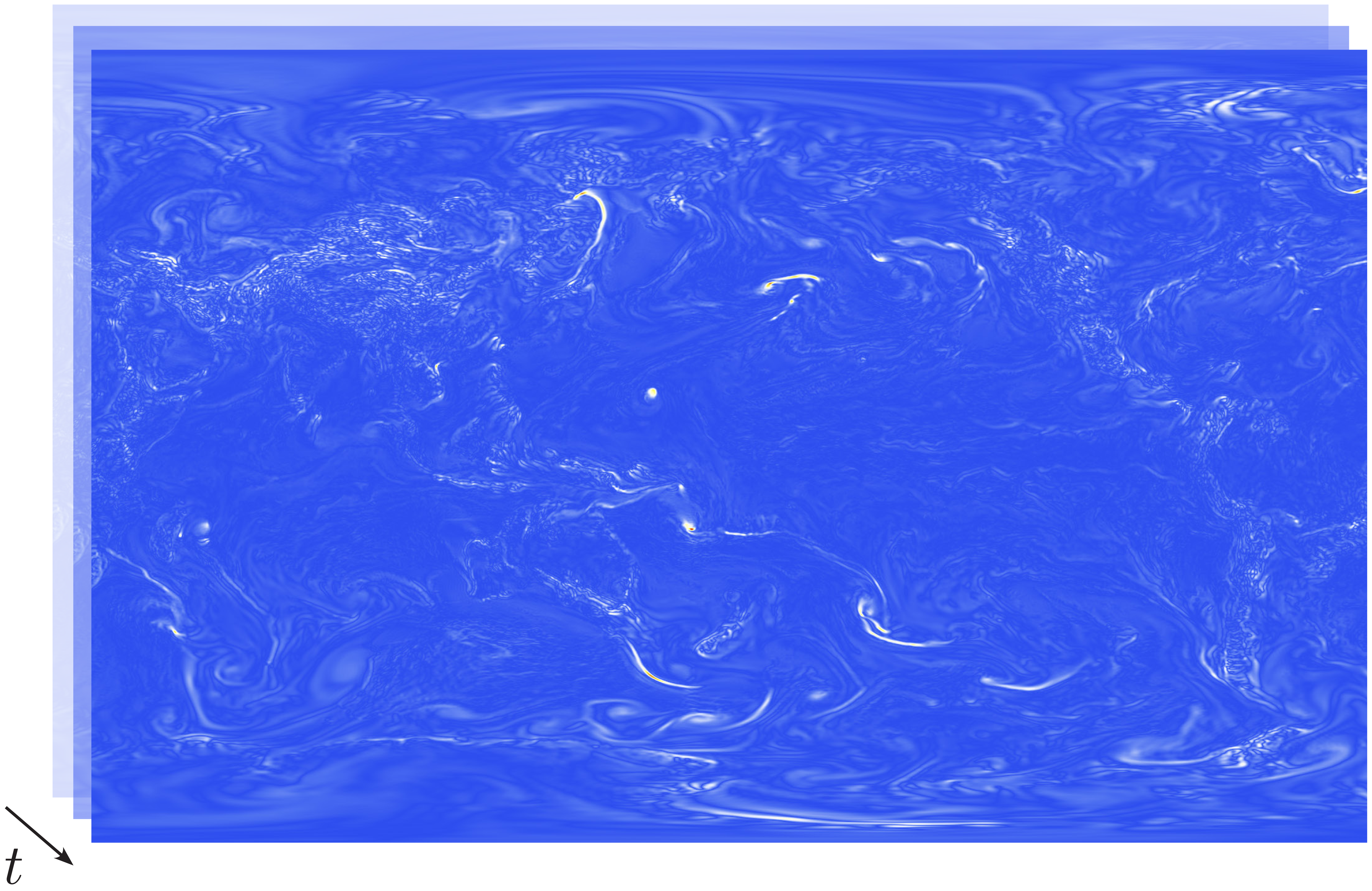
Fluid flow



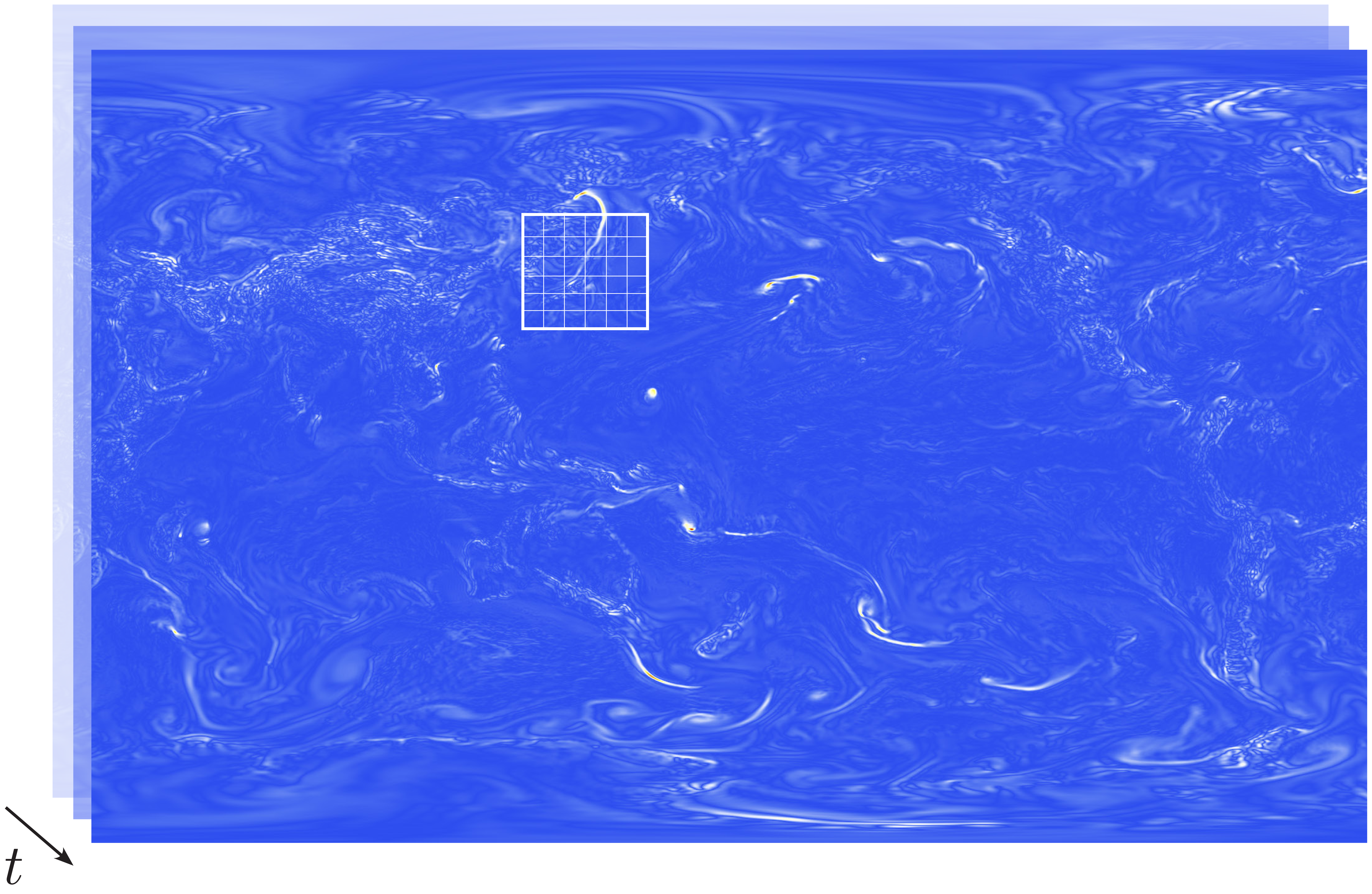
Atmospheric vorticity (ERA5)



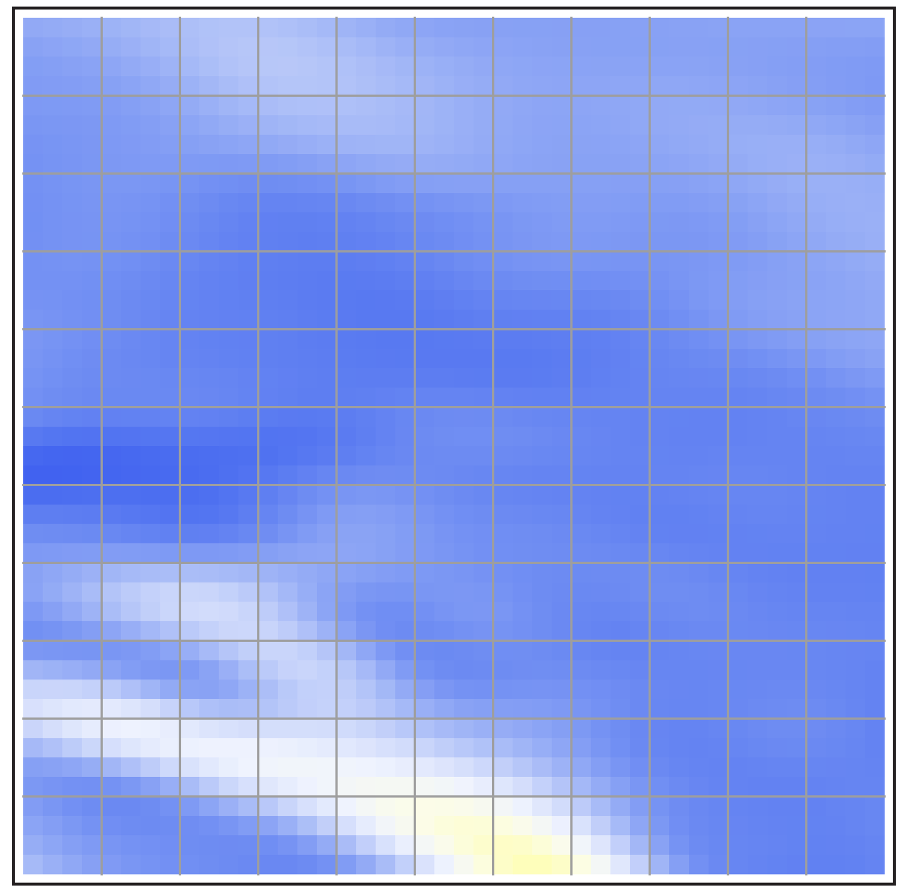
Atmospheric vorticity (ERA5)



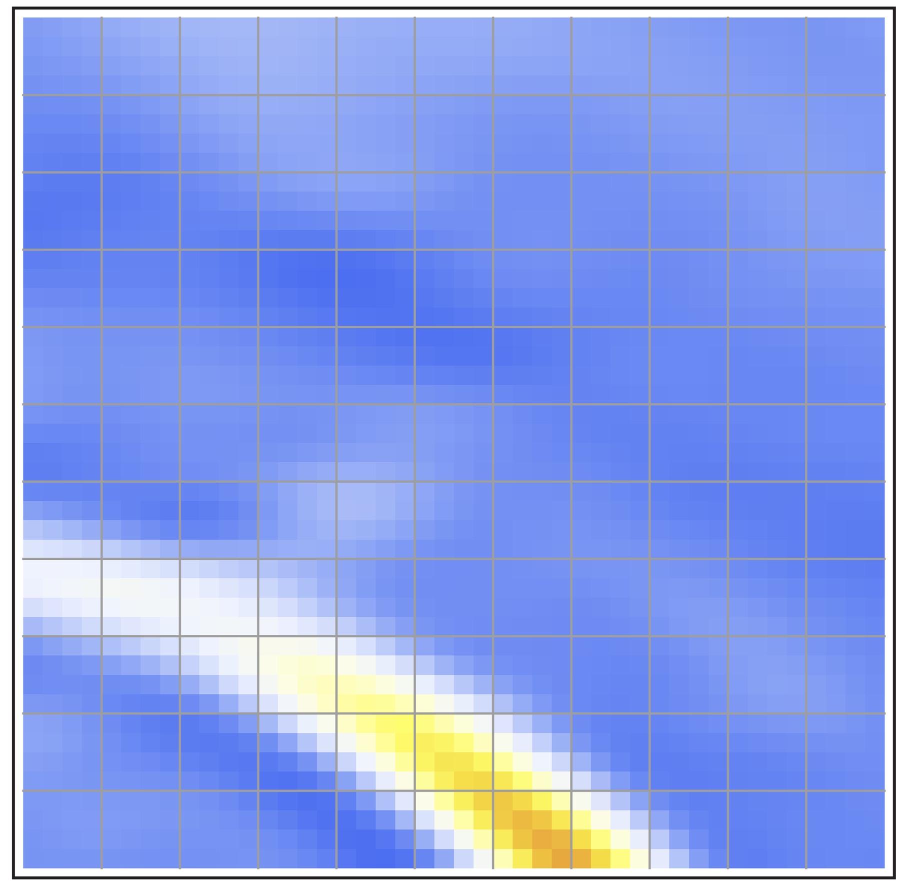
Atmospheric vorticity (ERA5)



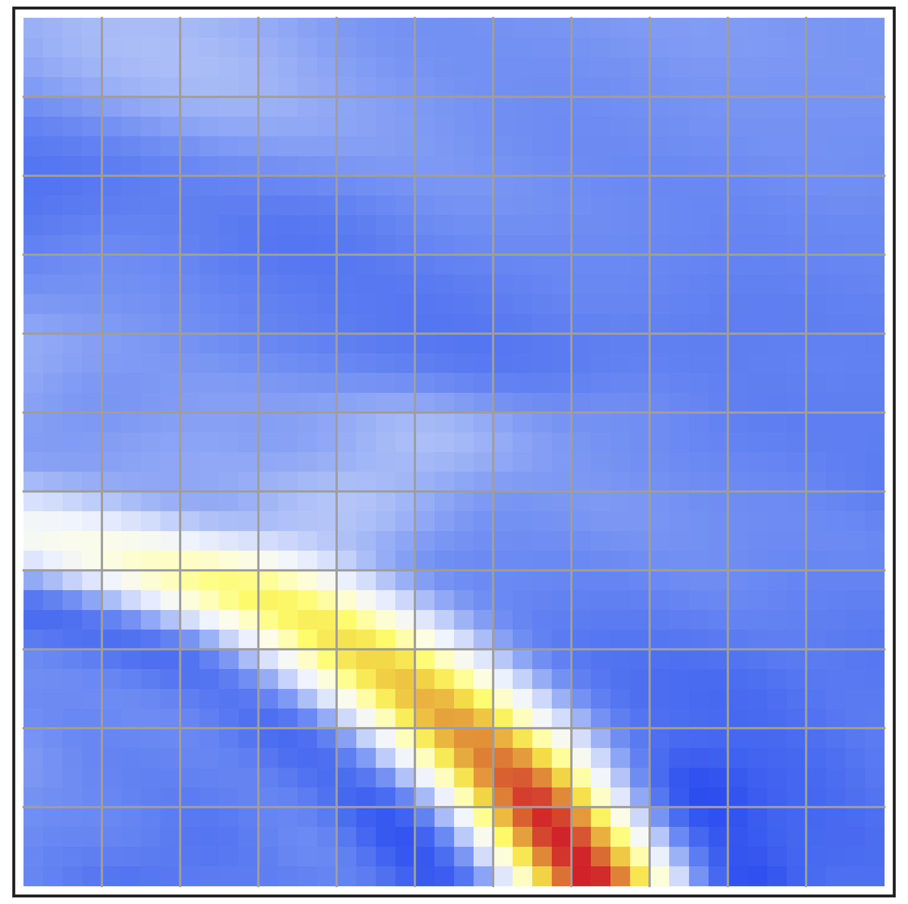
Atmospheric vorticity (ERA5)



t-2

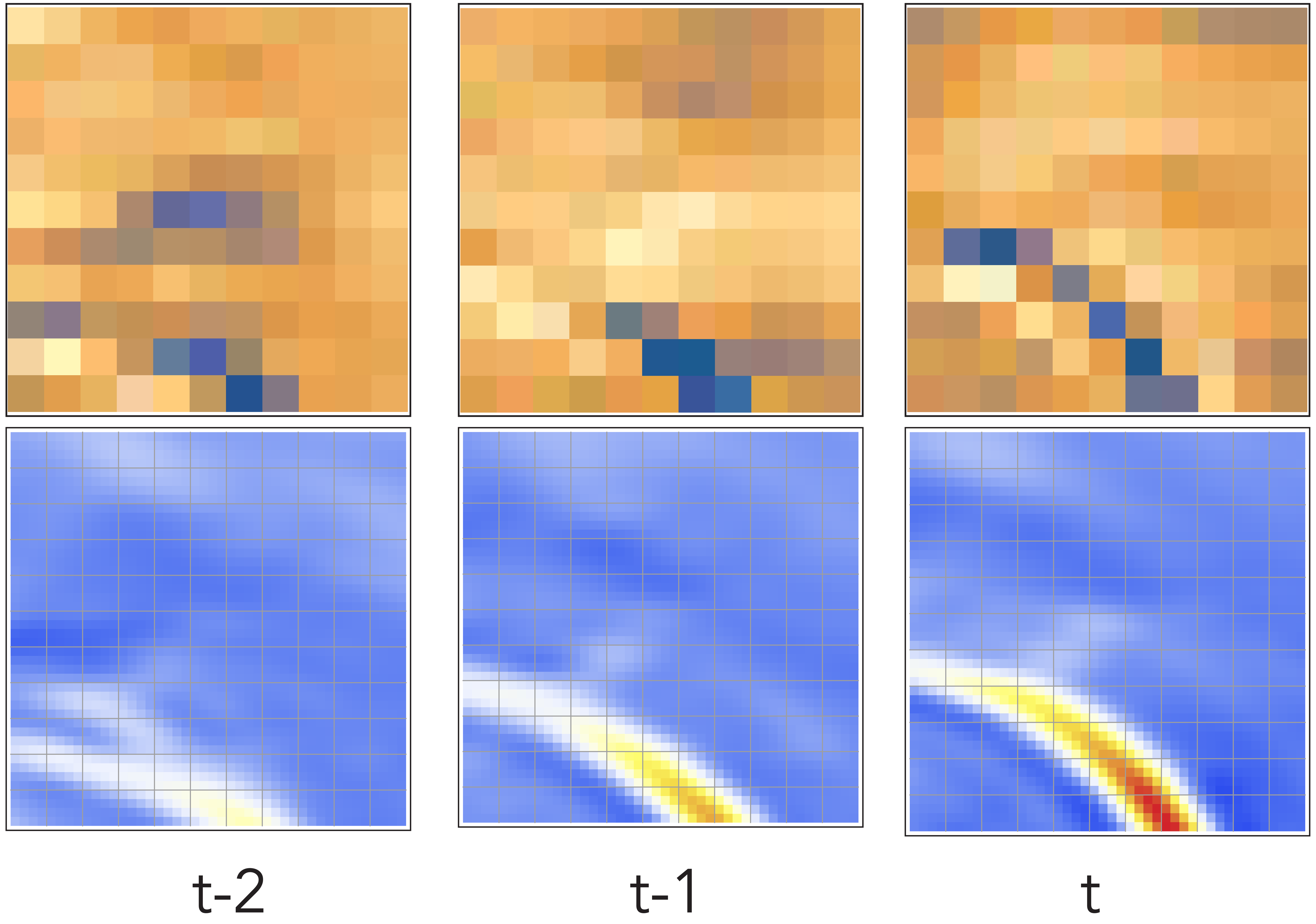


t-1

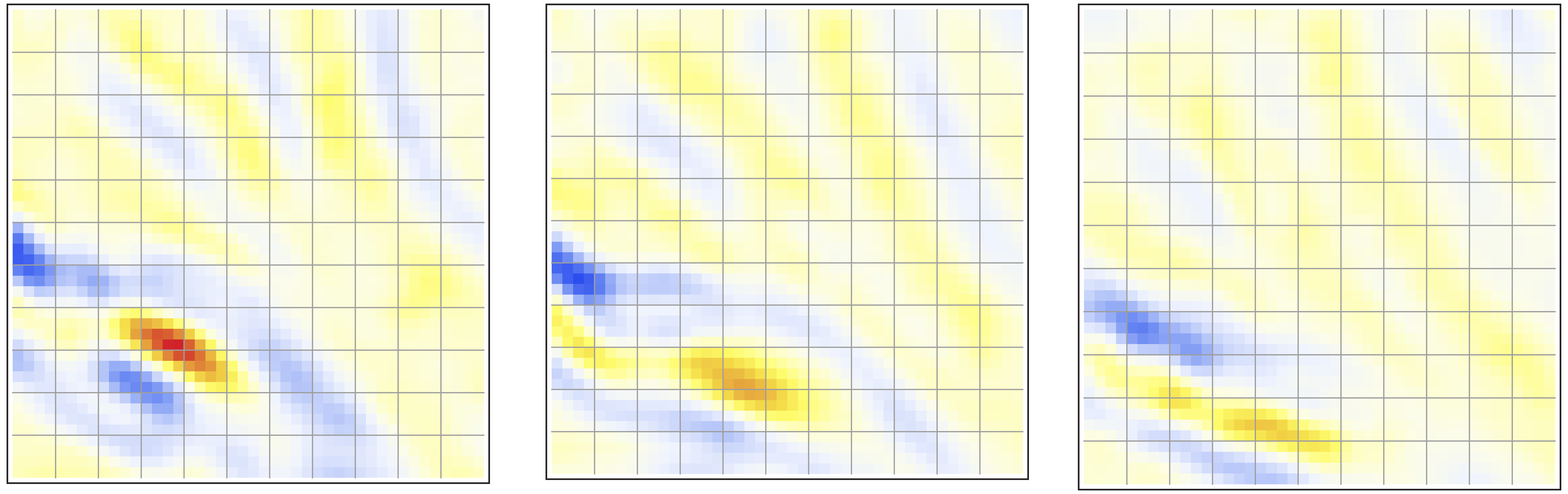


t

Atmospheric vorticity (ERA5)

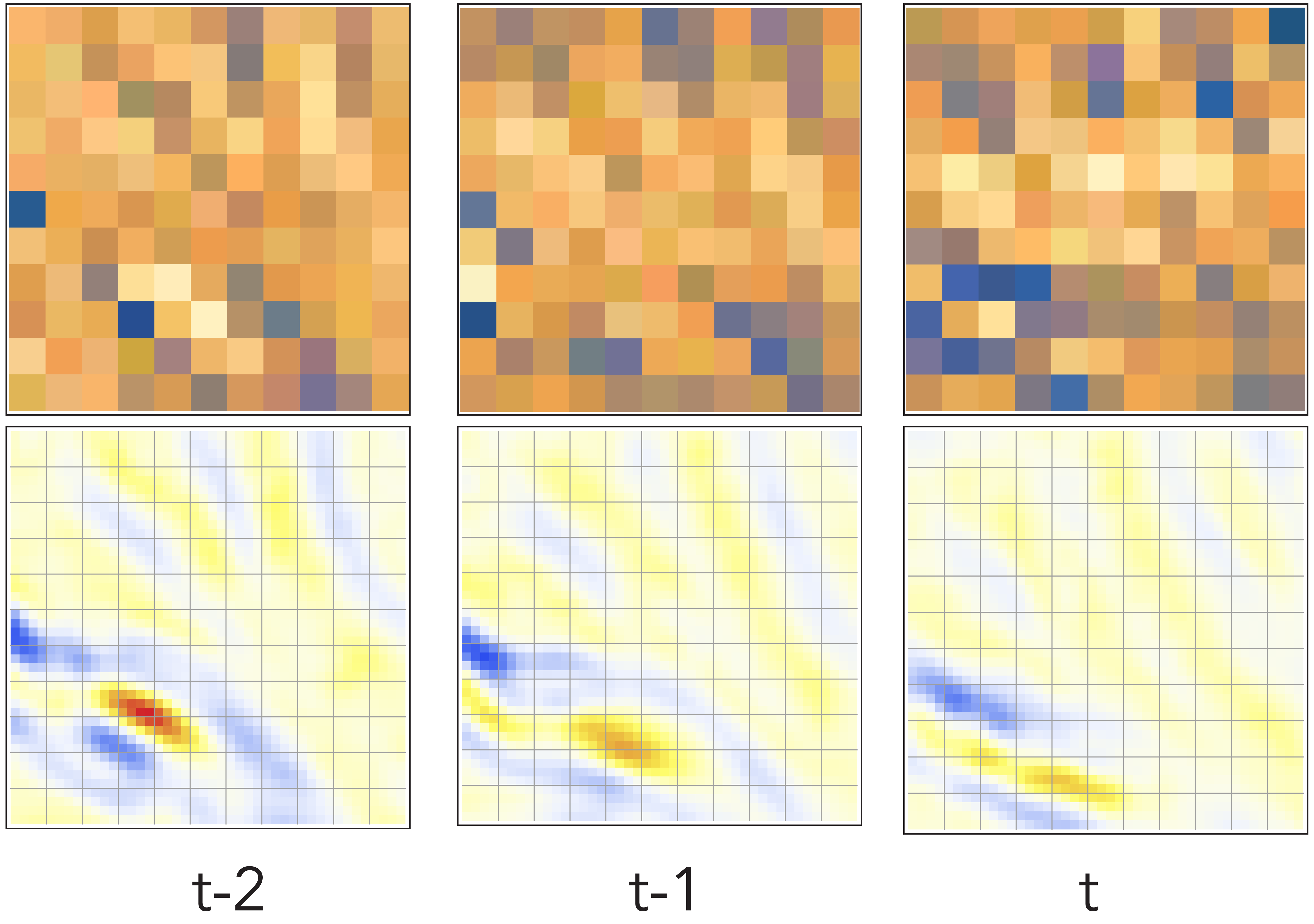


Atmospheric vorticity (ERA5)



t-2 t-1 t

Atmospheric vorticity (ERA5)



Physics and Learning?

Physics and Learning?

Scientific machine learning: use constraints from known models (e.g. symmetries) in the machine learning model

Physics and Learning?

Scientific machine learning: use constraints from known models (e.g. symmetries) in the machine learning model

Observational data: avoid the (inductive) biases and constraints we have in analytic models in the learning

Physics and Learning?

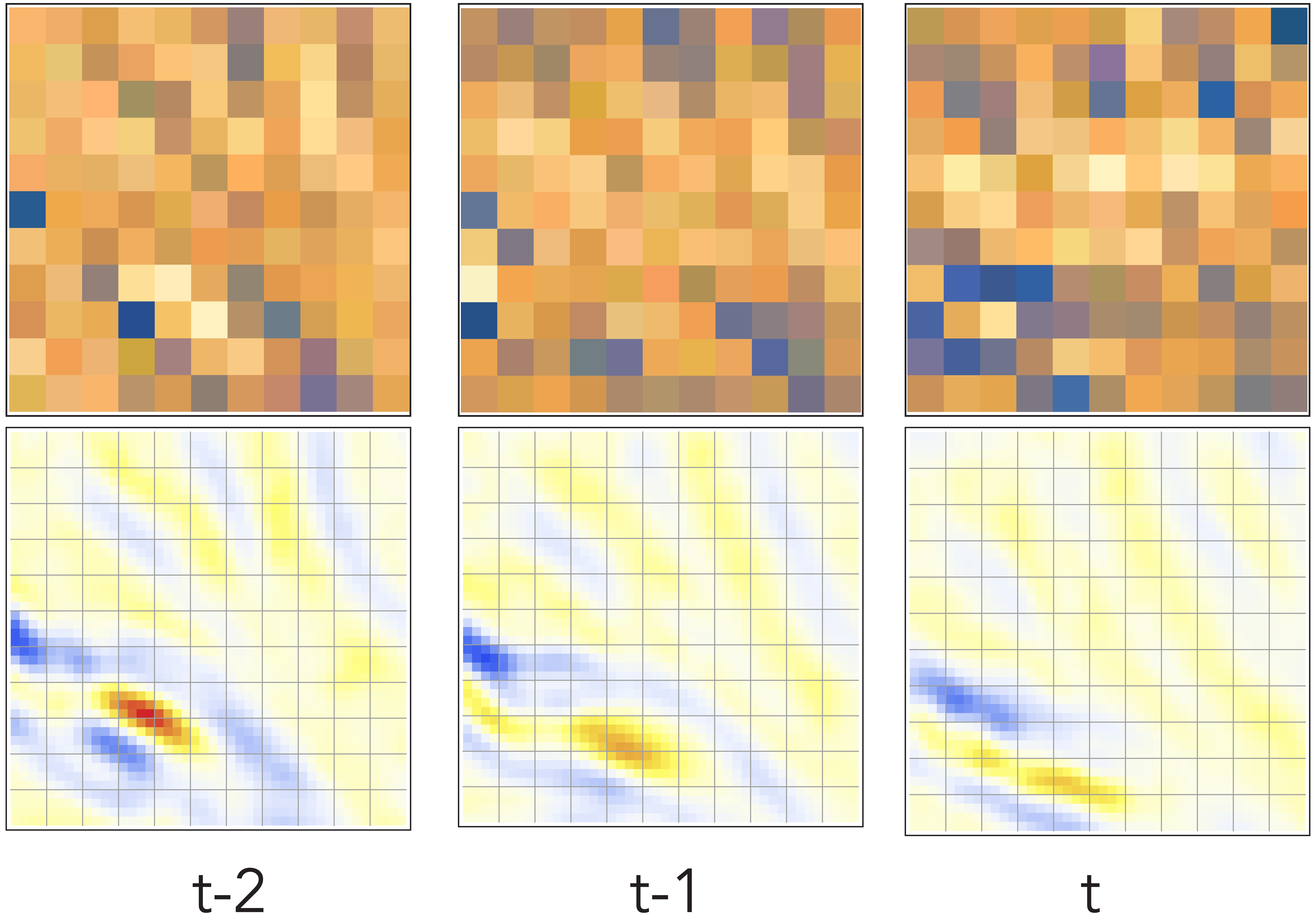
Scientific machine learning: use constraints from known models (e.g. symmetries) in the machine learning model

Observational data: avoid the (inductive) biases and constraints we have in analytic models in the learning

Attention maps:

- ensure physical validity of learned models
- understand physical systems?

Physics and Learning?



unknown physical relations?

AtmoRep

with Ilaria Luise, Maike Sonnewald, Martin Schultz, Aneesh Subramanian

Summary

- Self-supervised representation learning offers (in my opinion) great potential in the Earth sciences
  - Large amounts of unlabeled data (and fast growing)
  - Labeled data is scarce and difficult to obtain

Summary

- Self-supervised representation learning offers (in my opinion) great potential in the Earth sciences
  - Large amounts of unlabeled data (and fast growing)
  - Labeled data is scarce and difficult to obtain
- Representation learning has the potential to provide new insights into spatio-temporal interactions

Self-supervised representation learning

- DINO[1]
  - Self-supervised representation learning for computer vision tasks
  - Vision transformer as neural network
  - Training with unlabeled ImageNet dataset
  - Student-teacher training with virtual prediction task

[1] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. CoRR, 2021.

Self-supervised representation learning

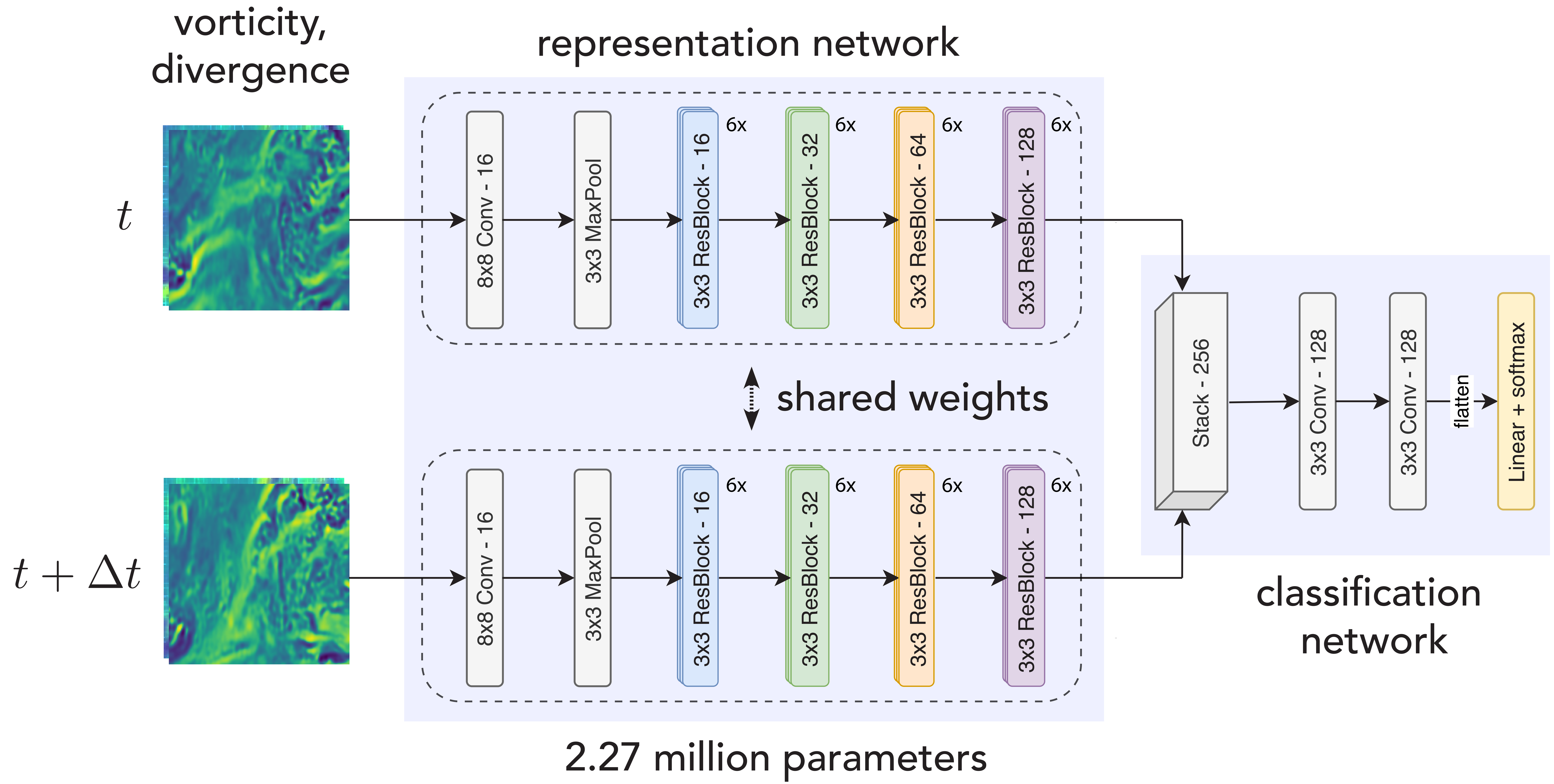
- DINO[1]

	Cifar_{10}	Cifar_{100}	INat_{18}	INat_{19}	Flwrs	Cars	INet
ViT-S/16							
Sup. [69]	**99.0**	89.5	70.7	76.6	98.2	92.1	79.9
DINO	**99.0**	**90.5**	**72.0**	**78.2**	**98.5**	**93.0**	**81.5**
ViT-B/16							
Sup. [69]	99.0	90.8	**73.2**	77.7	98.4	92.1	81.8
DINO	**99.1**	**91.7**	72.6	**78.6**	**98.8**	**93.0**	**82.8**

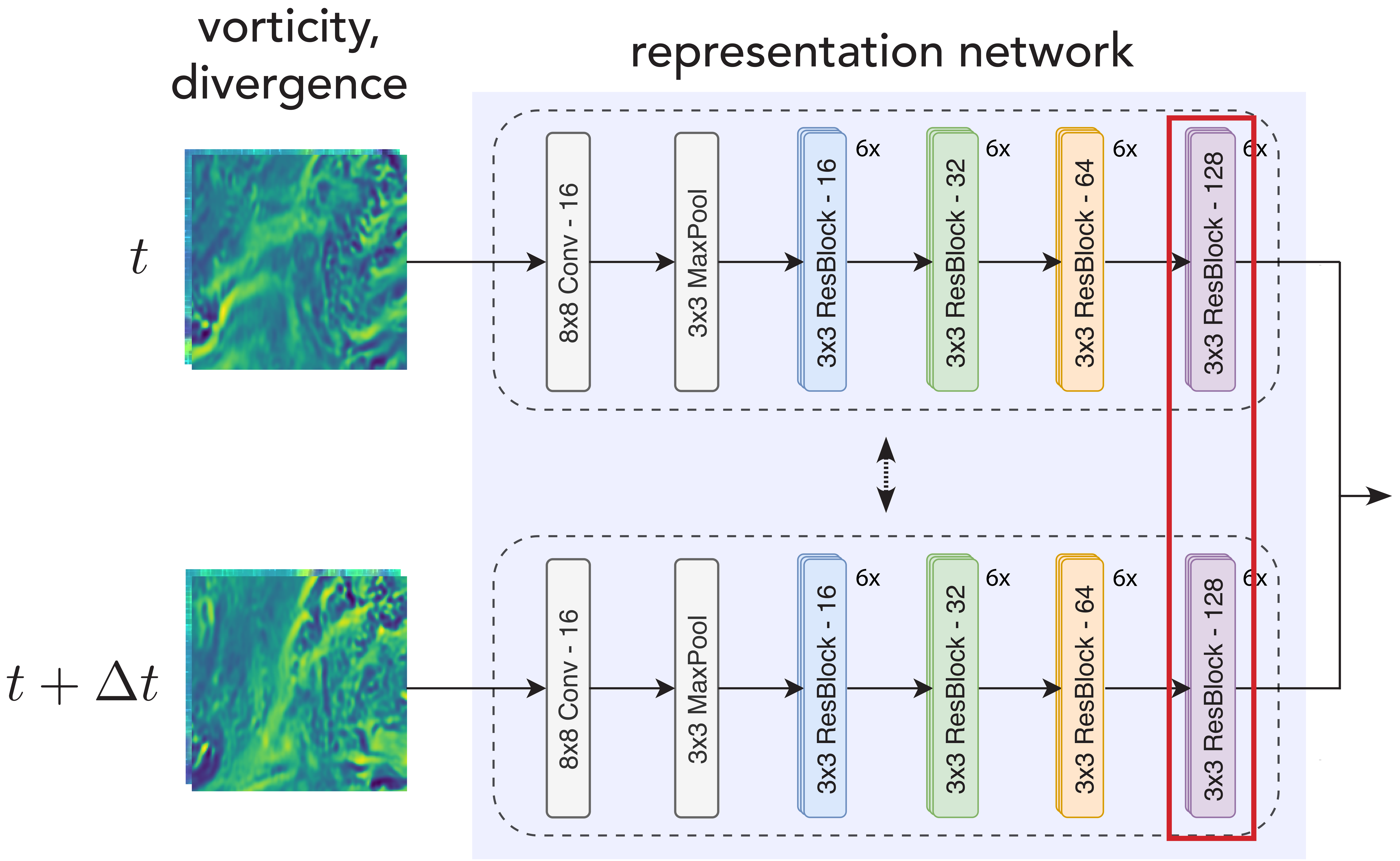
Performance of fine-tuned model on classi-fication

[1] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. CoRR, 2021.

AtmoDist: network



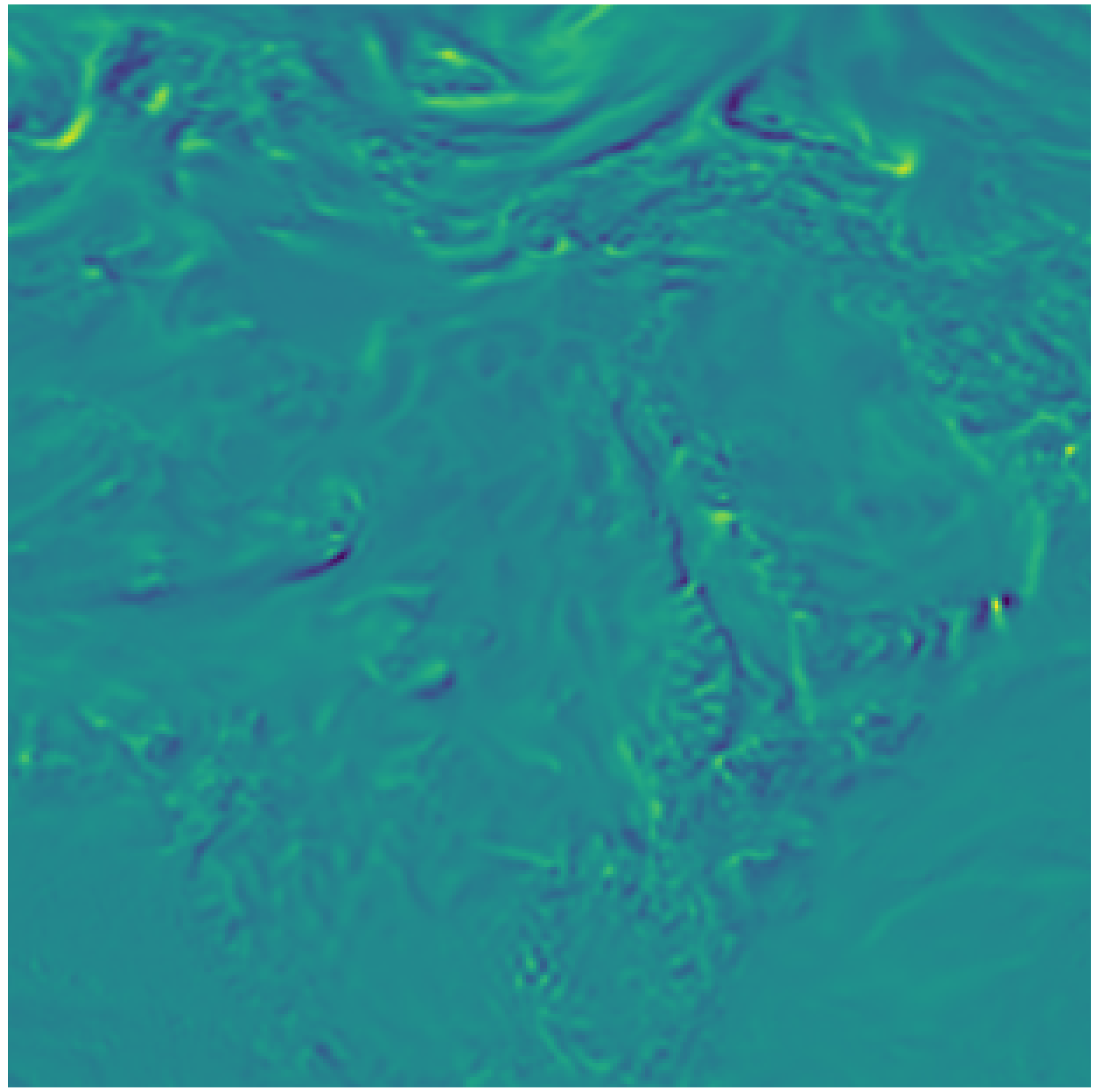
AtmoDist: network



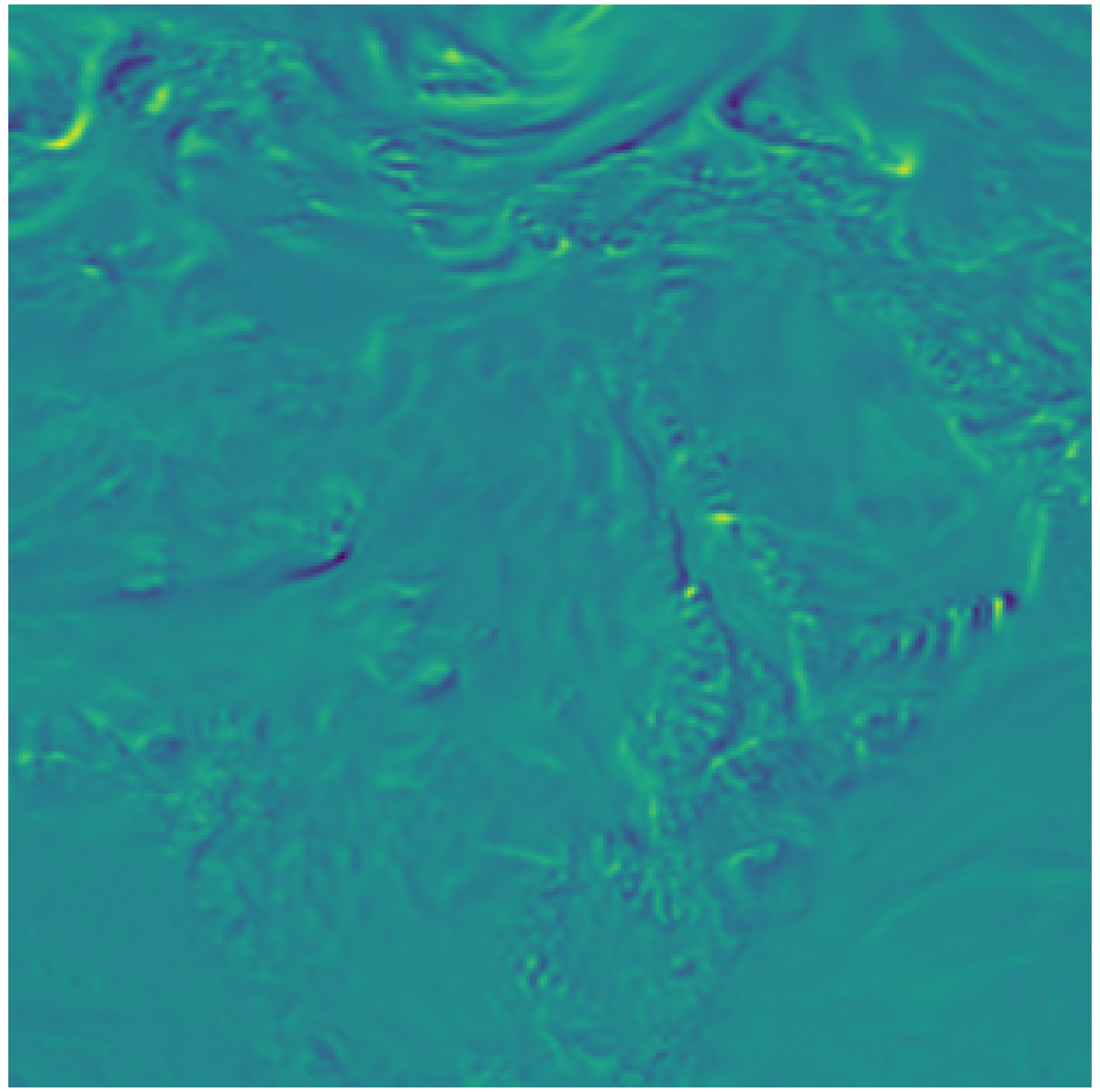
$$\left\| h_J(\zeta_t) - h_J(\zeta_{t+\Delta t}) \right\|_2^2$$

Super-resolution using AtmoDist

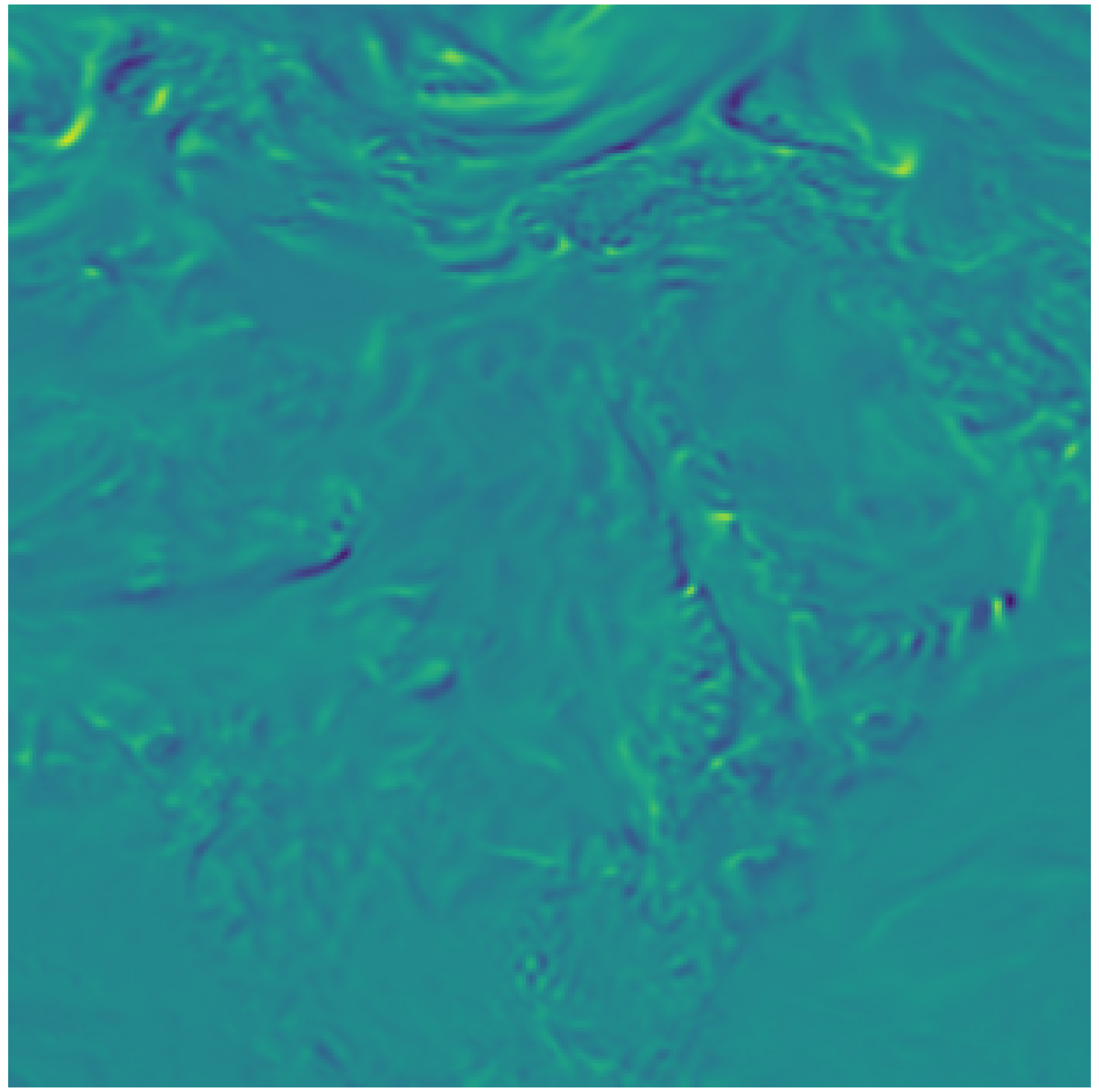
ours



ground thruth

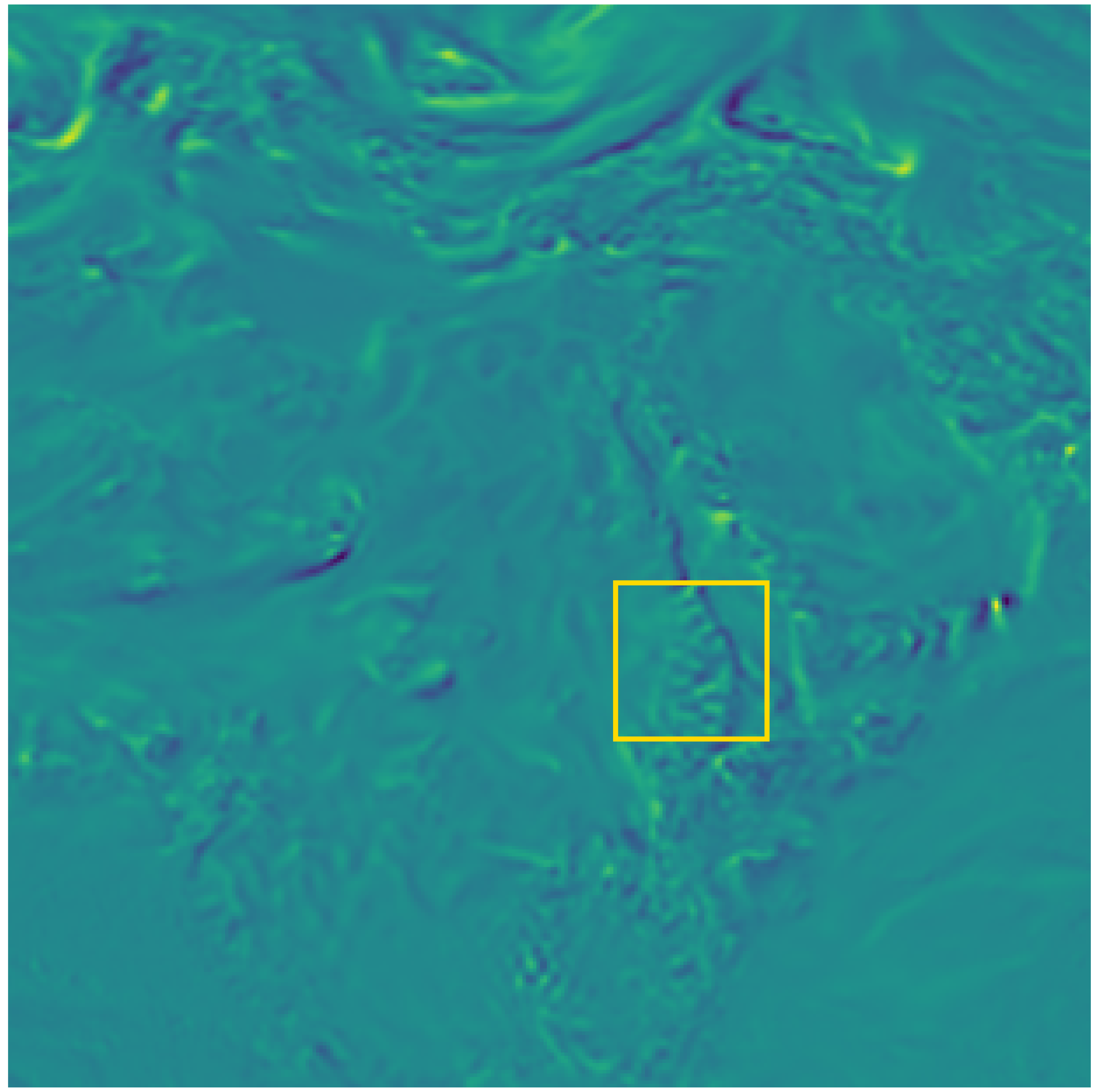


mse

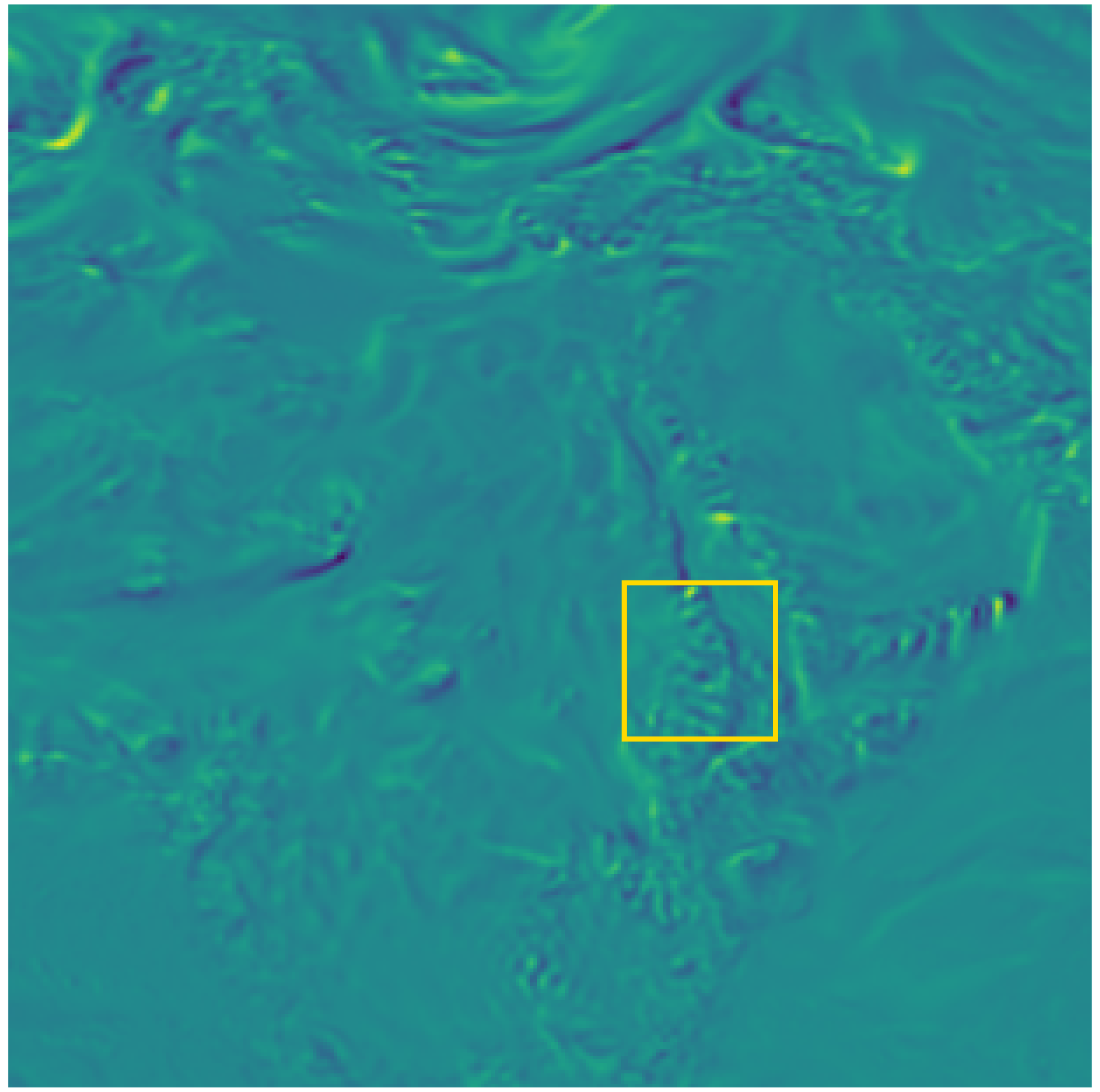


Super-resolution using AtmoDist

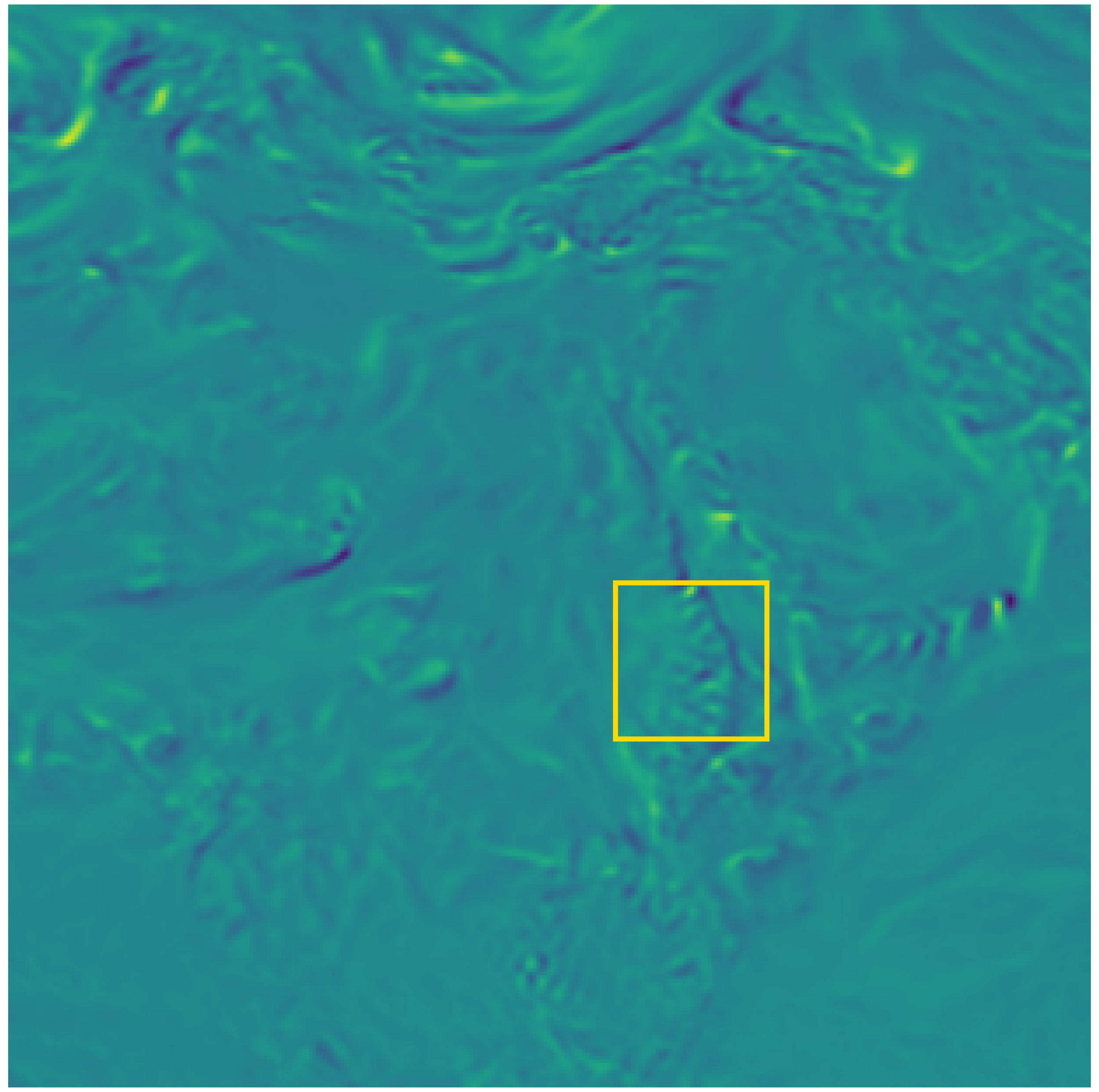
ours



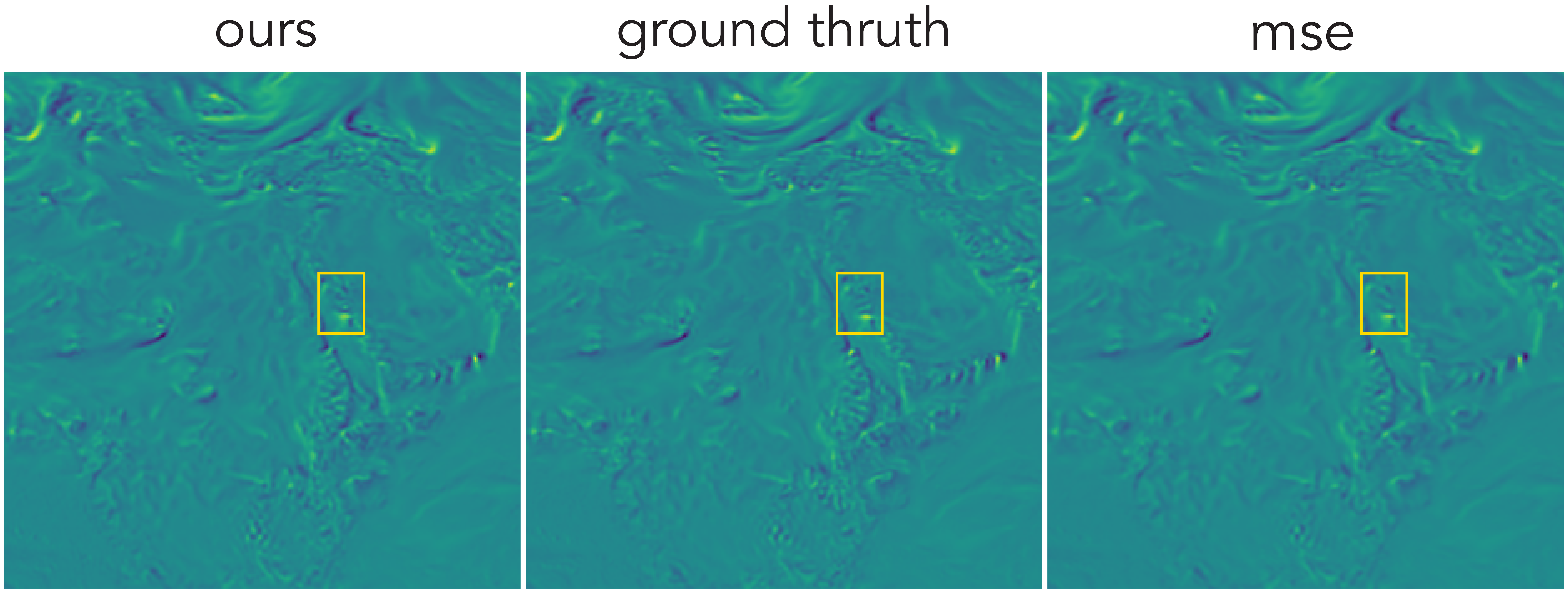
ground thruth



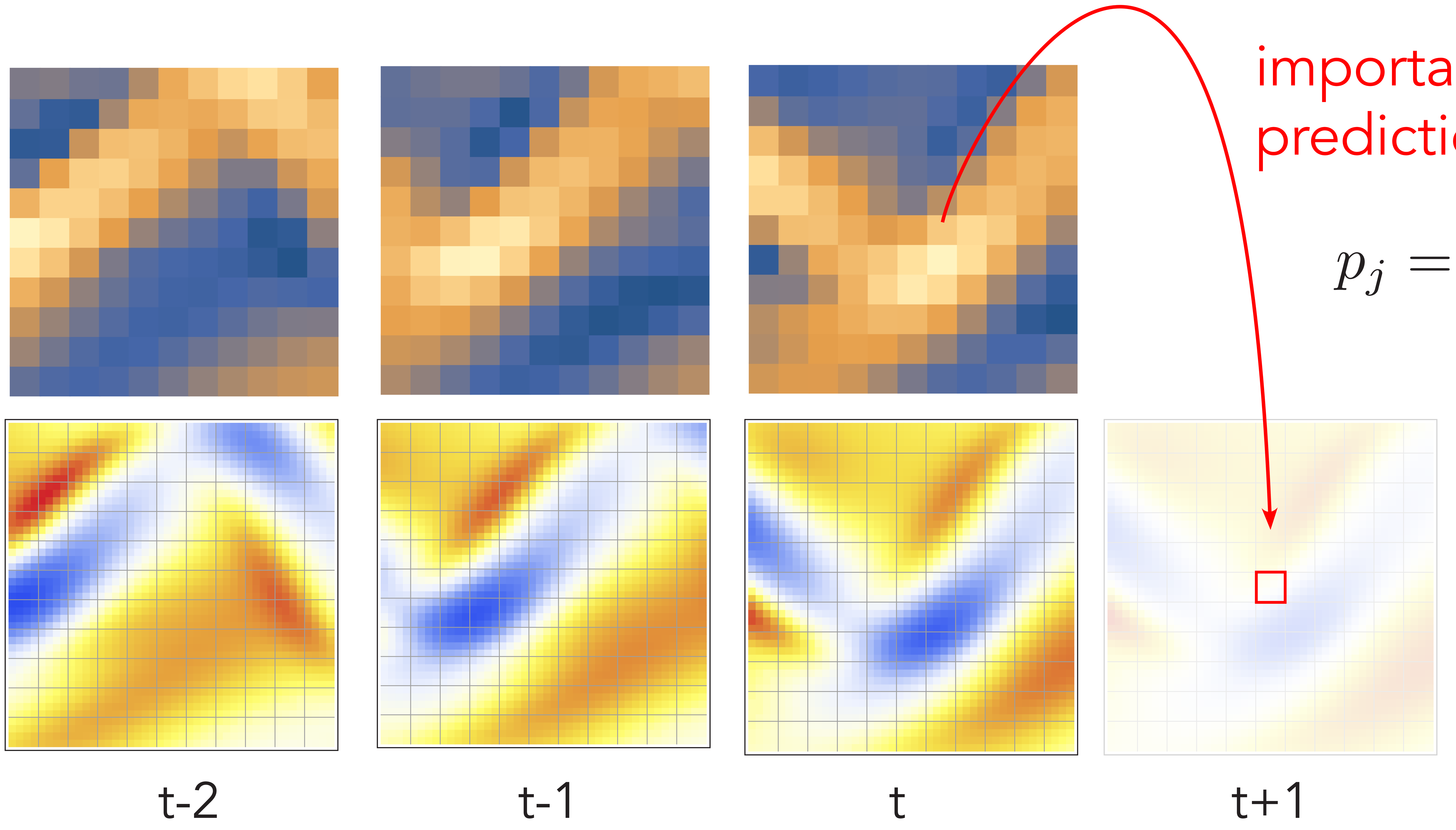
mse



Super-resolution using AtmoDist



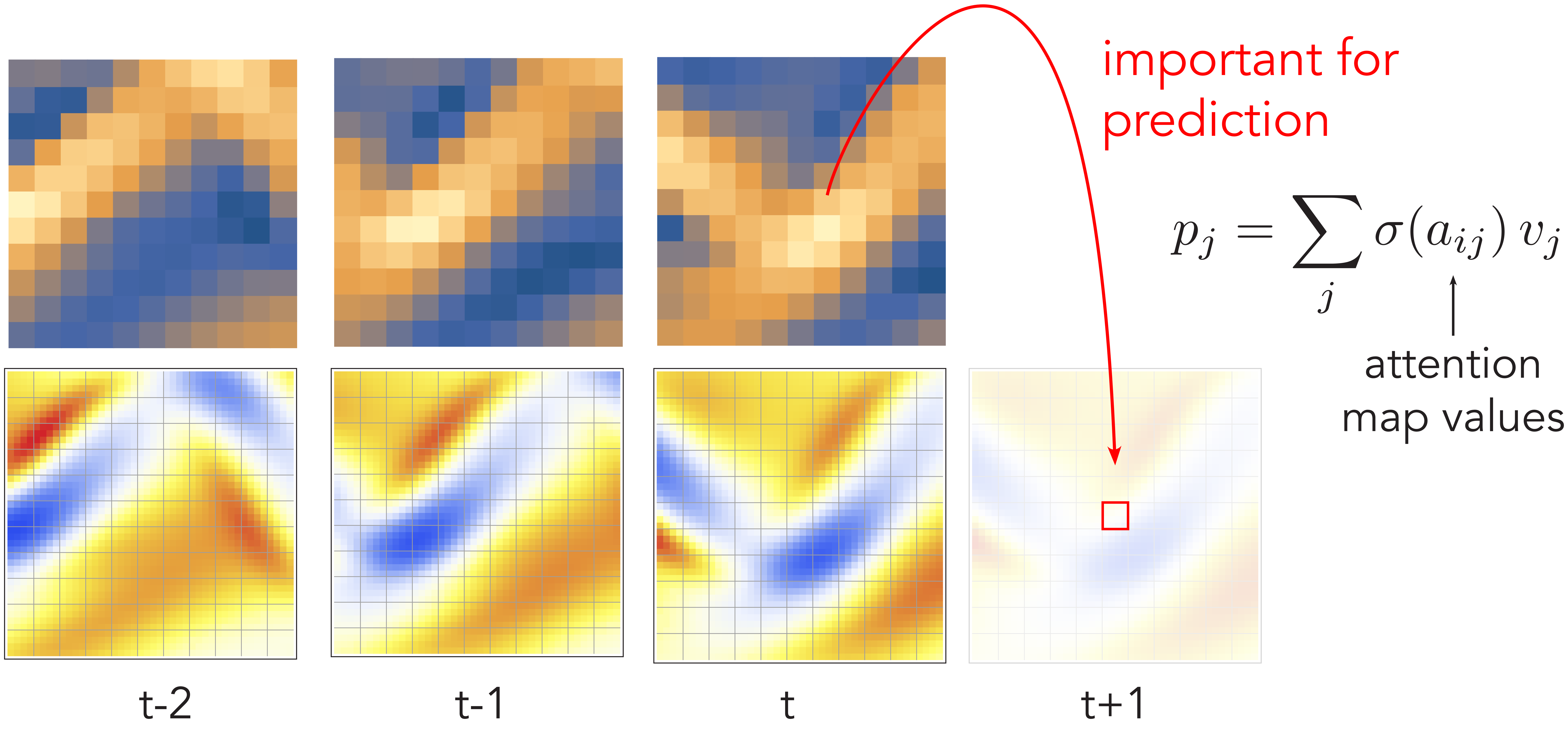
Fluid flow



important for prediction

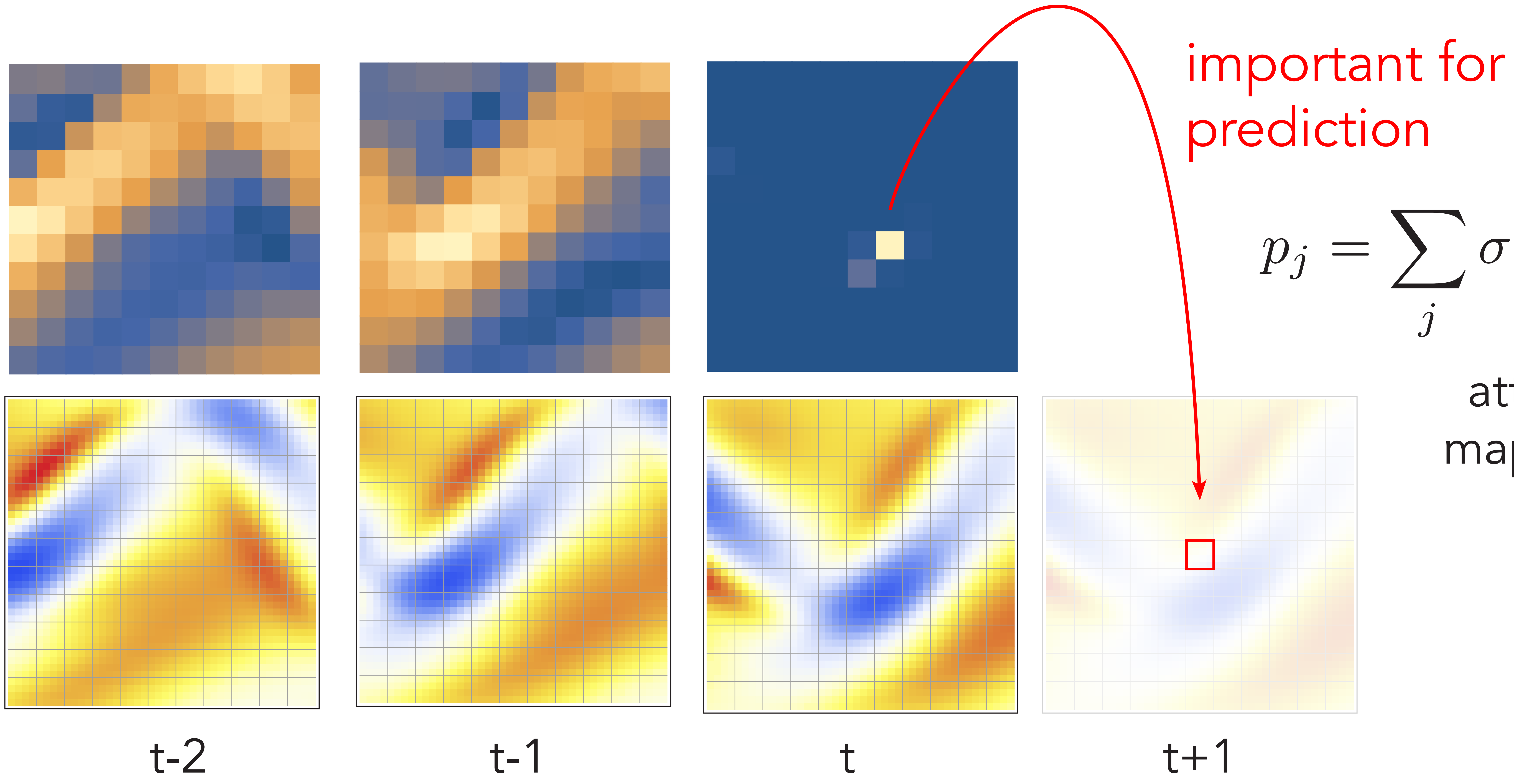
$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

Fluid flow



$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

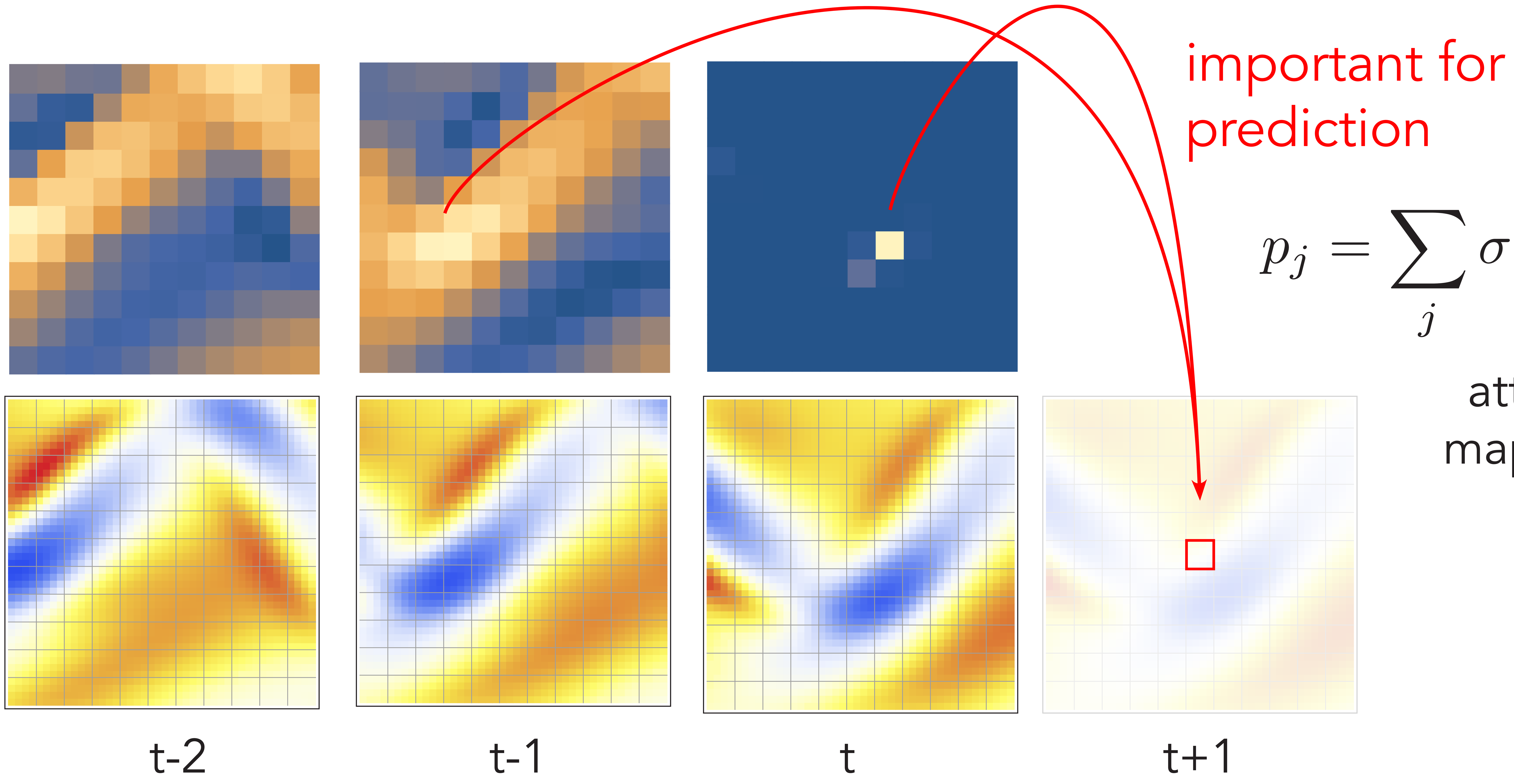
Fluid flow



$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

attention map values

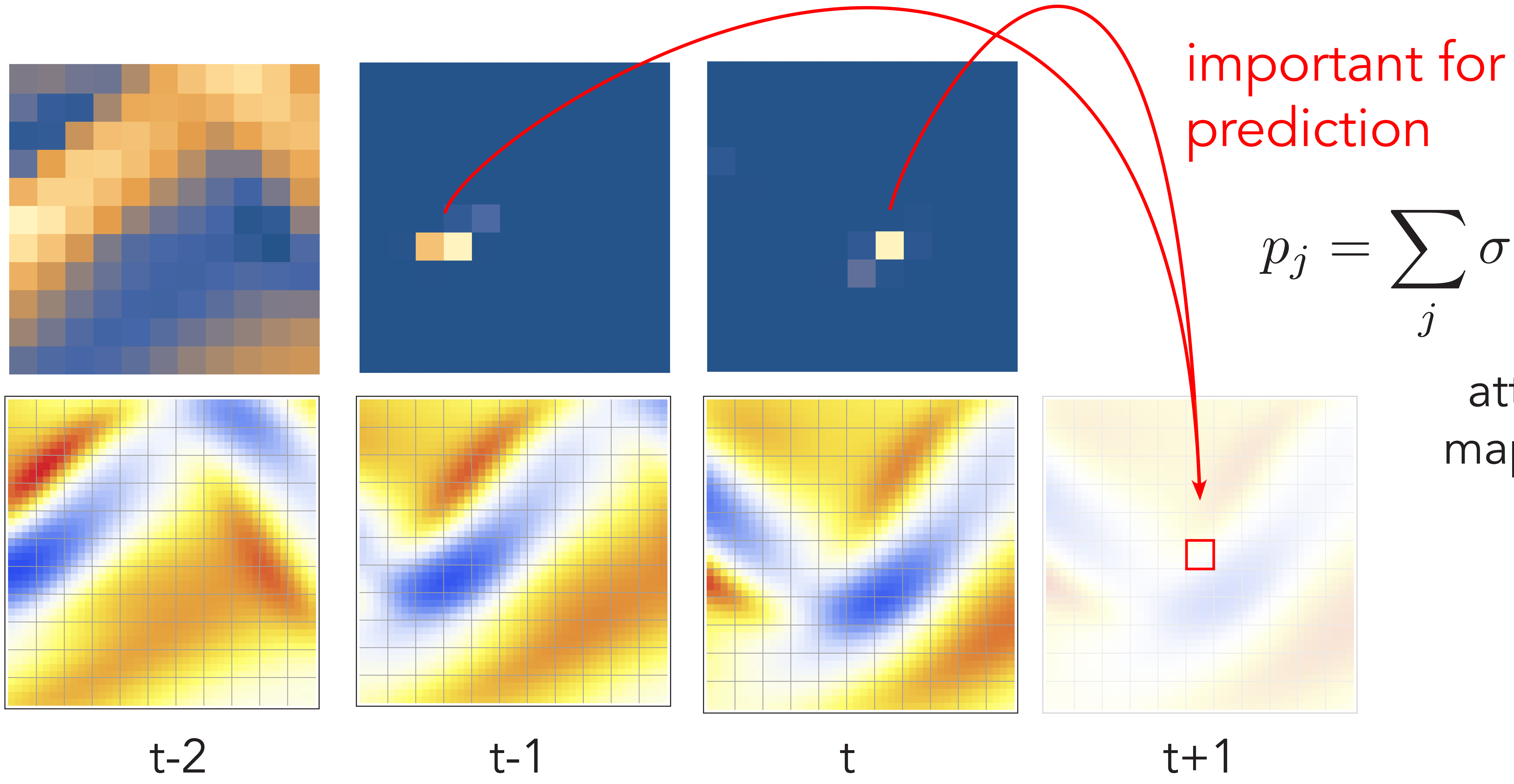
Fluid flow



$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

attention map values

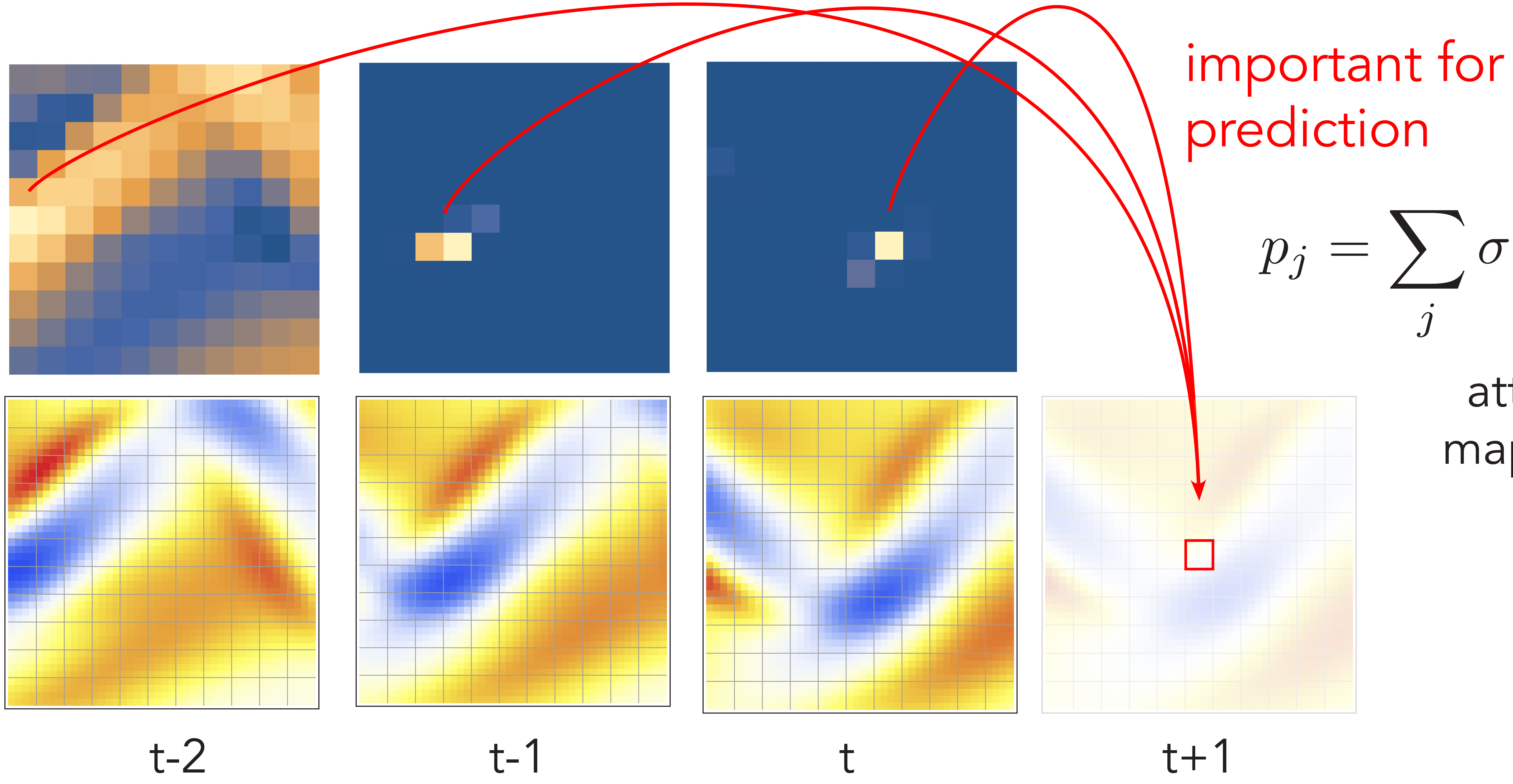
Fluid flow



$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

attention map values

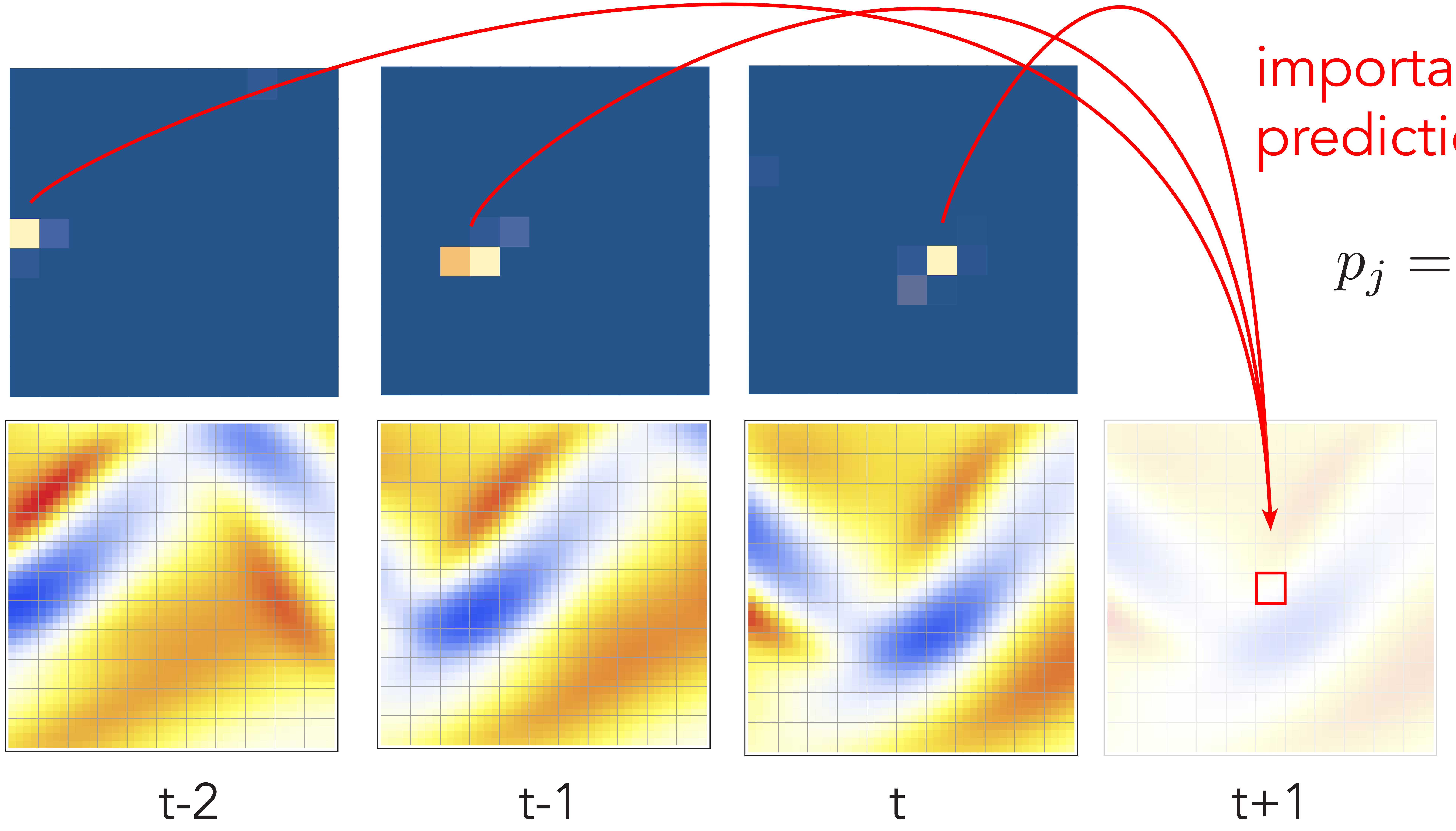
Fluid flow



$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

attention map values

Fluid flow

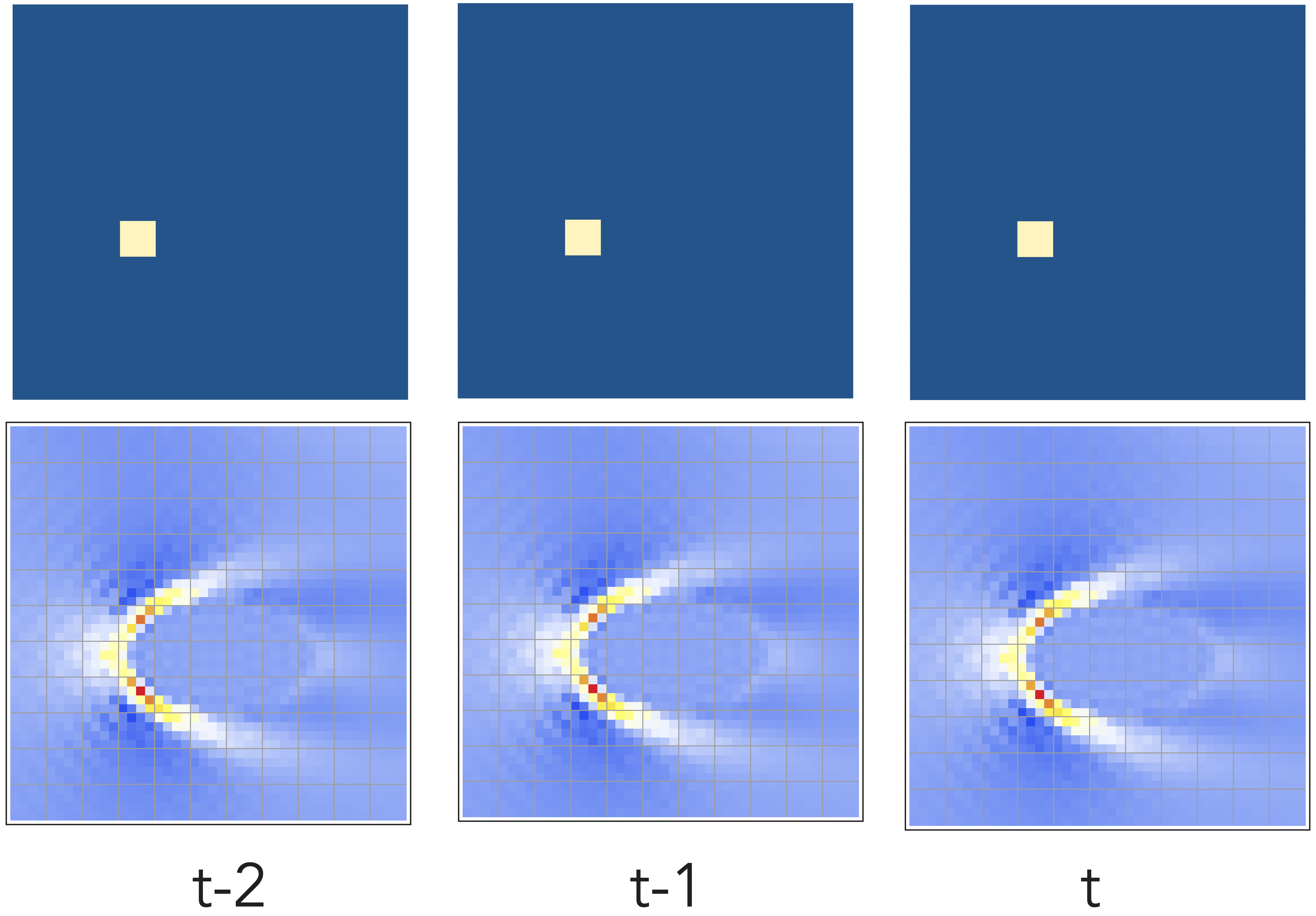


important for prediction

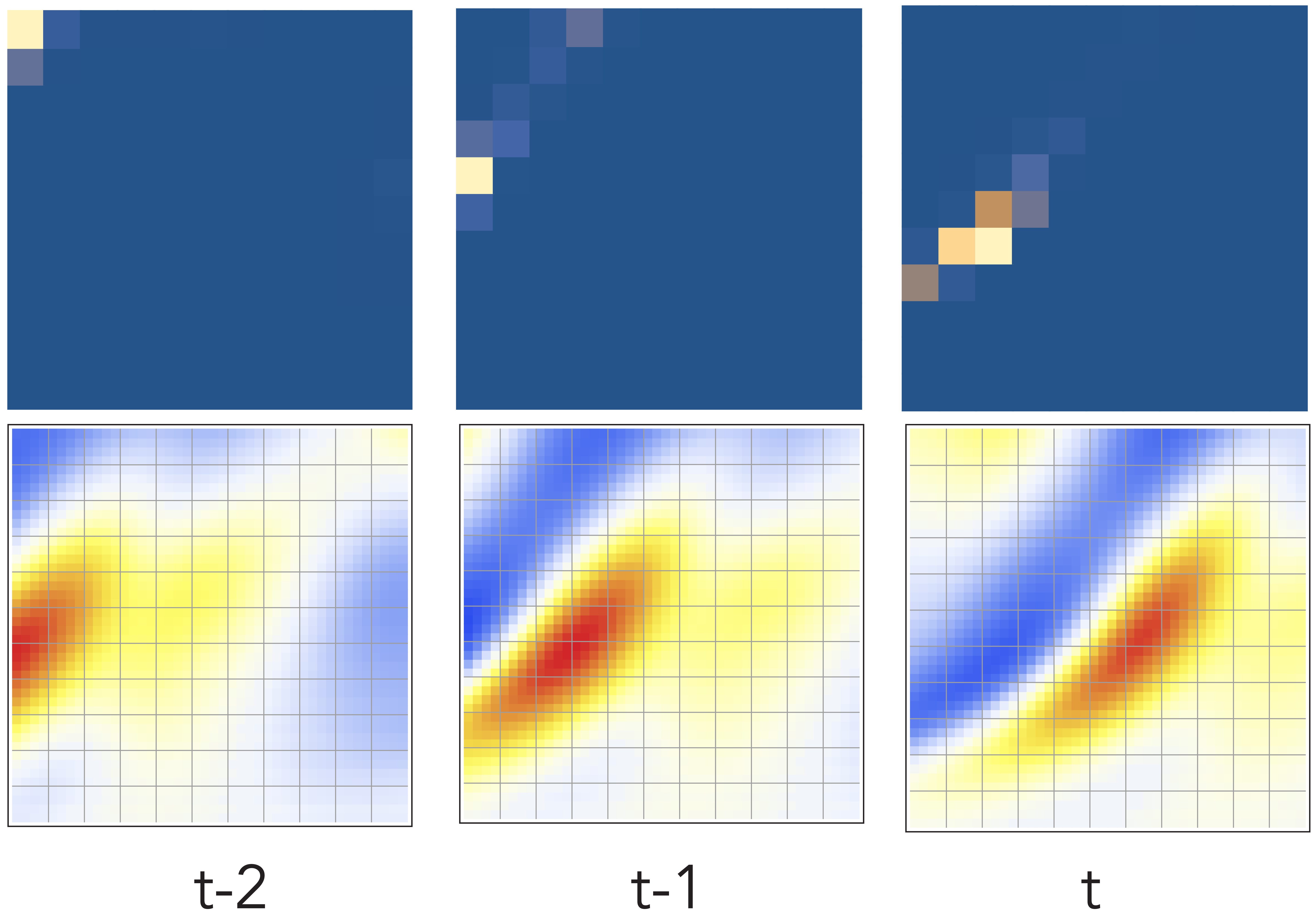
$$p_j = \sum_j \sigma(a_{ij})\, v_j$$

attention map values

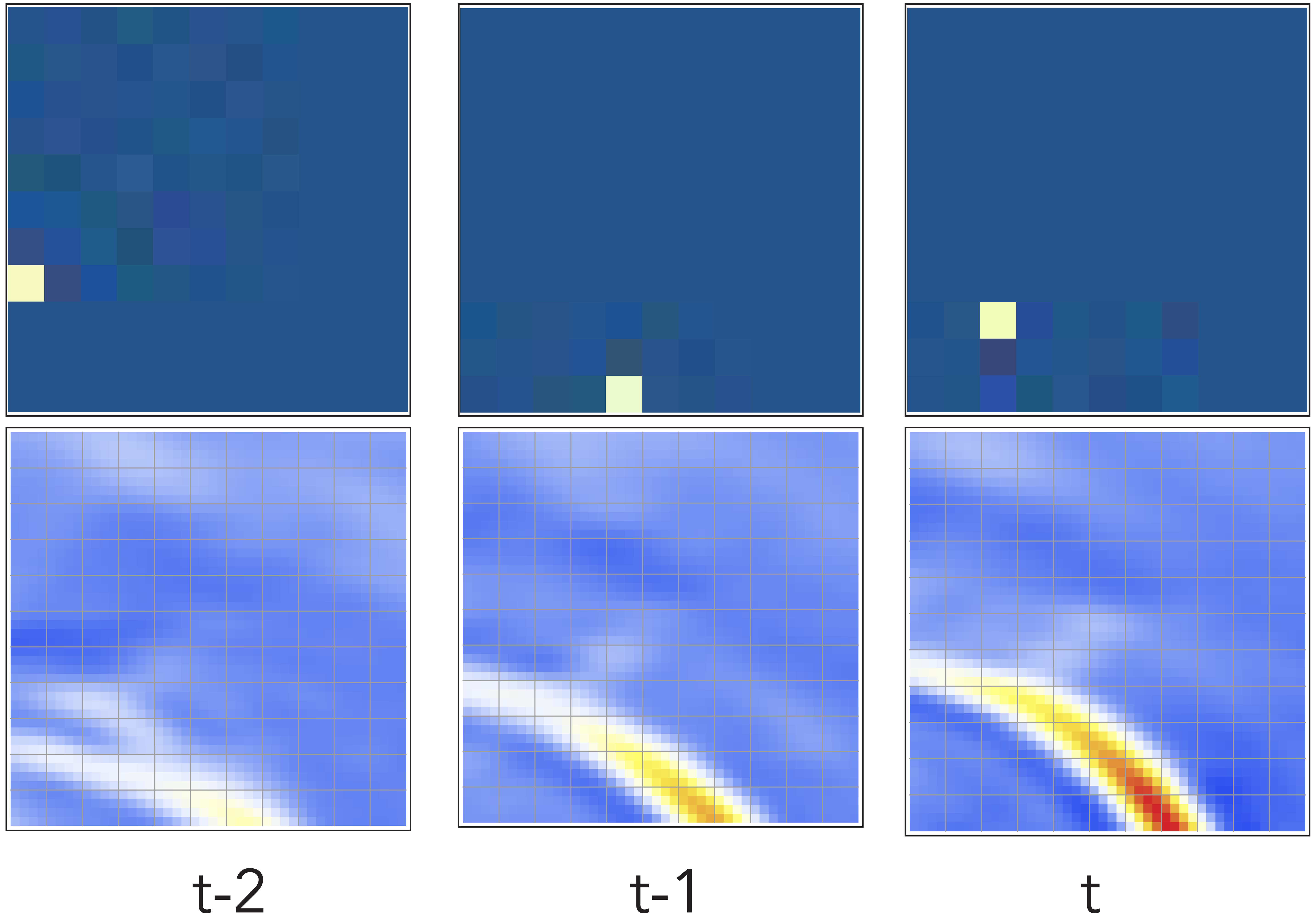
Fluid flow



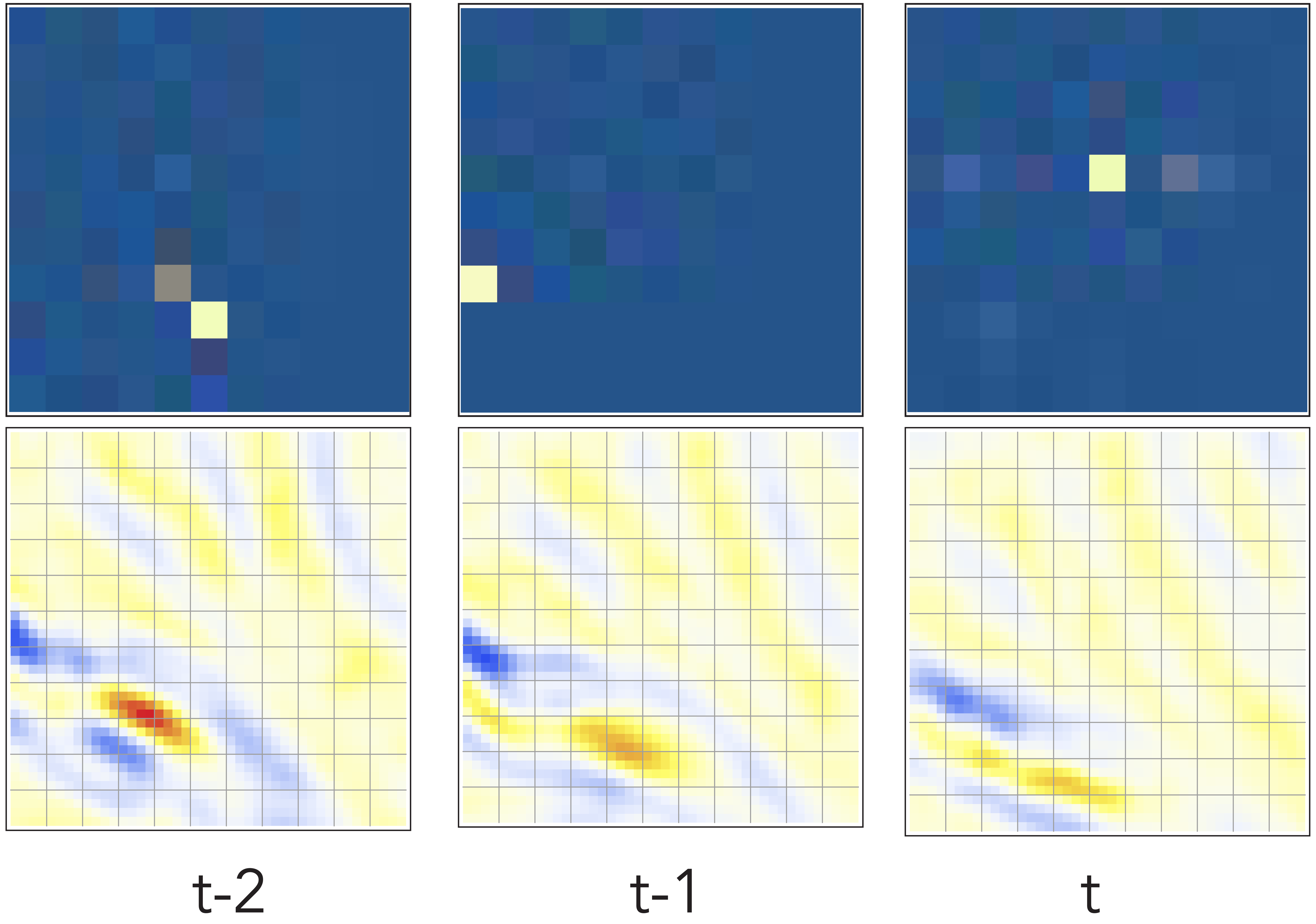
Fluid flow



Atmospheric vorticity (ERA5)



Atmospheric vorticity (ERA5)



AtmoDist: evaluation