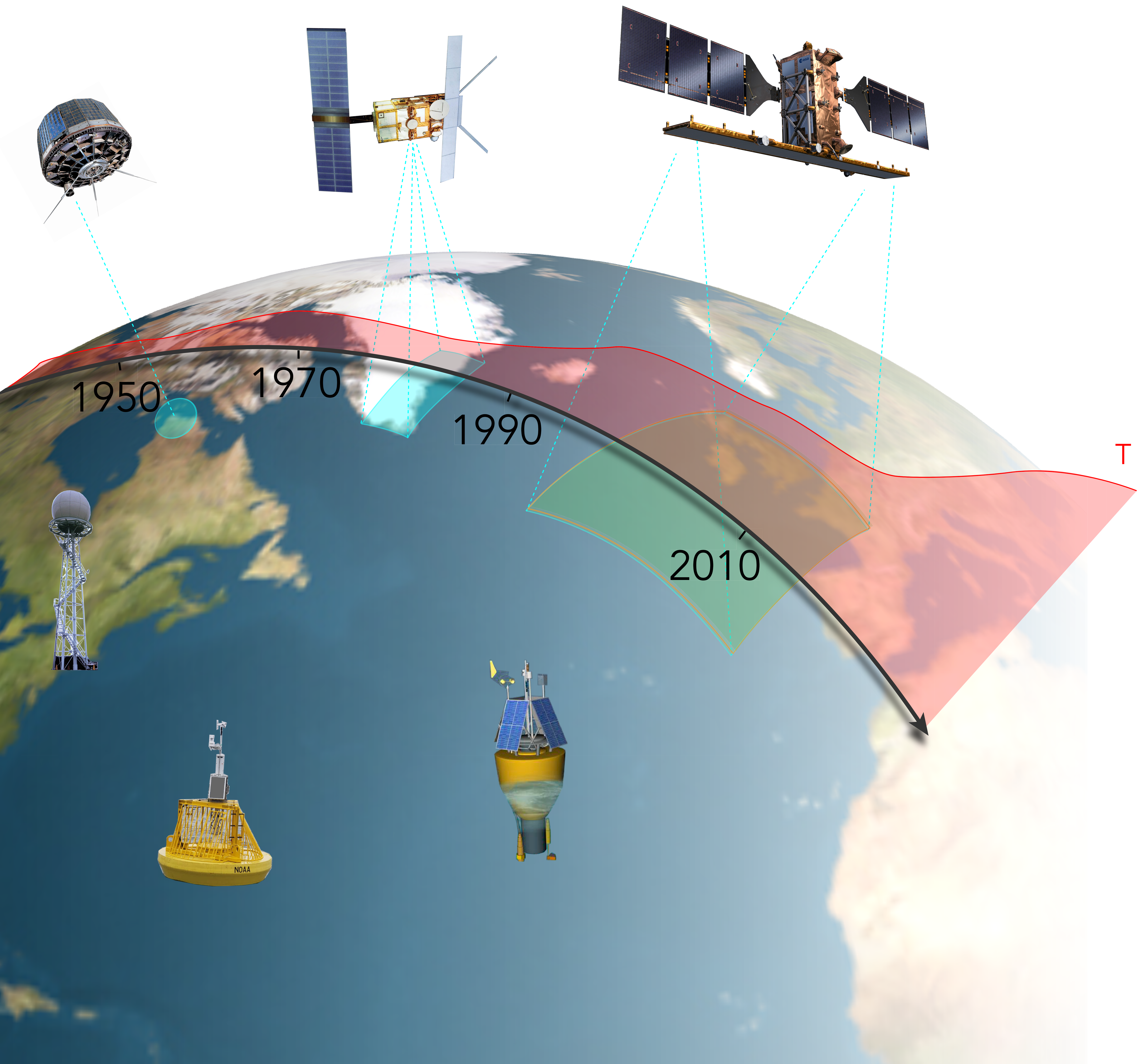
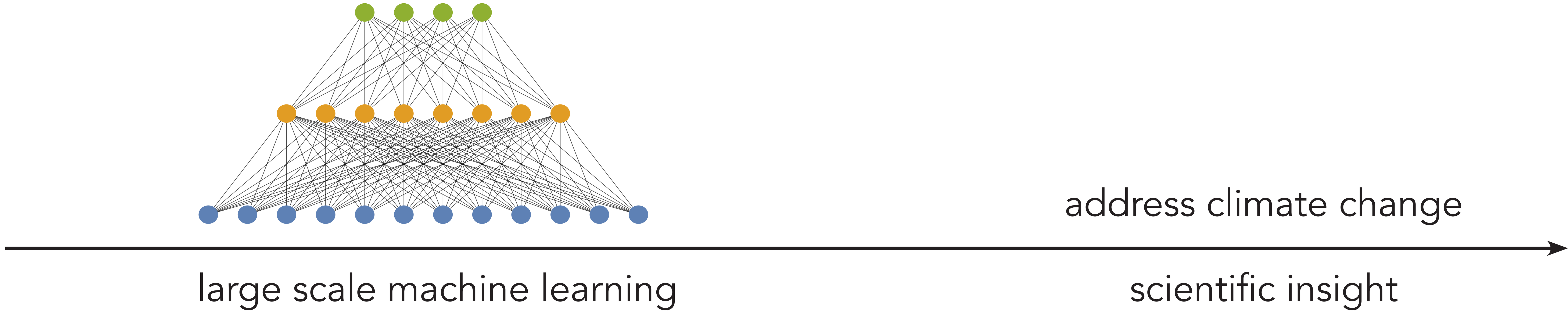
AtmoRep

Large scale representation learning of atmospheric dynamics

Christian Lessig, Ilaria Luise, Martin Schultz





Representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications

Representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications
  - Representation network provides transformation of network input to effective feature spaces
  - Self-supervised training on very large amounts of data (O(PB)) with very large networks ($O(10^9)$ parameters)
  - Useful for downstream applications using tail network or fine-tuning

Can we perform representation learning in the Earth sciences?

Representation learning for the Earth sciences?

- Very large amounts of observational data
  - ERA5 reanalysis: 6+ PB
  - ESA's MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

Representation learning for the Earth sciences?

- Very large amounts of **observational** data
  - ERA5 reanalysis: 6+ PB
  - ESA's MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

Representation learning for the Earth sciences?

- Very large amounts of **observational** data
- No complete classical model for system and dynamics
  - Central issue for forecasting and climate projections

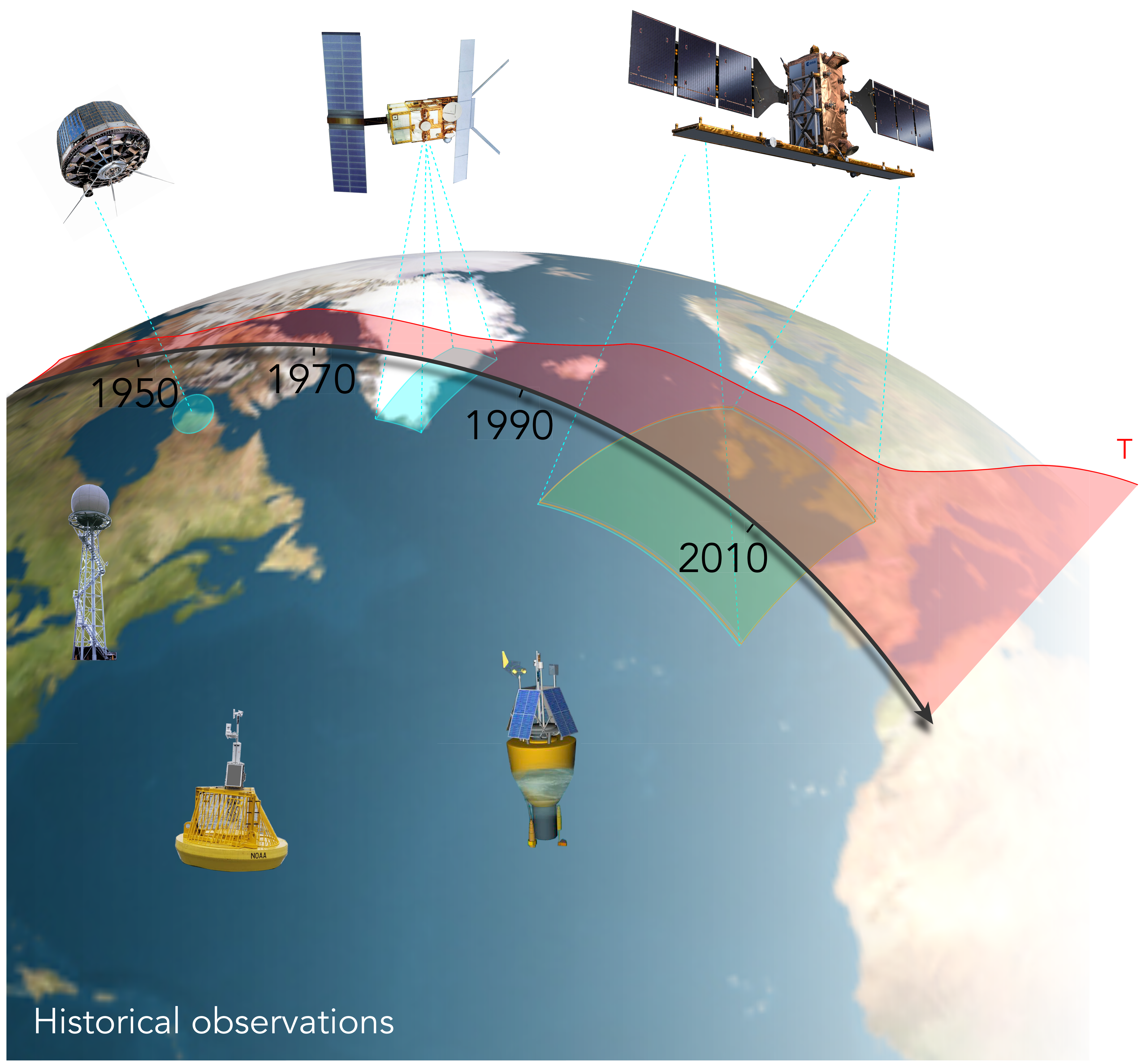
Representation learning for the Earth sciences?

- Very large amounts of **observational** data
- No complete classical model for system and dynamics
- Chaosticisty in atmospheric dynamics leads to ambiguity
  - There is often not one “correct answer“
  - Large networks learn statistical representations

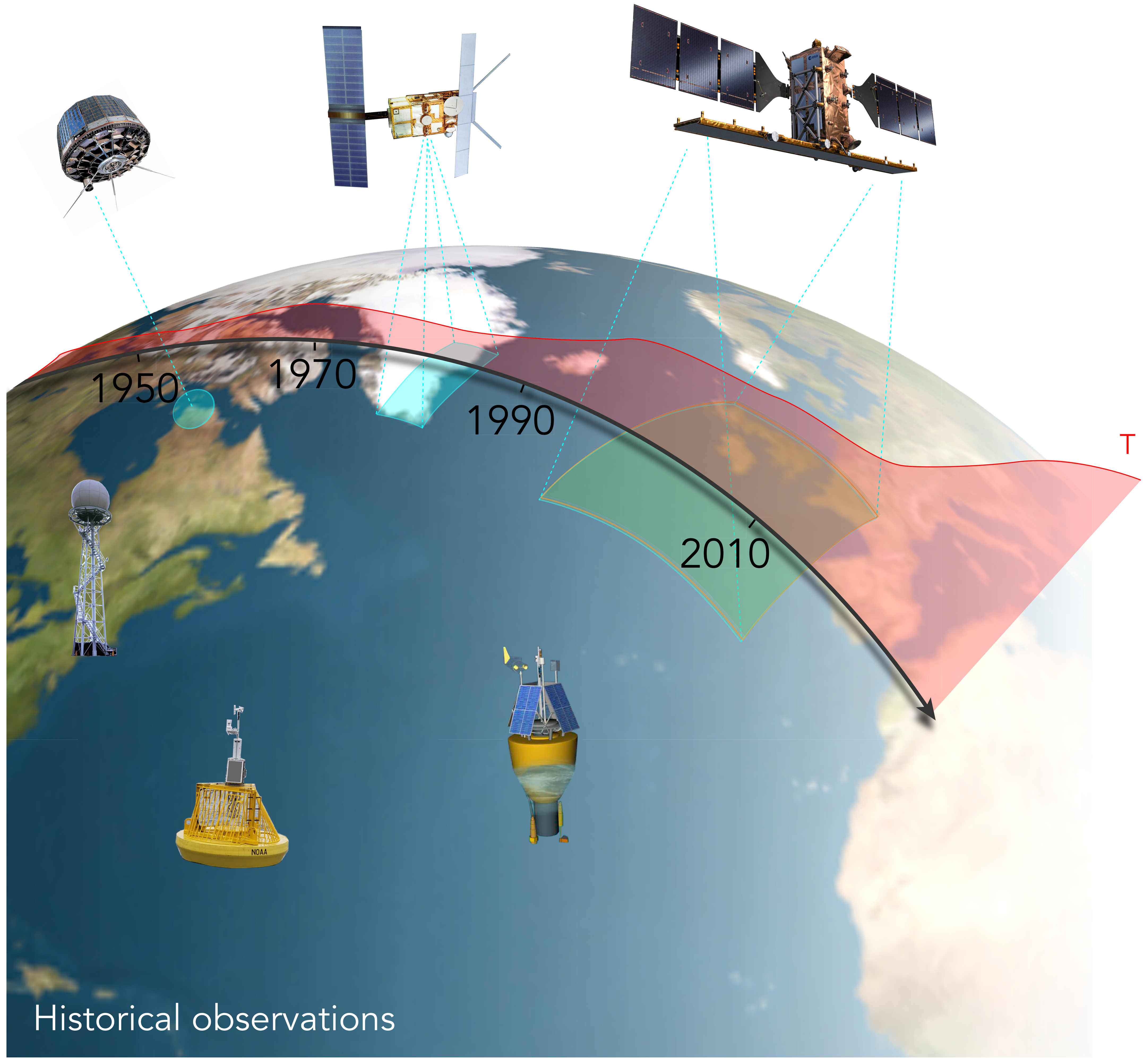
AtmoRep

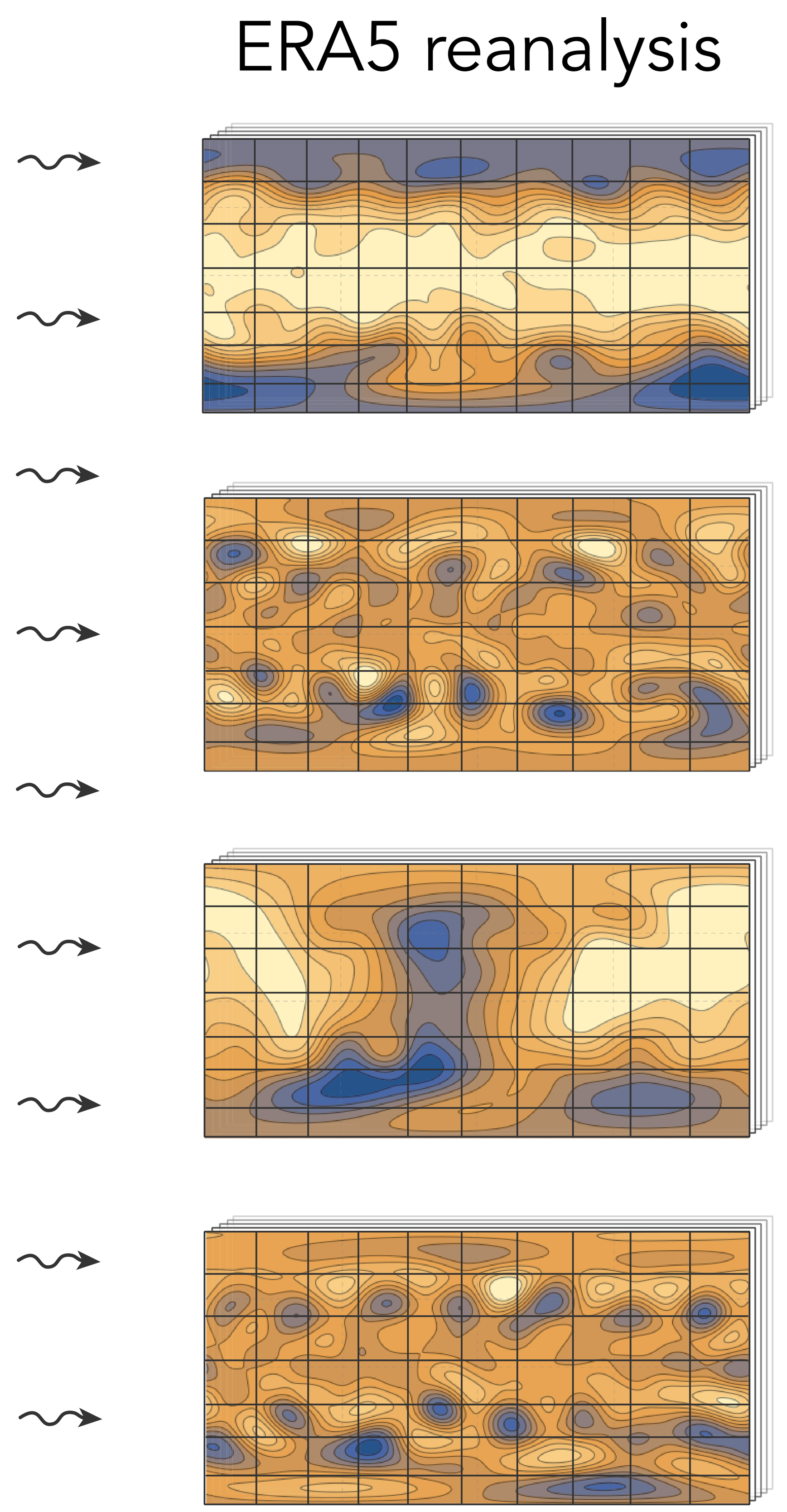
Large scale representation learning of atmospheric dynamics

AtmoRep

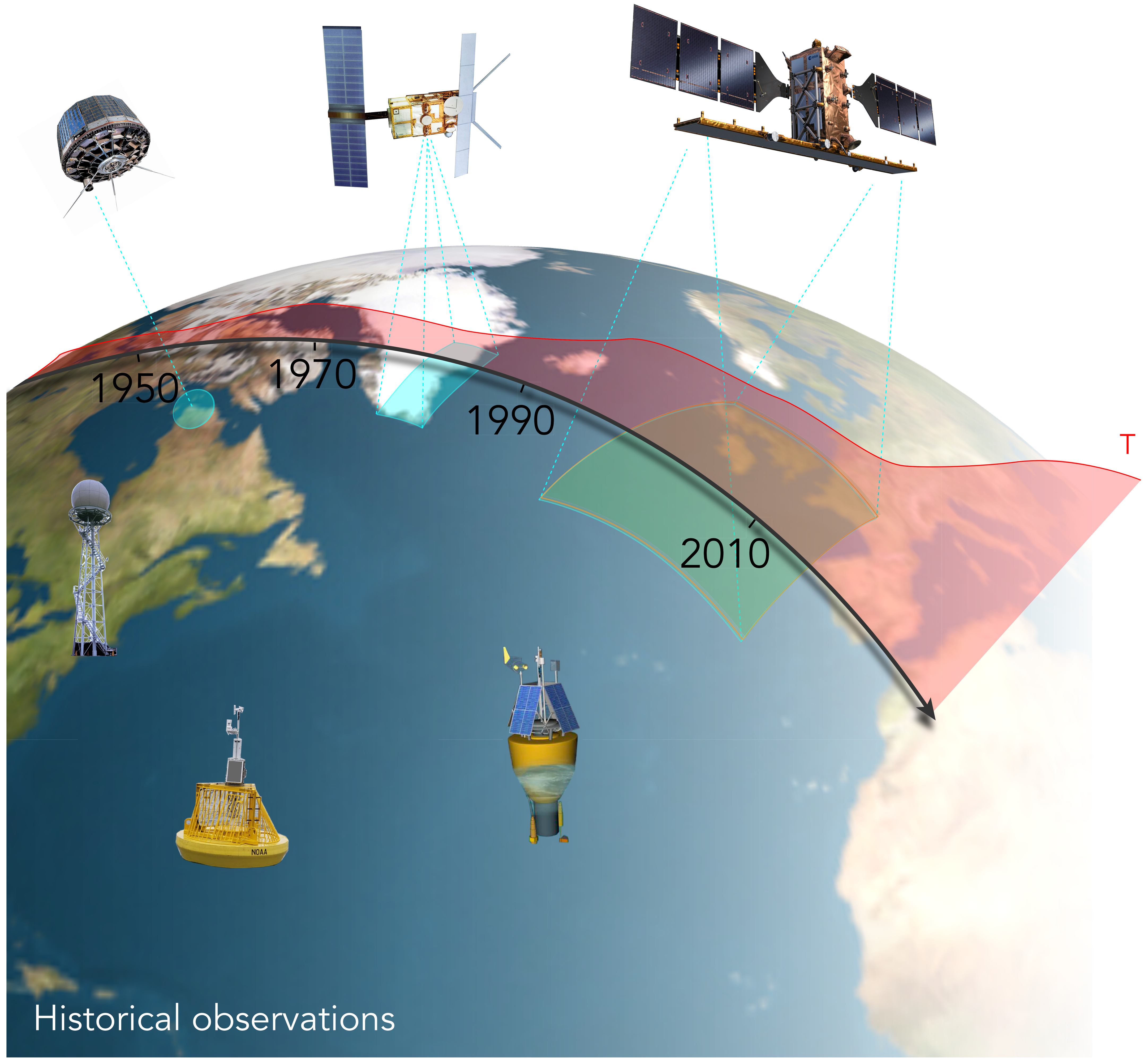


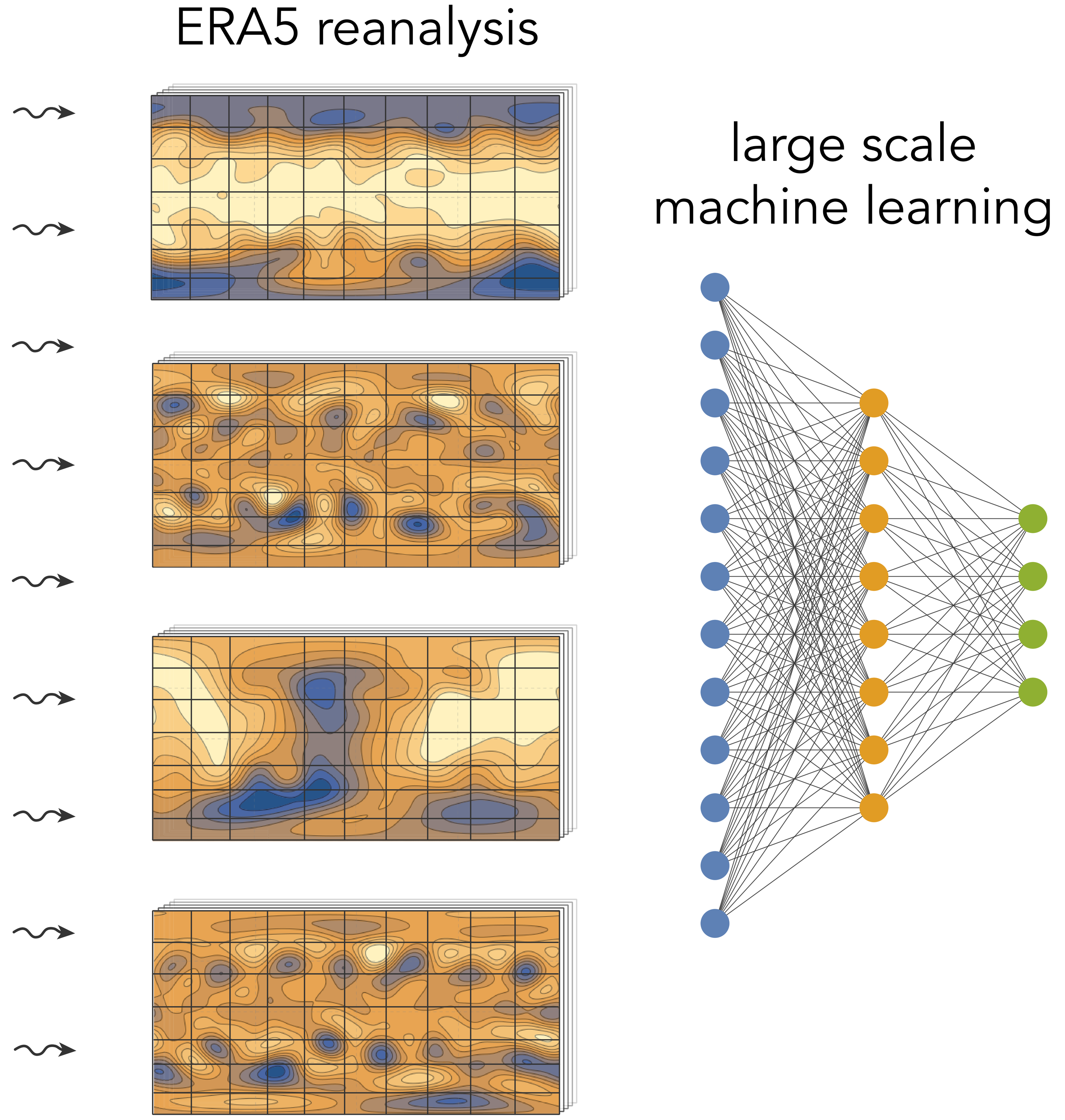
AtmoRep


1950
1970
1990
2010
T
Historical observations

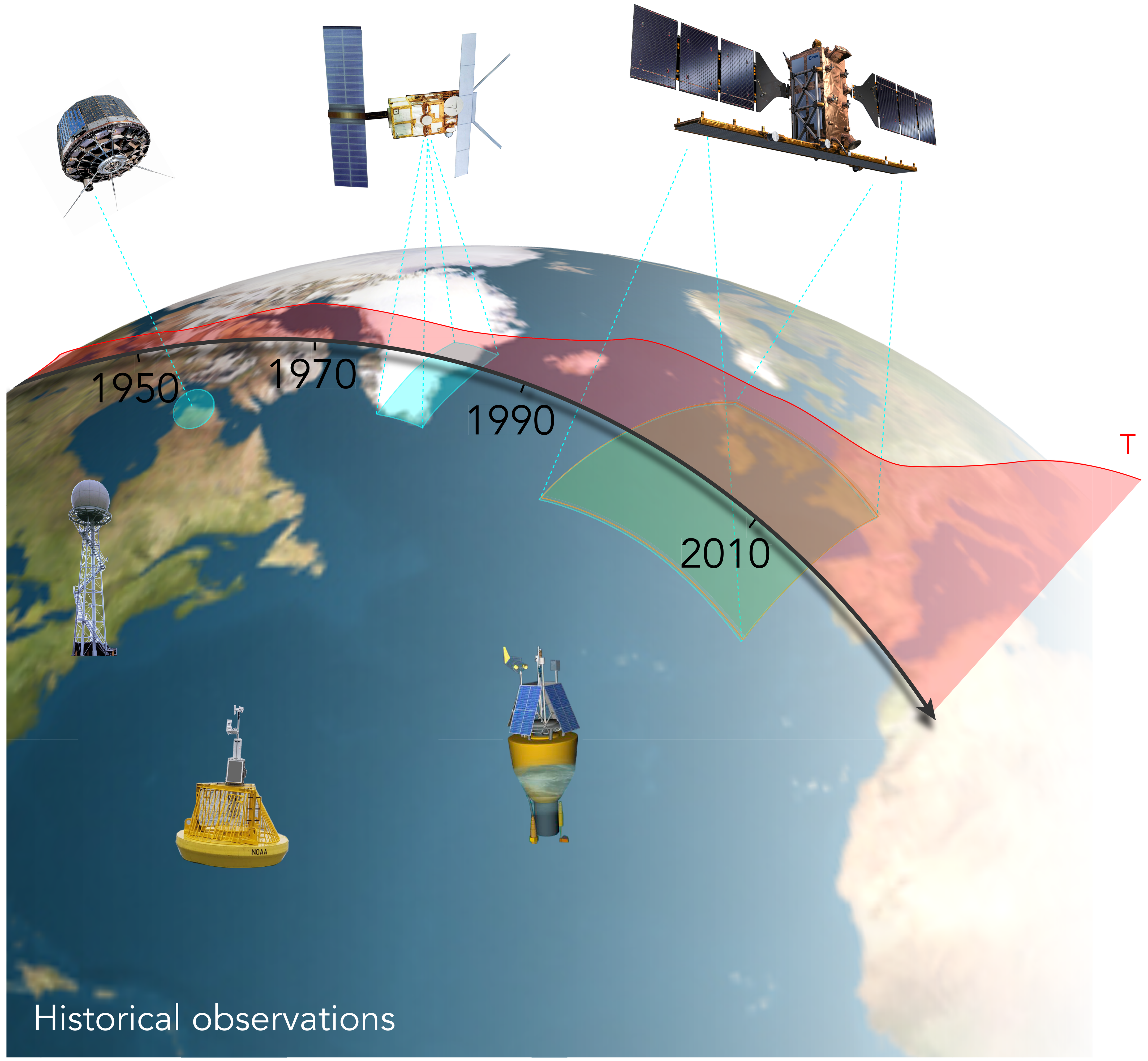

ERA5 reanalysis

AtmoRep

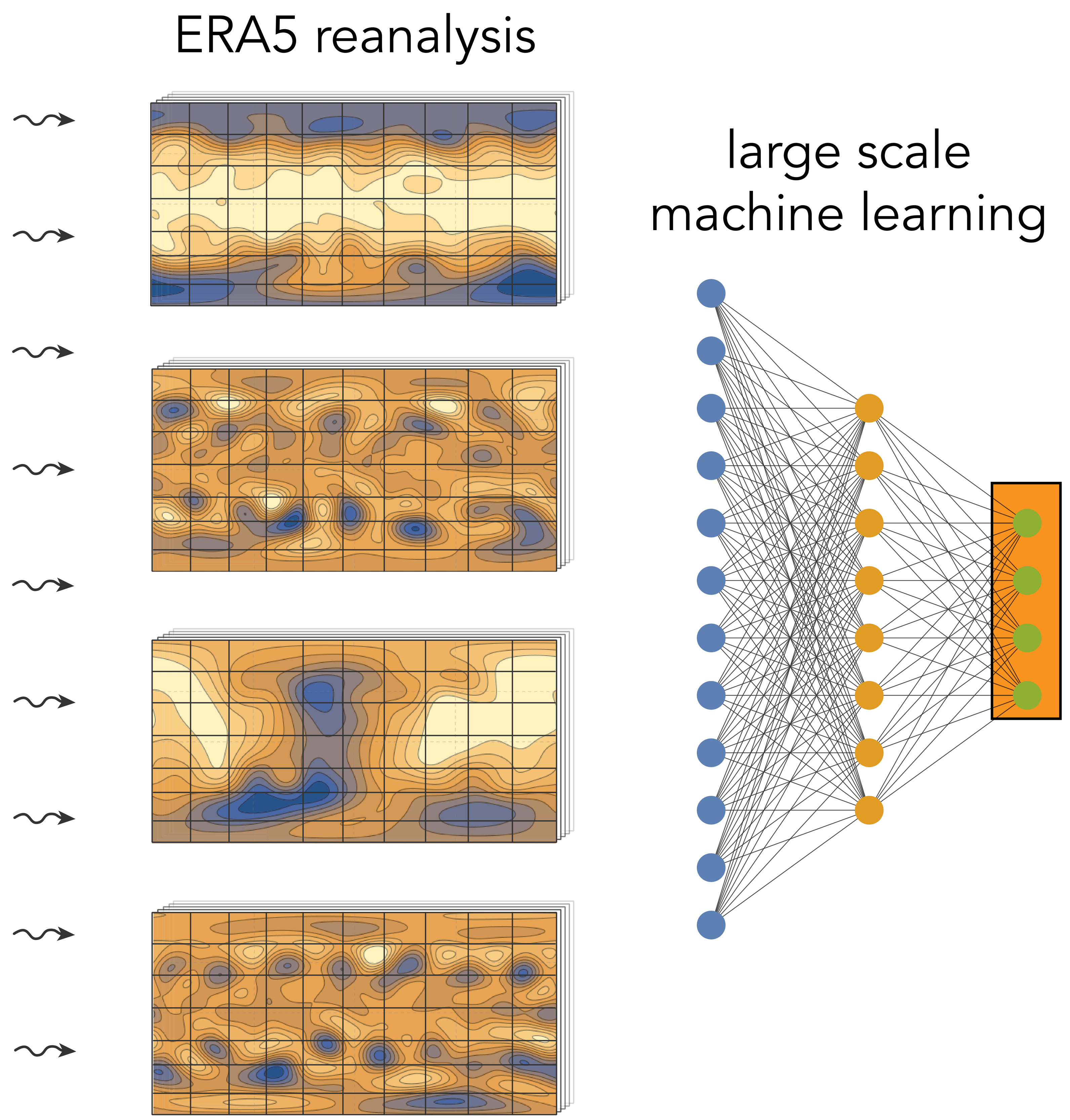

1950
1970
1990
2010
T
Historical observations


ERA5 reanalysis
large scale
machine learning

AtmoRep



1950
1970
1990
2010
T
Historical observations



ERA5 reanalysis
large scale
machine learning

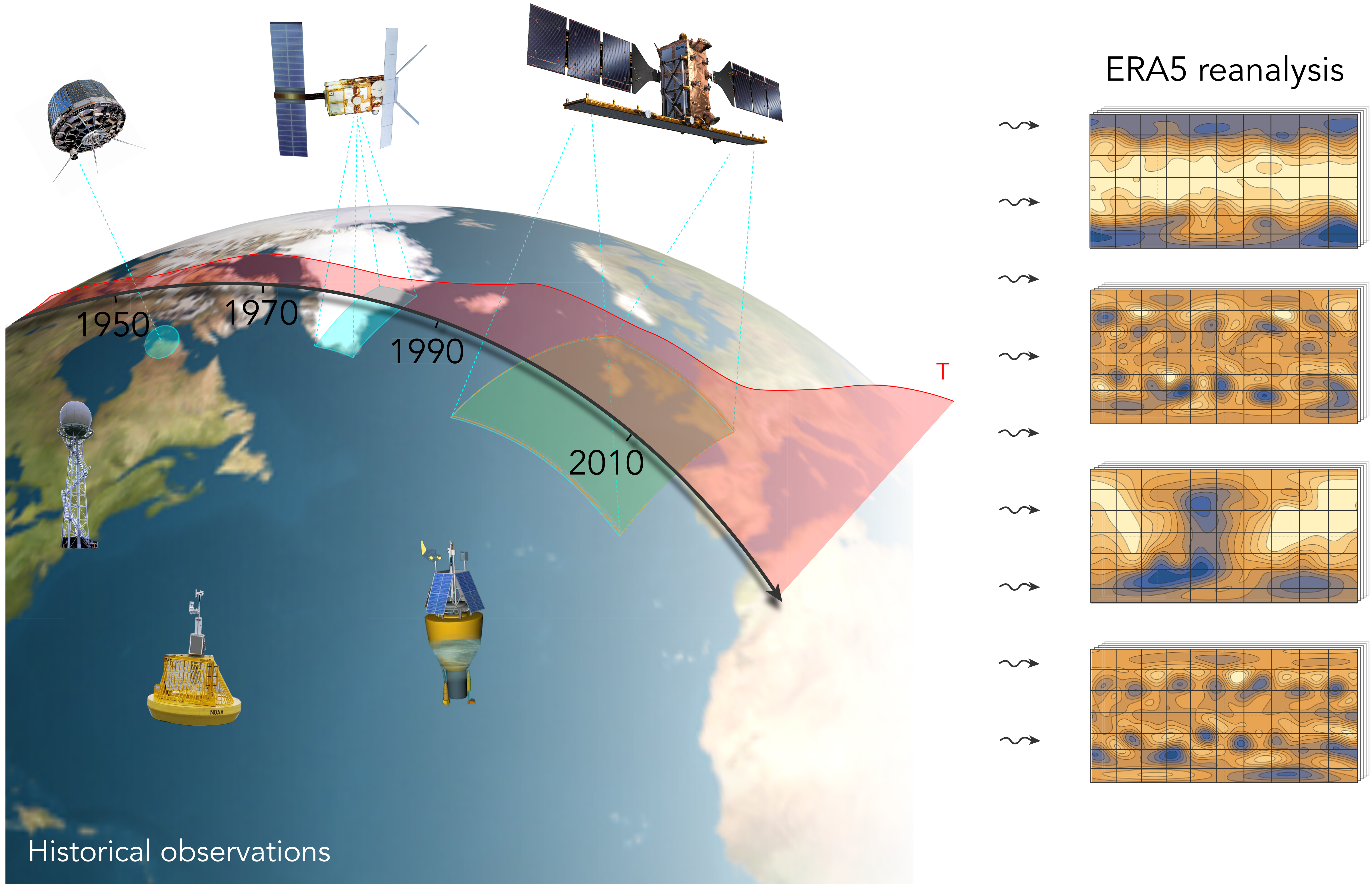
AtmoRep


1950
1970
1990
2010
T
Historical observations
ERA5 reanalysis
large scale
machine learning
applications
Forecasting
Impact analysis
Climate projections
Downscaling

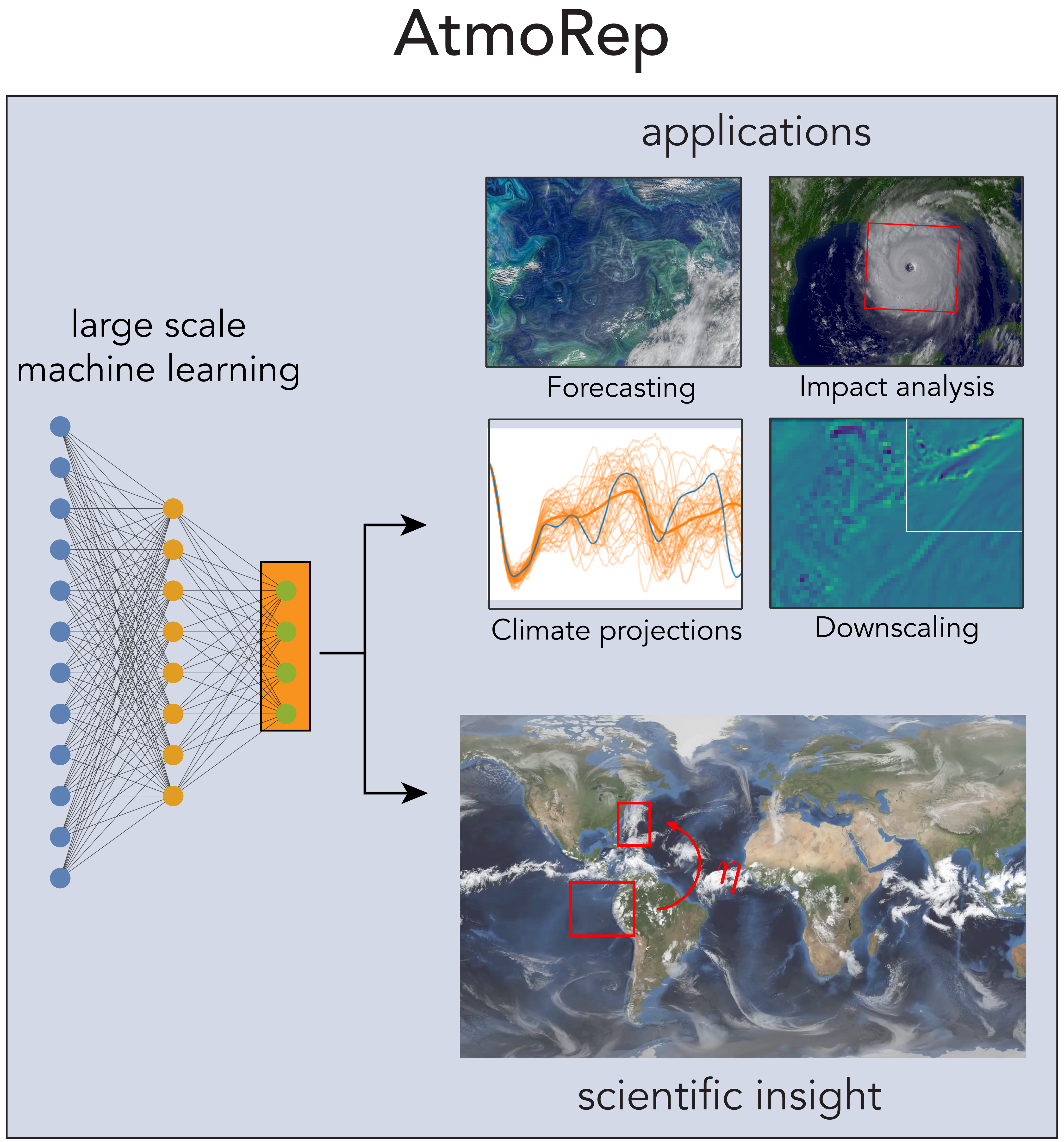
AtmoRep

1950
1970
1990
2010
T
Historical observations
ERA5 reanalysis
large scale
machine learning
applications
Forecasting
Impact analysis
Climate projections
Downscaling
scientific insight

AtmoRep



1950
1970
1990
2010
T
Historical observations
ERA5 reanalysis



AtmoRep
applications
large scale
machine learning
Forecasting
Impact analysis
Climate projections
Downscaling
scientific insight

AtmoRep: longer term objectives

AtmoRep: longer term objectives

- Weather forecasting
  - Representation network as holistic closure for unresolved effects/basis for fully network based model
  - Address in particular challenging objectives such as pre-cipitation, extreme events, sub-seasonal to seasonal

AtmoRep: longer term objectives

- Weather forecasting
- Climate projections
  - Representation network as holistic closure/parametrization for unresolved effects in a hybrid model
  - Ergodic hypothesis: use large spatial diversity (with local model) as substitute for lack of historical data
  - Stochastic climate models:[1] representation network learns fast, stochastic scale of overall dynamics

[1] K. Hasselmann. Stochastic climate models part I. Theory. Tellus, 28(6):473–485, 1976.

AtmoRep: longer term objectives

- Weather forecasting
- Climate projections
- Coupled Earth system
  - Atmosphere, ocean, infrastructure, …

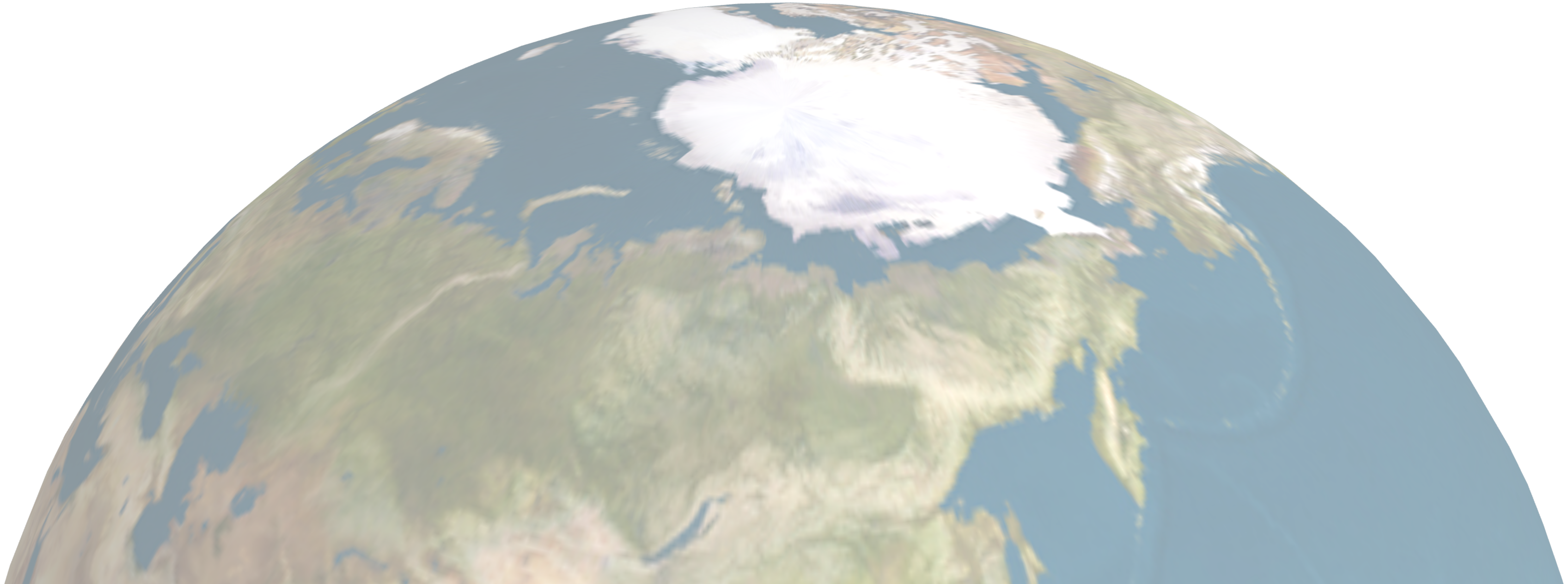
AtmoRep: longer term objectives

- Weather forecasting
- Climate projections
- Coupled Earth system
- Scientific model
  - Can we use a large representation network as a scientifical model complementing existing ones such as partial differential equations?

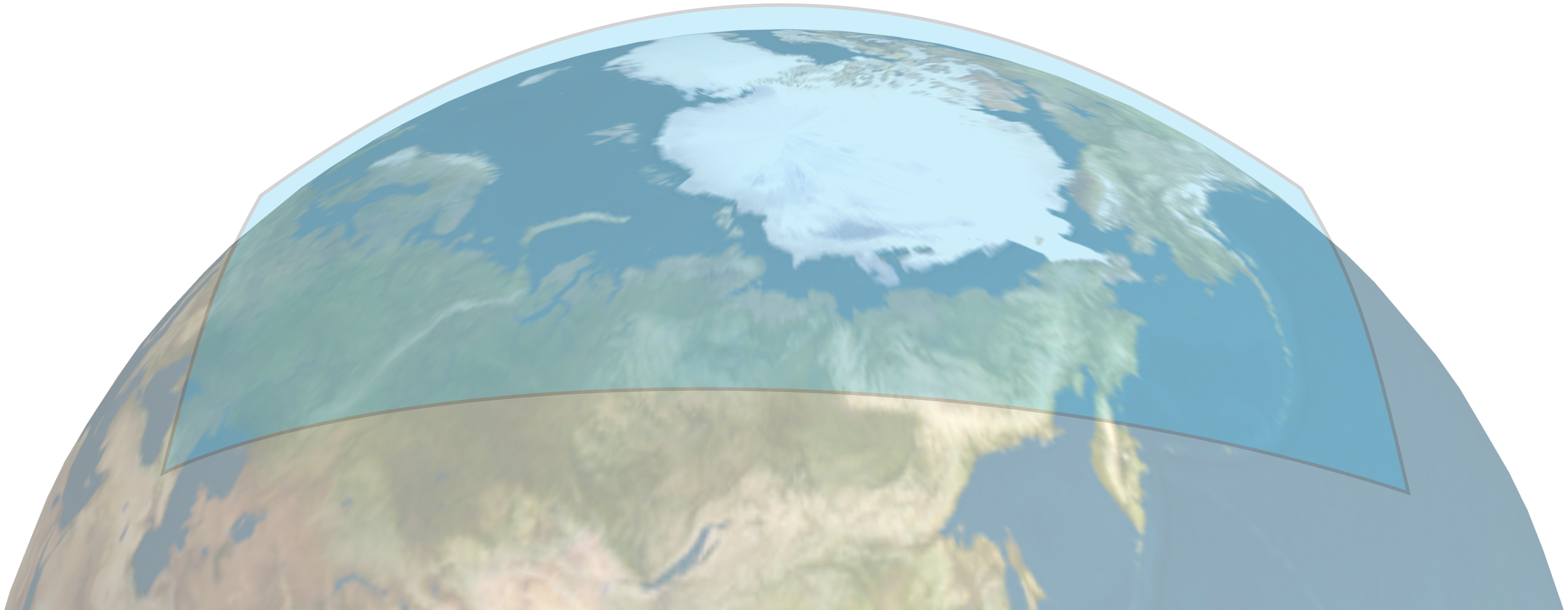
AtmoRep: longer term objectives

- Weather forecasting
- Climate projections
- Coupled Earth system
- Scientific model
- Training/fine-tuning on direct observational data
  - Learn directly from satellite observations and other mea-surments

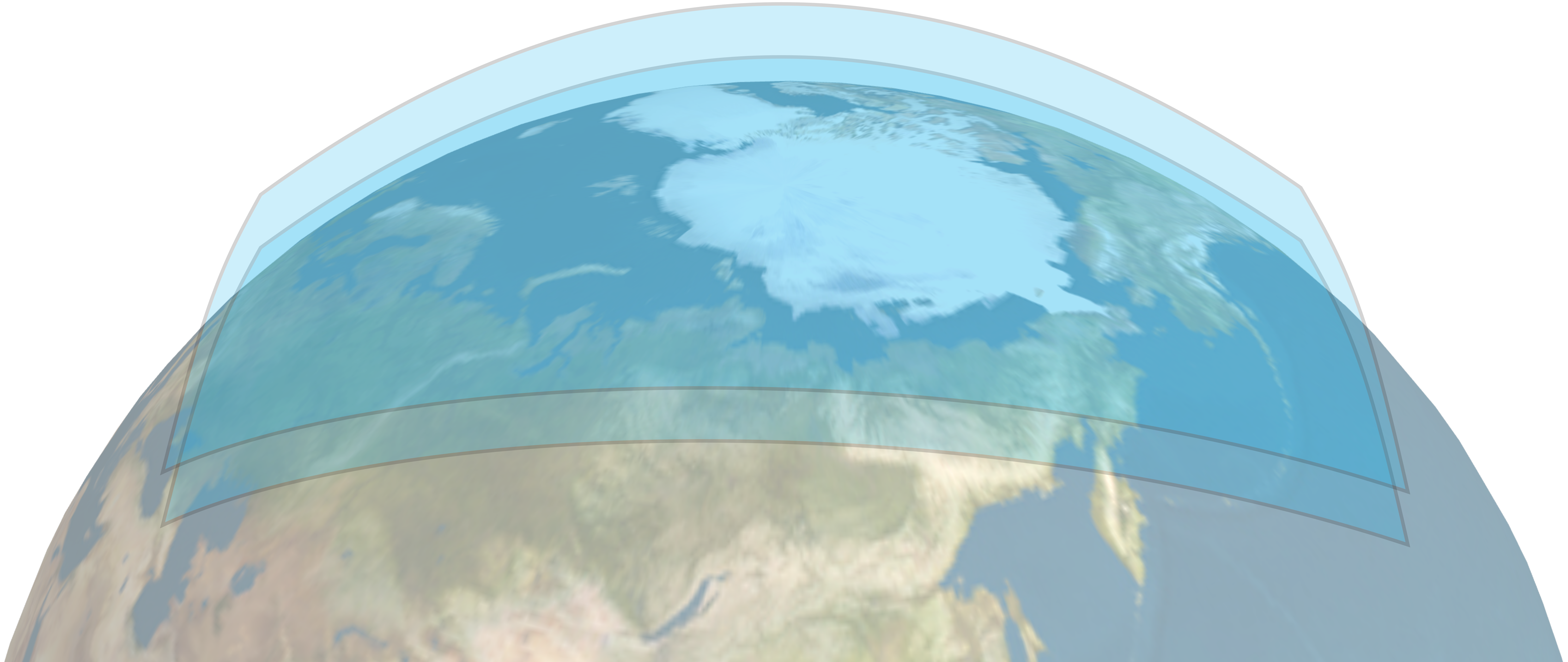
Data: ERA5 reanalysis



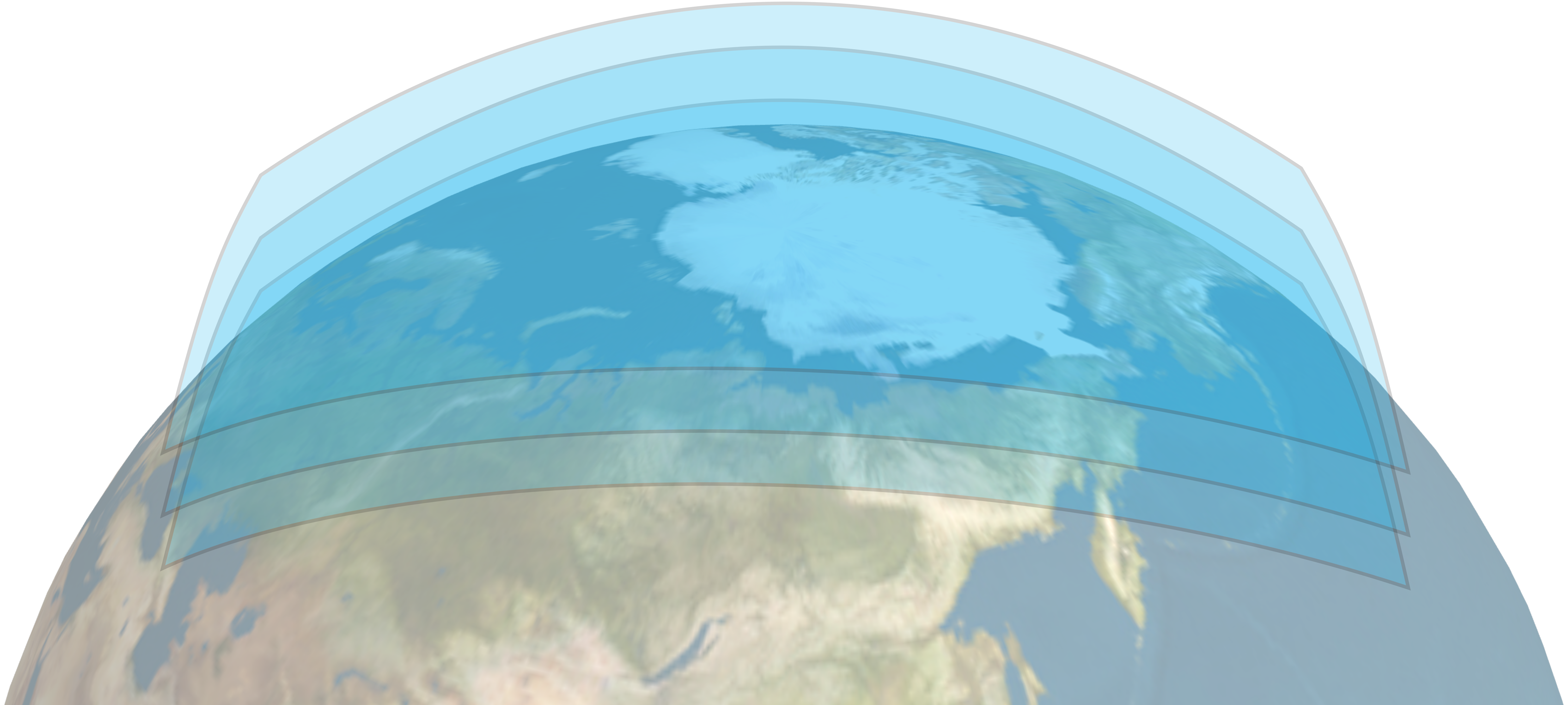
Data: ERA5 reanalysis



Data: ERA5 reanalysis

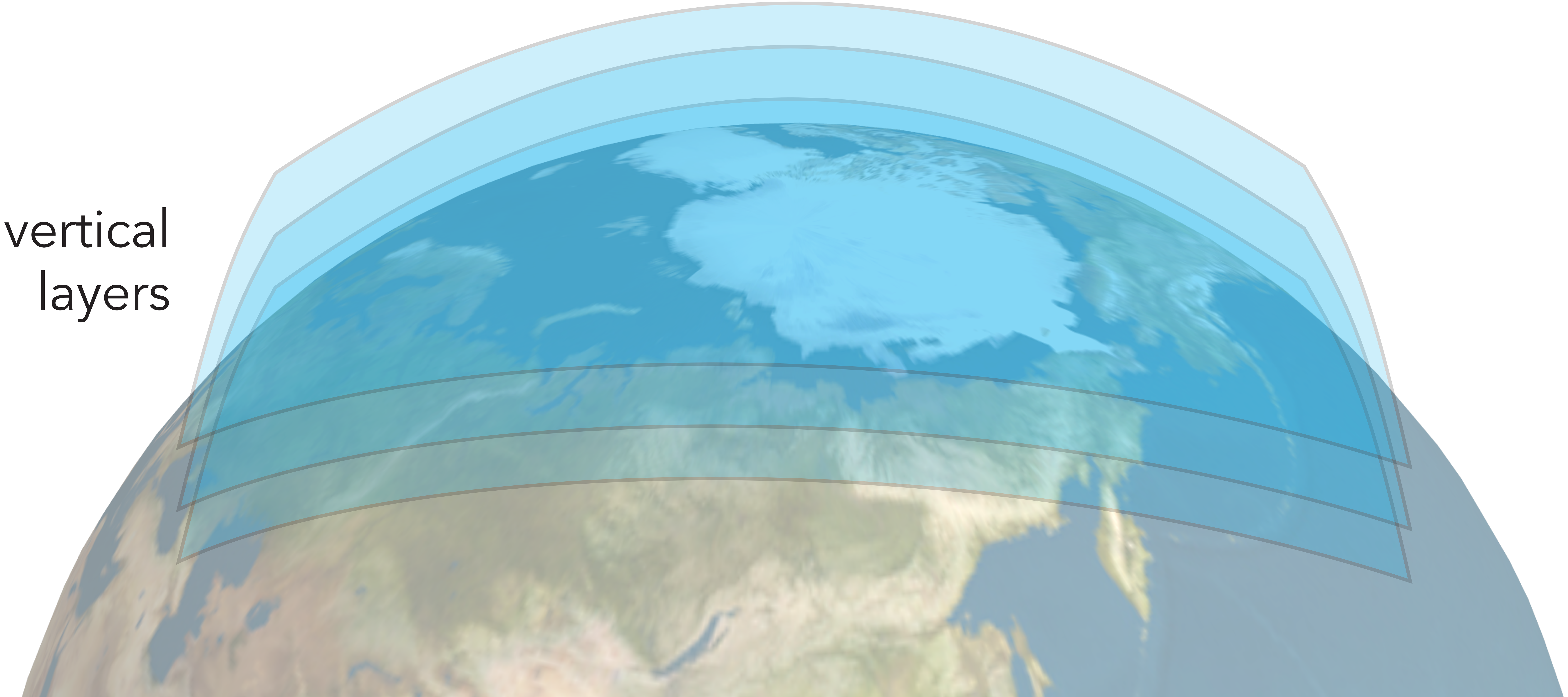


Data: ERA5 reanalysis



Data: ERA5 reanalysis

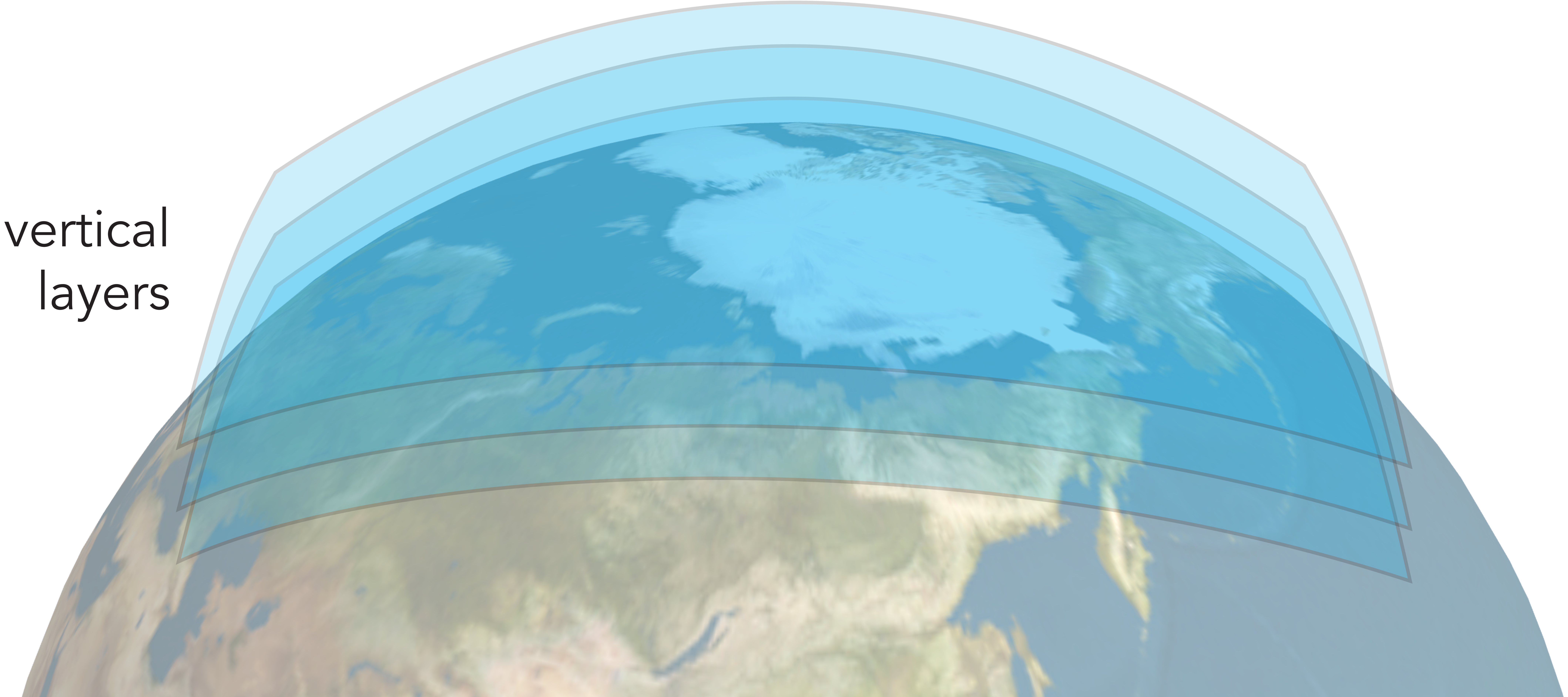
137 vertical
layers



Data: ERA5 reanalysis

721x1440 horizontal grid (0.25 degree)

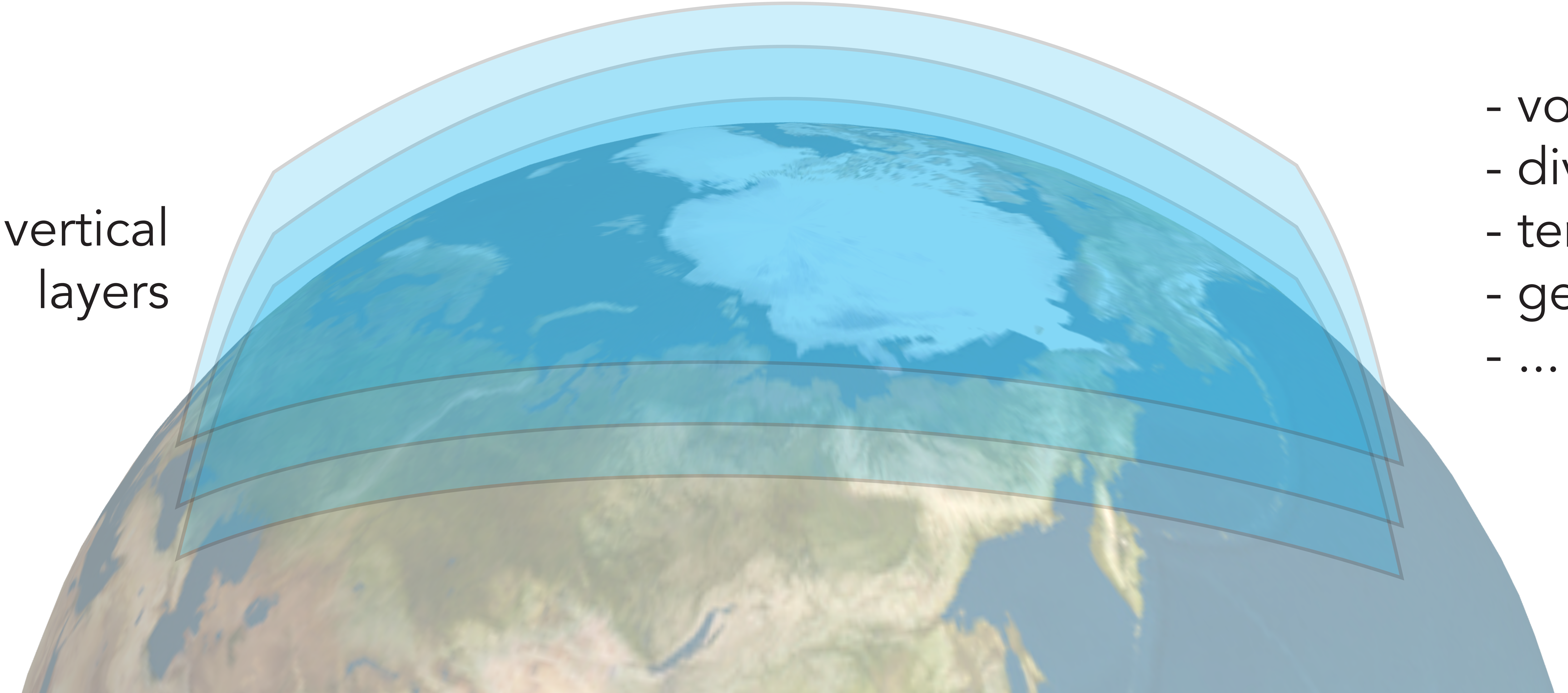
137 vertical
layers



Data: ERA5 reanalysis

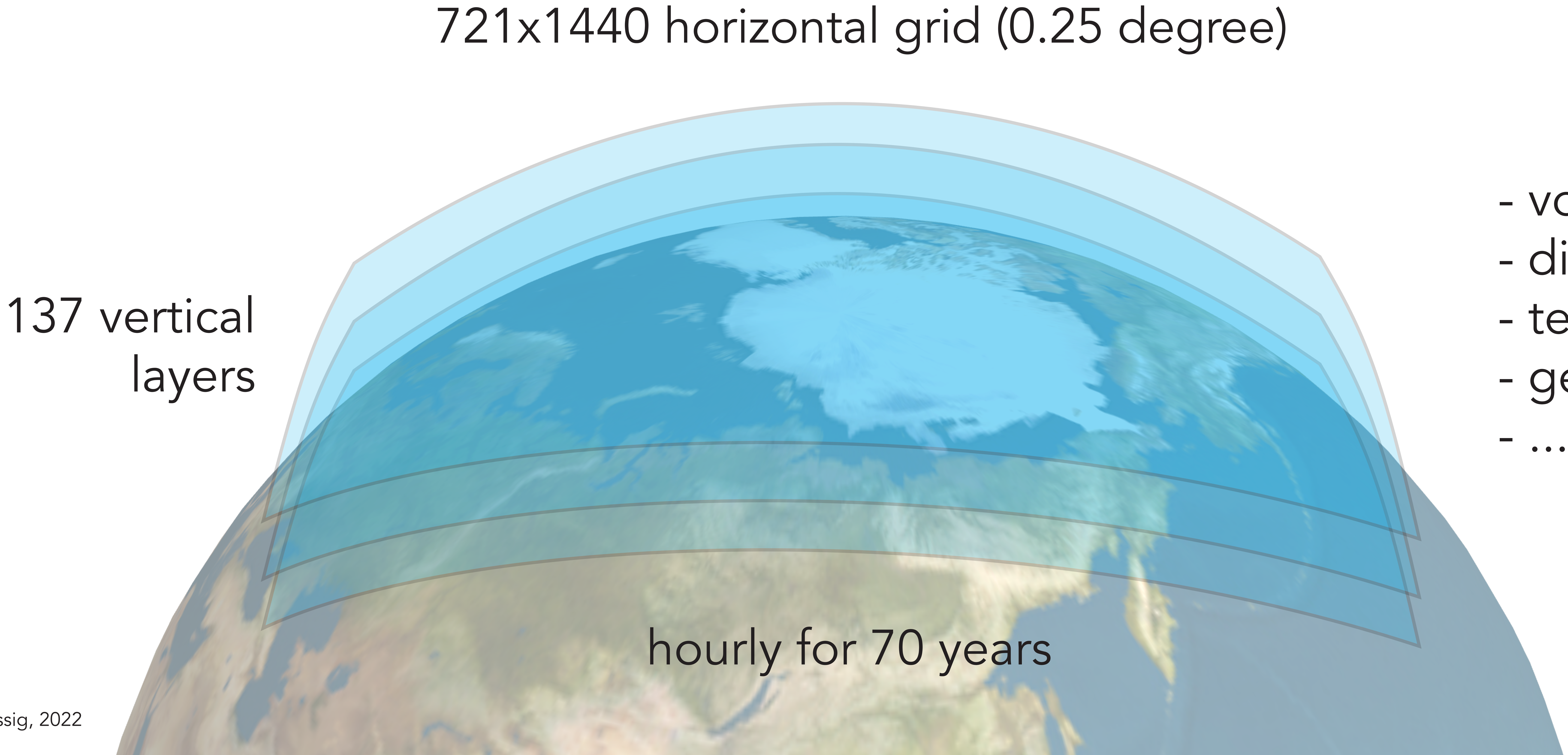
721x1440 horizontal grid (0.25 degree)

137 vertical layers



- vorticity
- divergence
- temperature
- geopotential
- ...

Data: ERA5 reanalysis

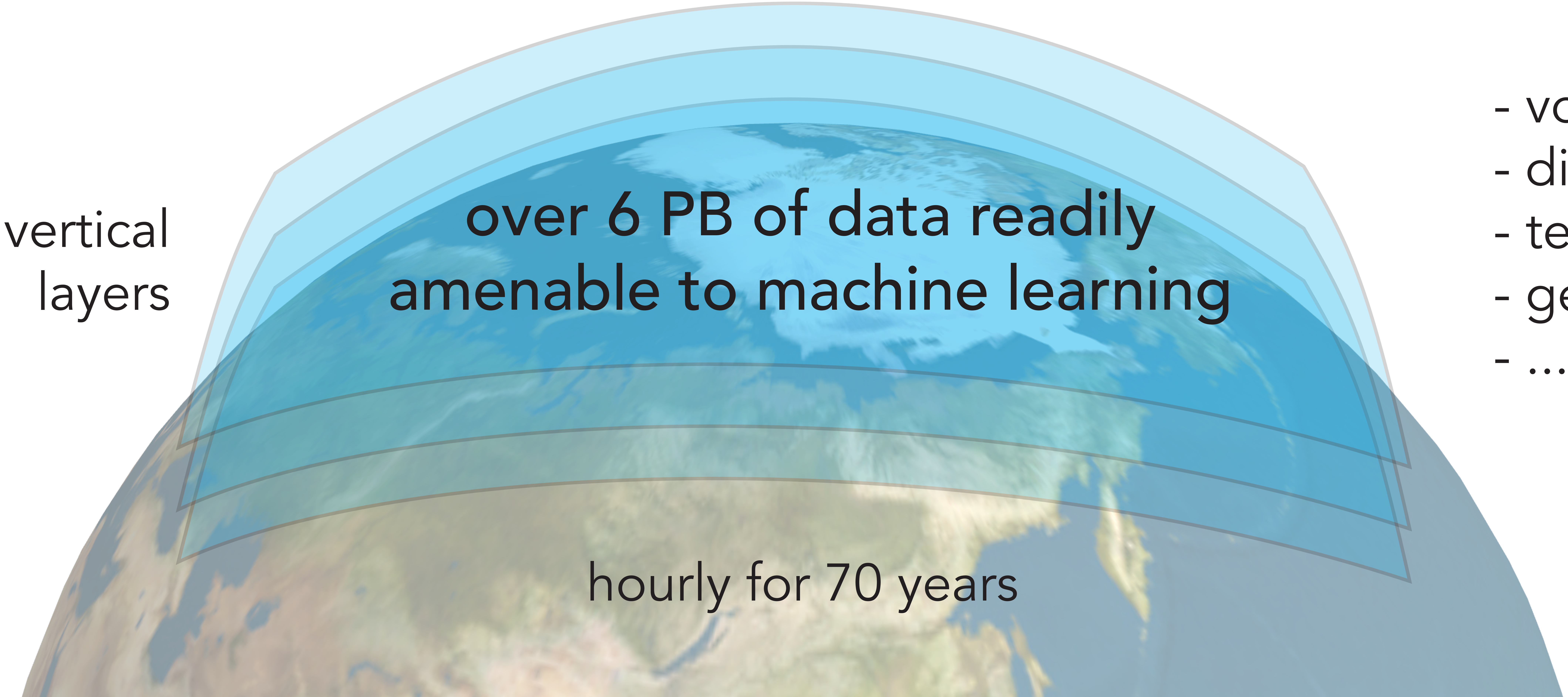


- vorticity
- divergence
- temperature
- geopotential
- ...

Data: ERA5 reanalysis

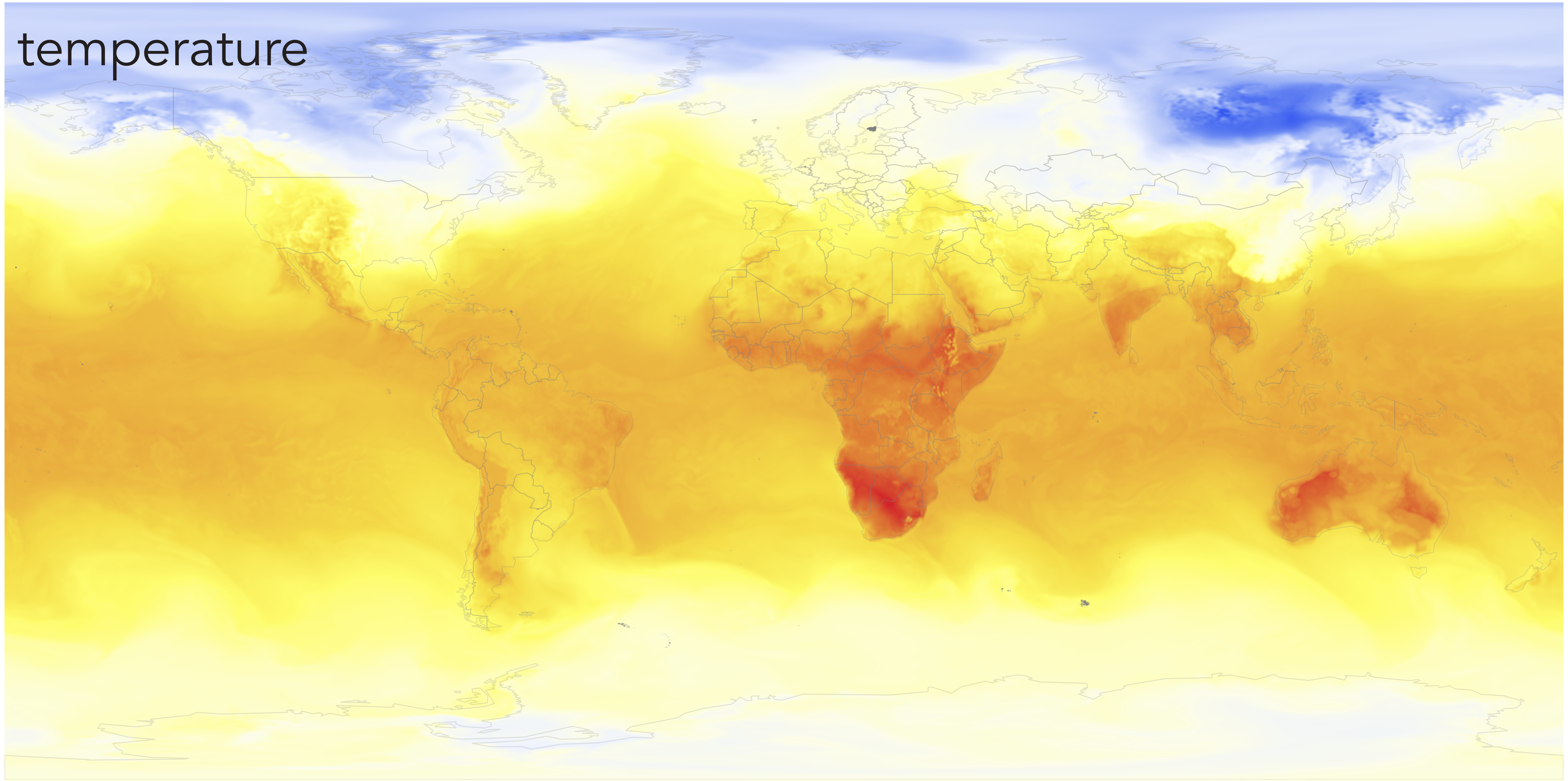
721x1440 horizontal grid (0.25 degree)

137 vertical layers

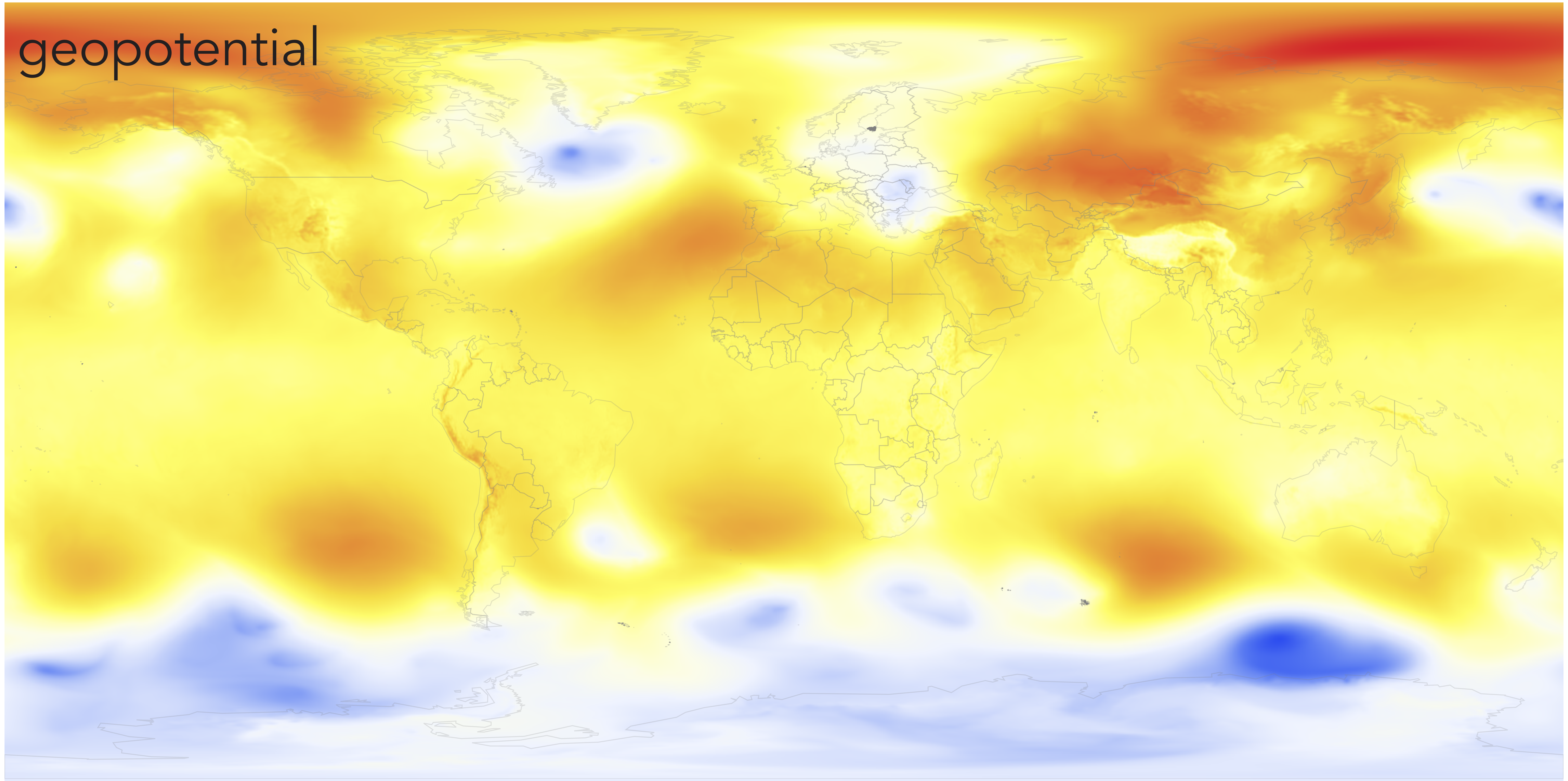


- vorticity
- divergence
- temperature
- geopotential
- ...

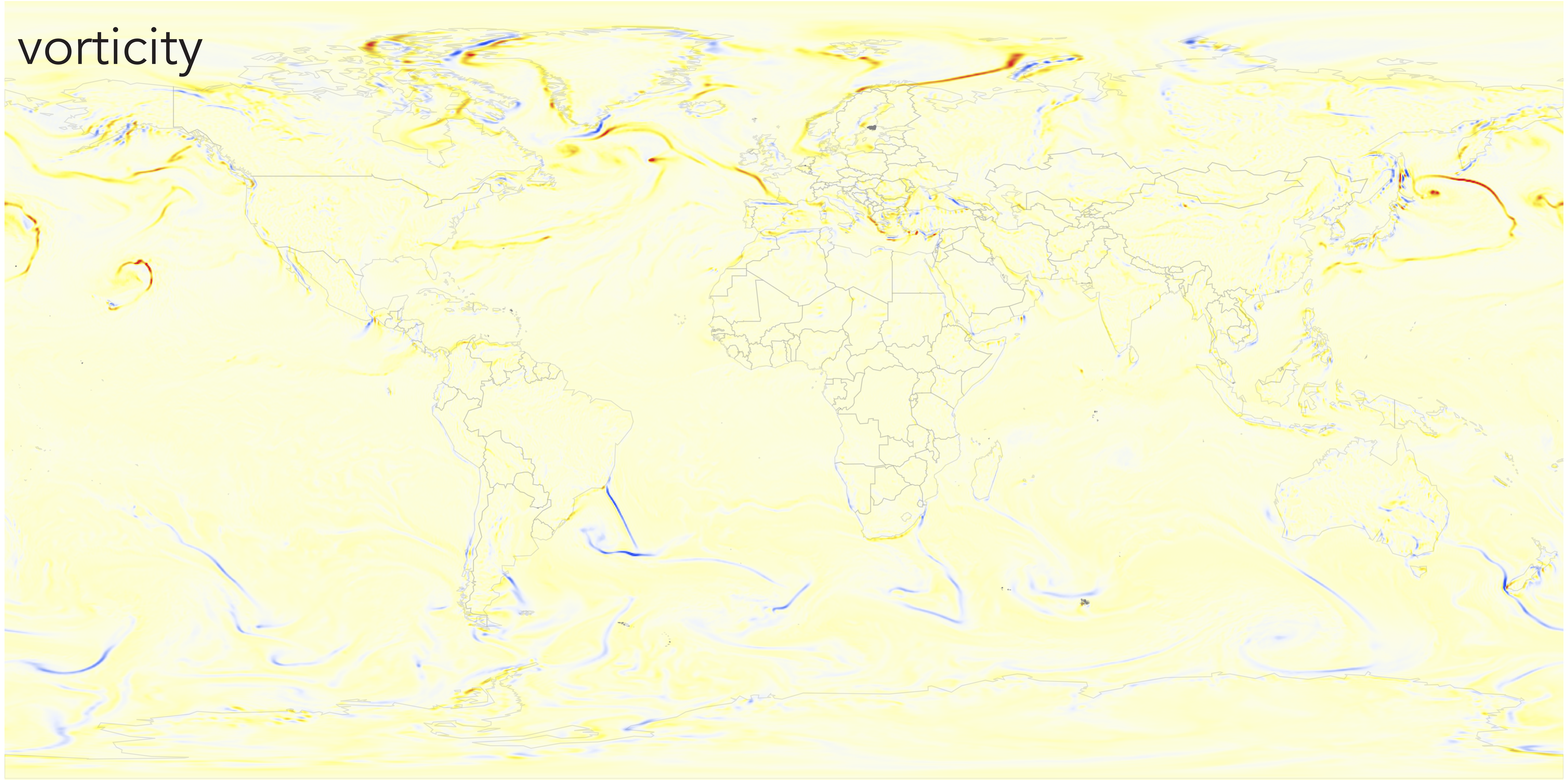
AtmoRep data



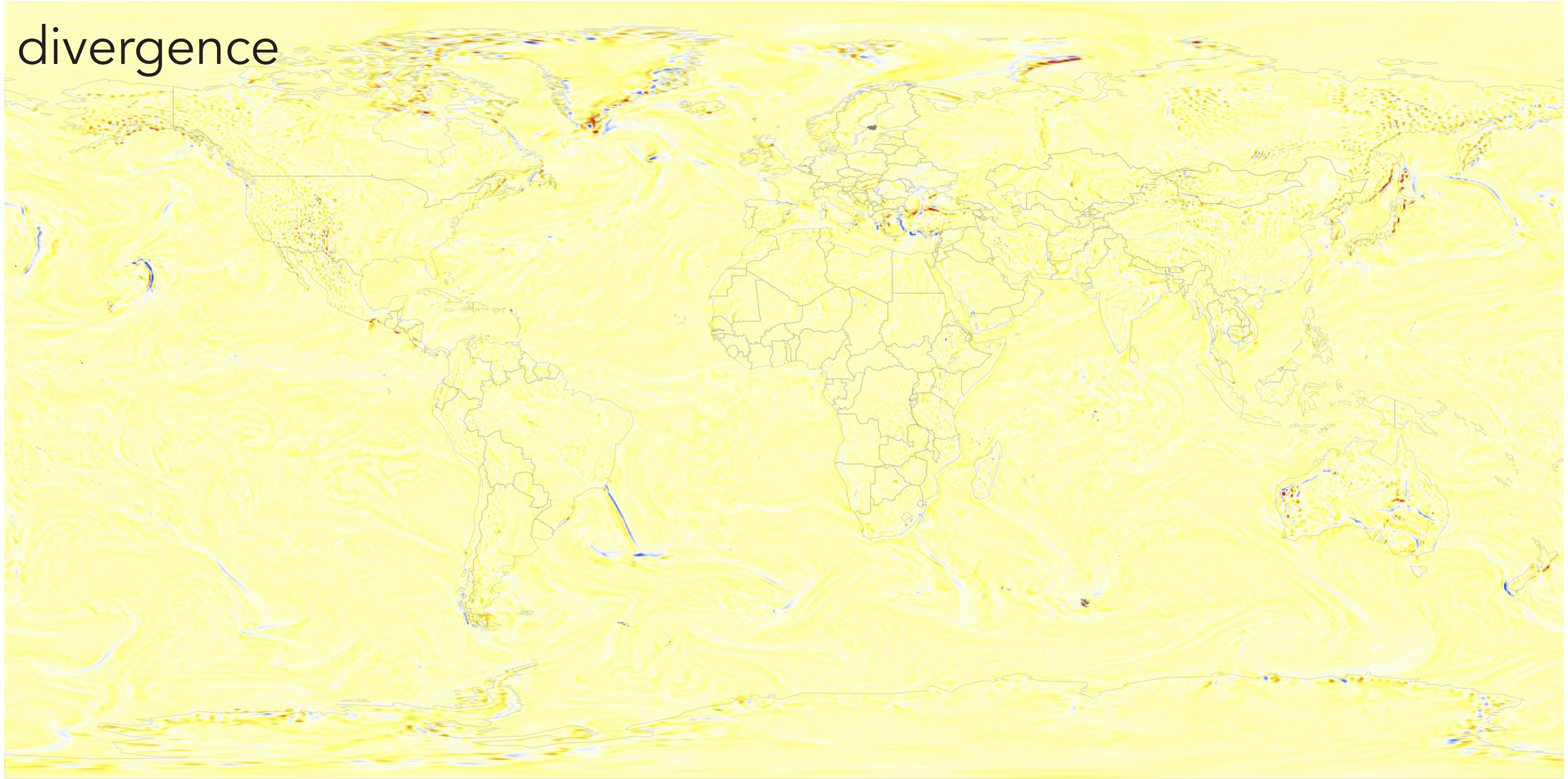
AtmoRep data



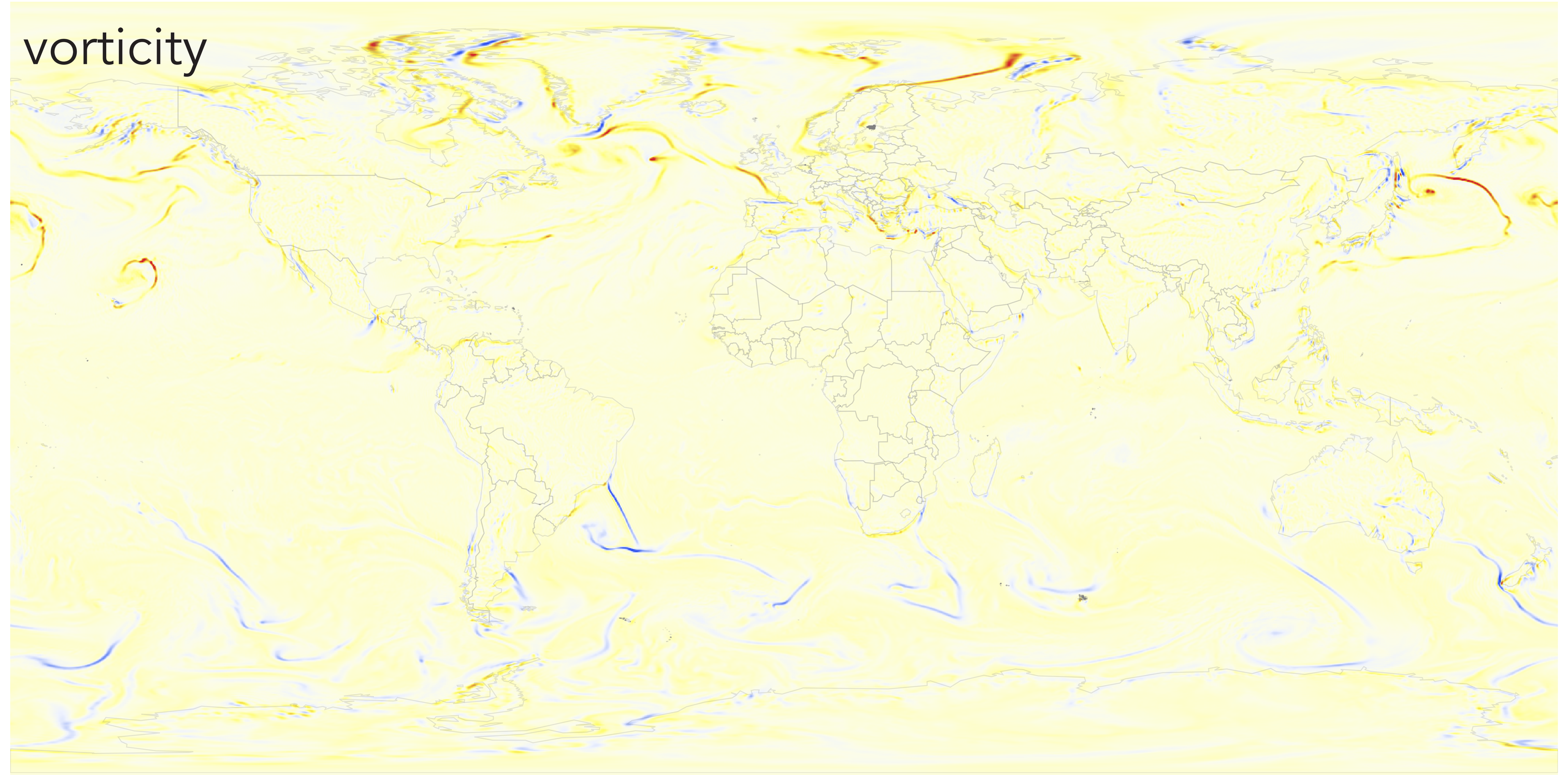
AtmoRep data

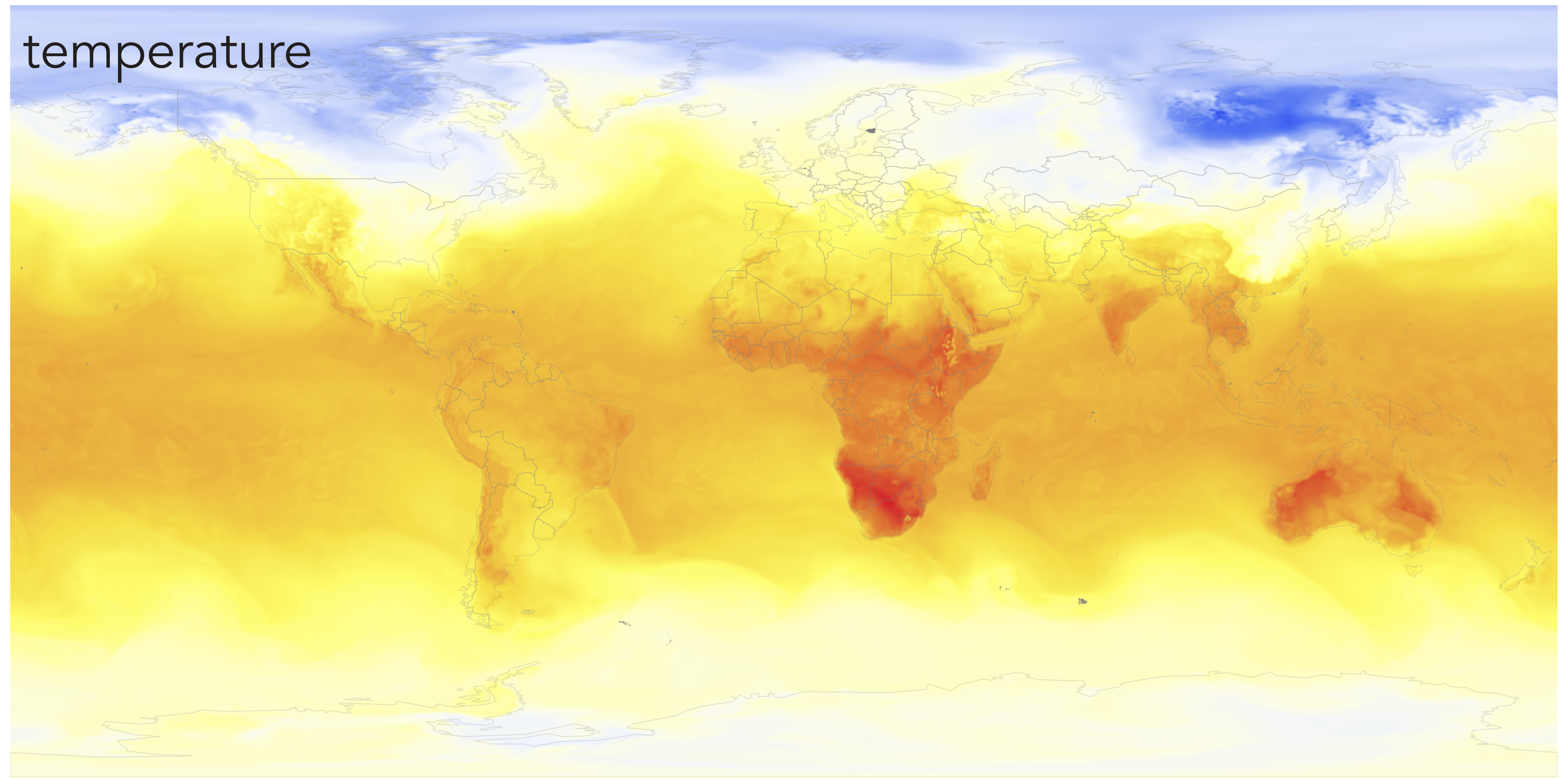


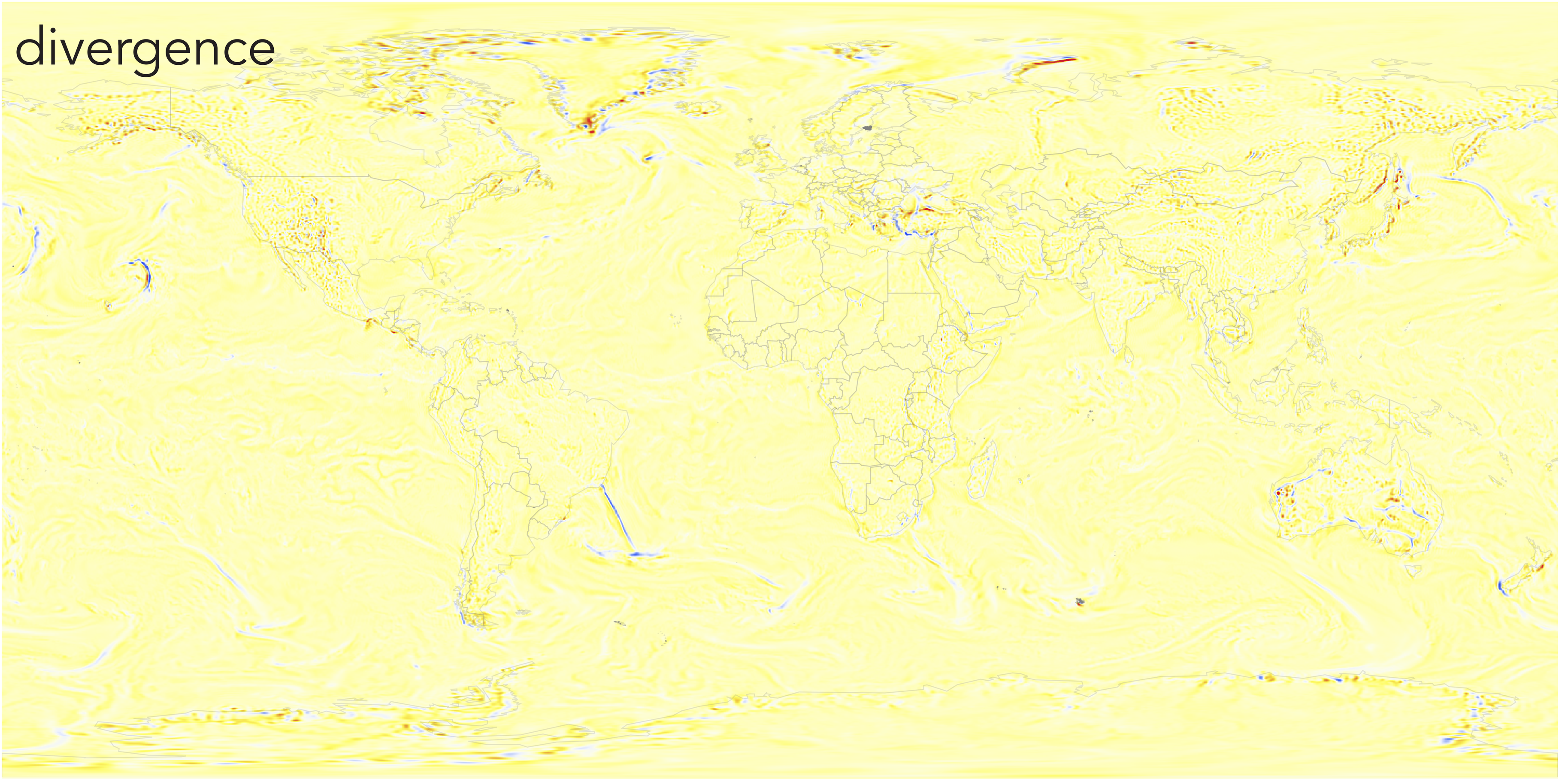
AtmoRep data

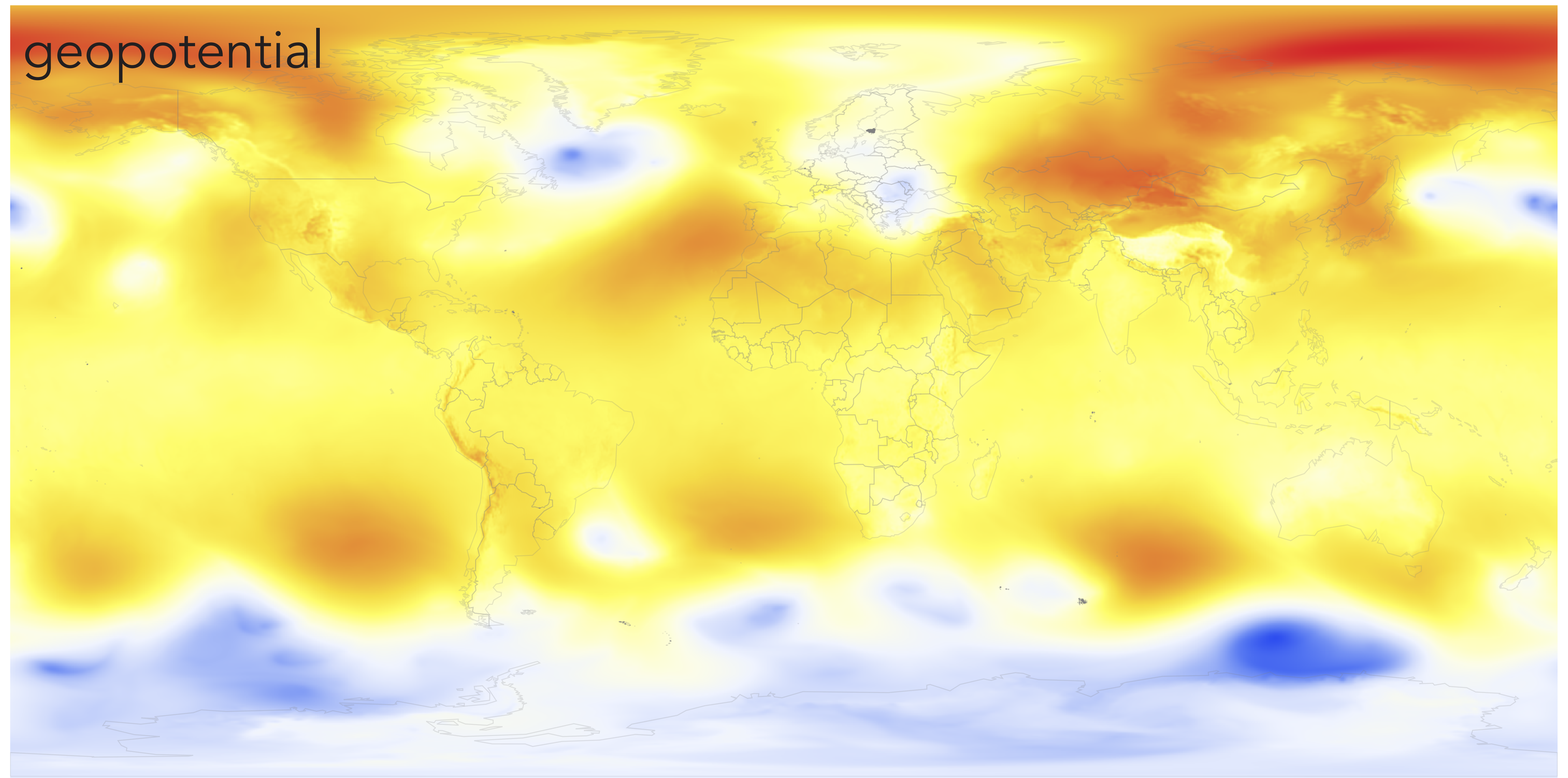


AtmoRep data

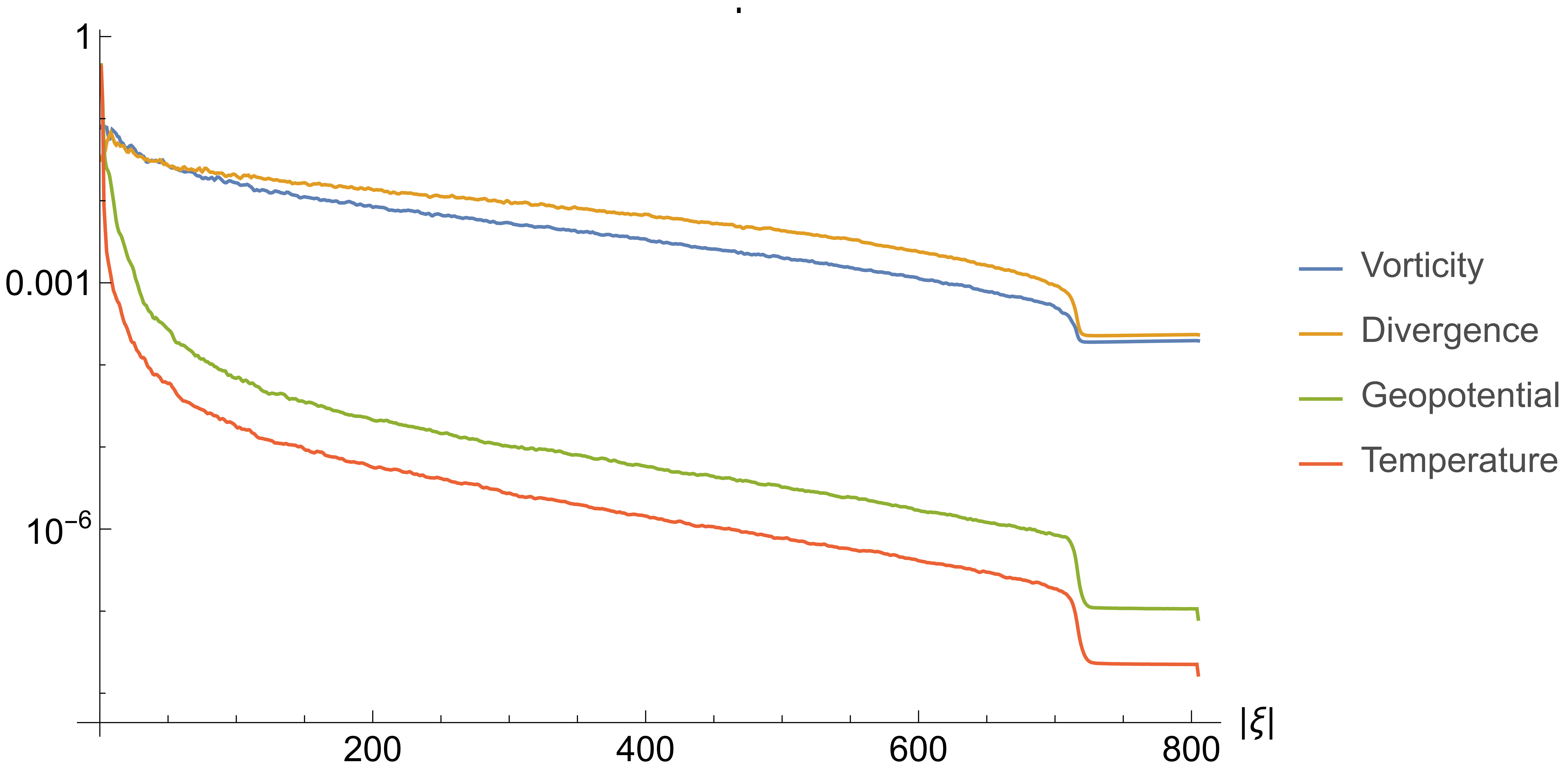




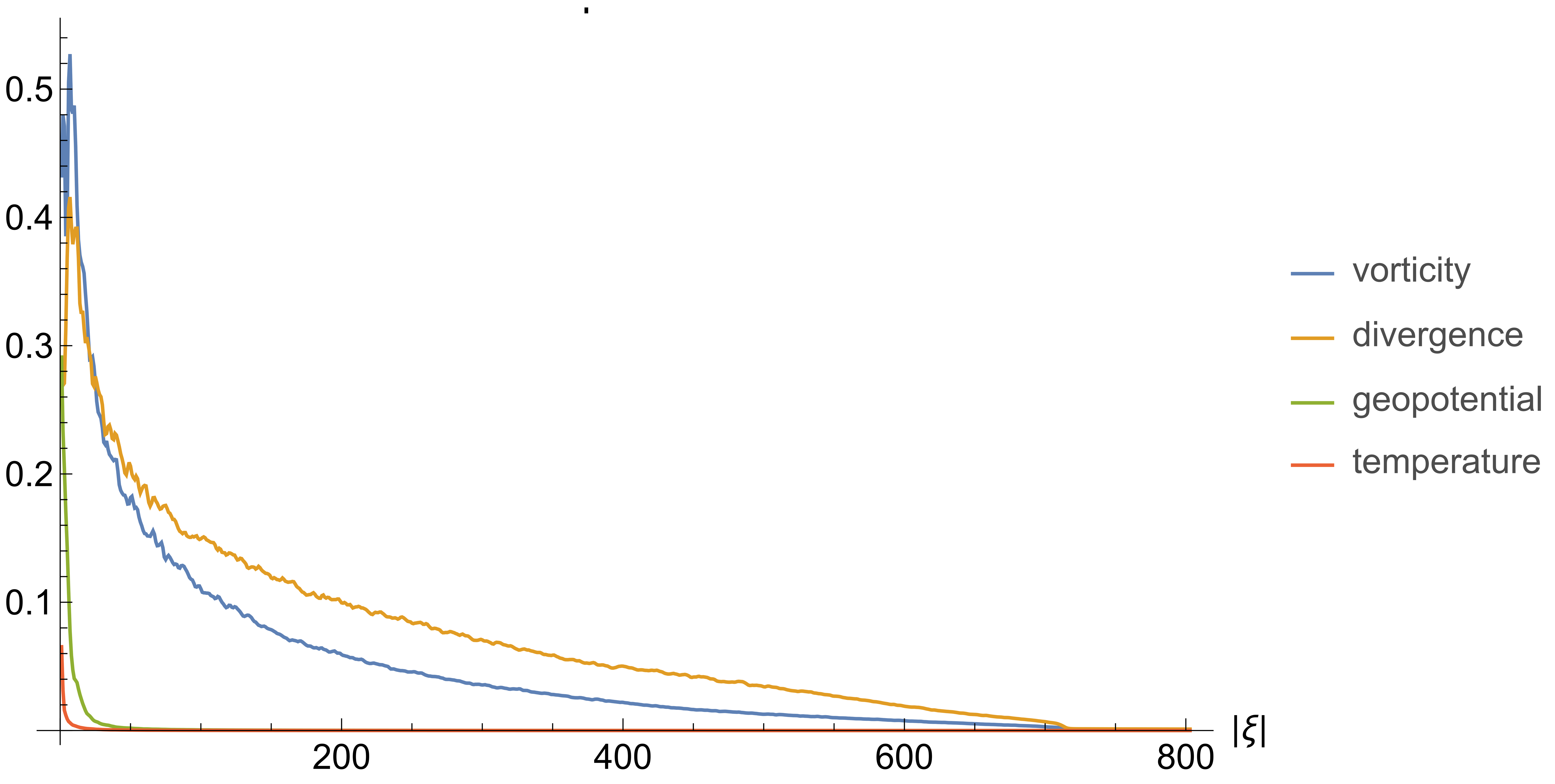




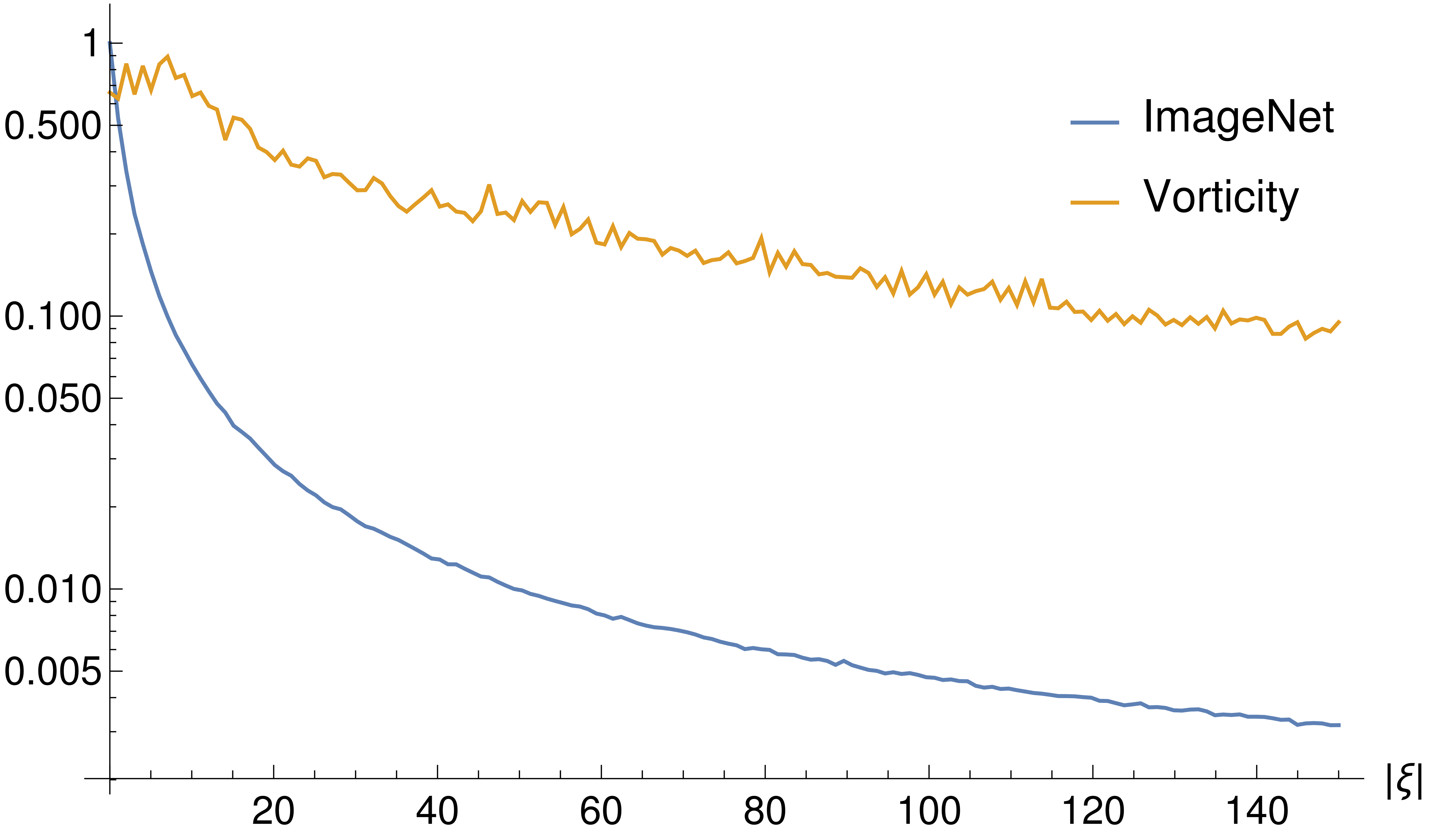
AtmoRep data



AtmoRep data



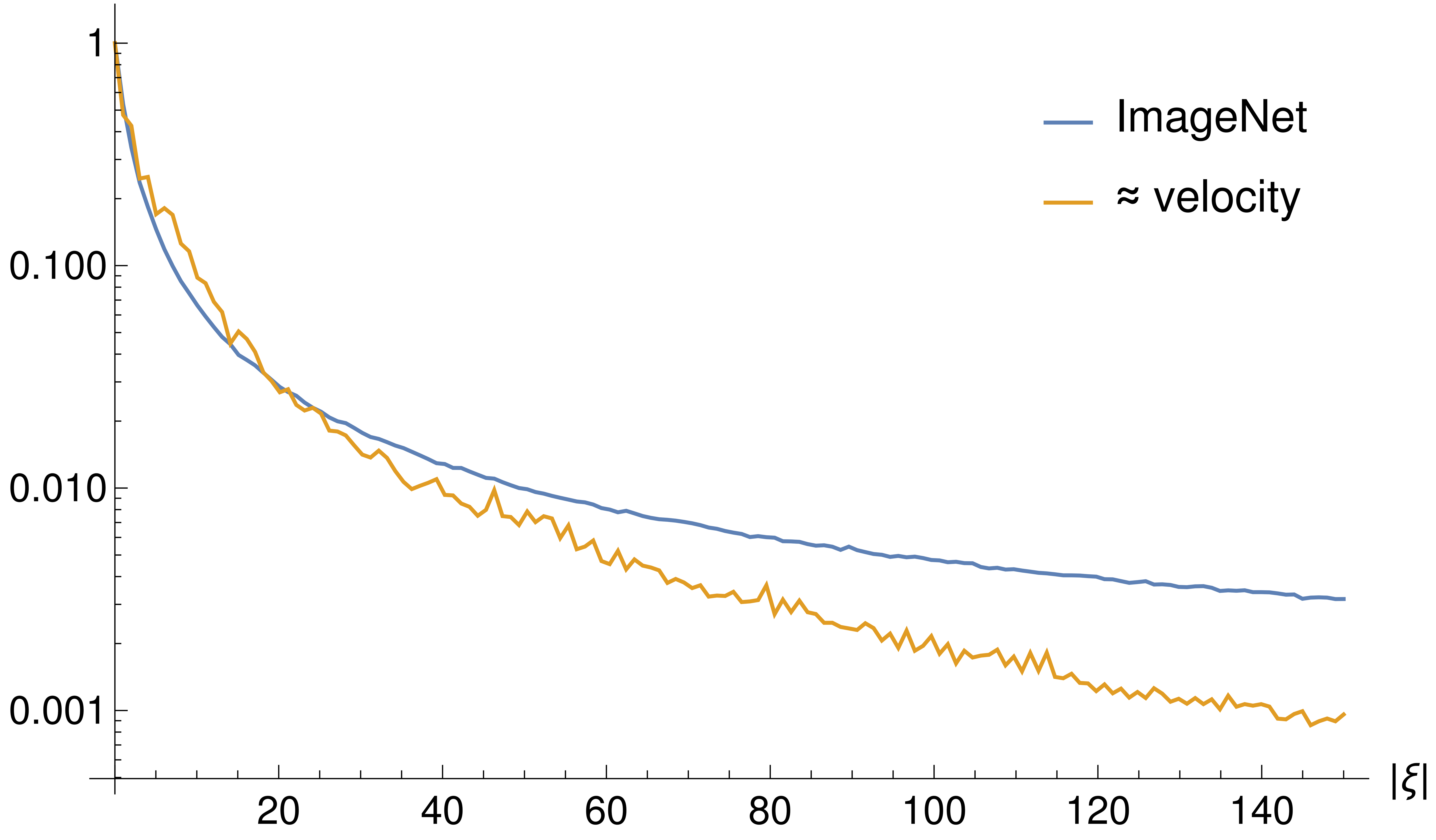
ERA5 versus ImageNet



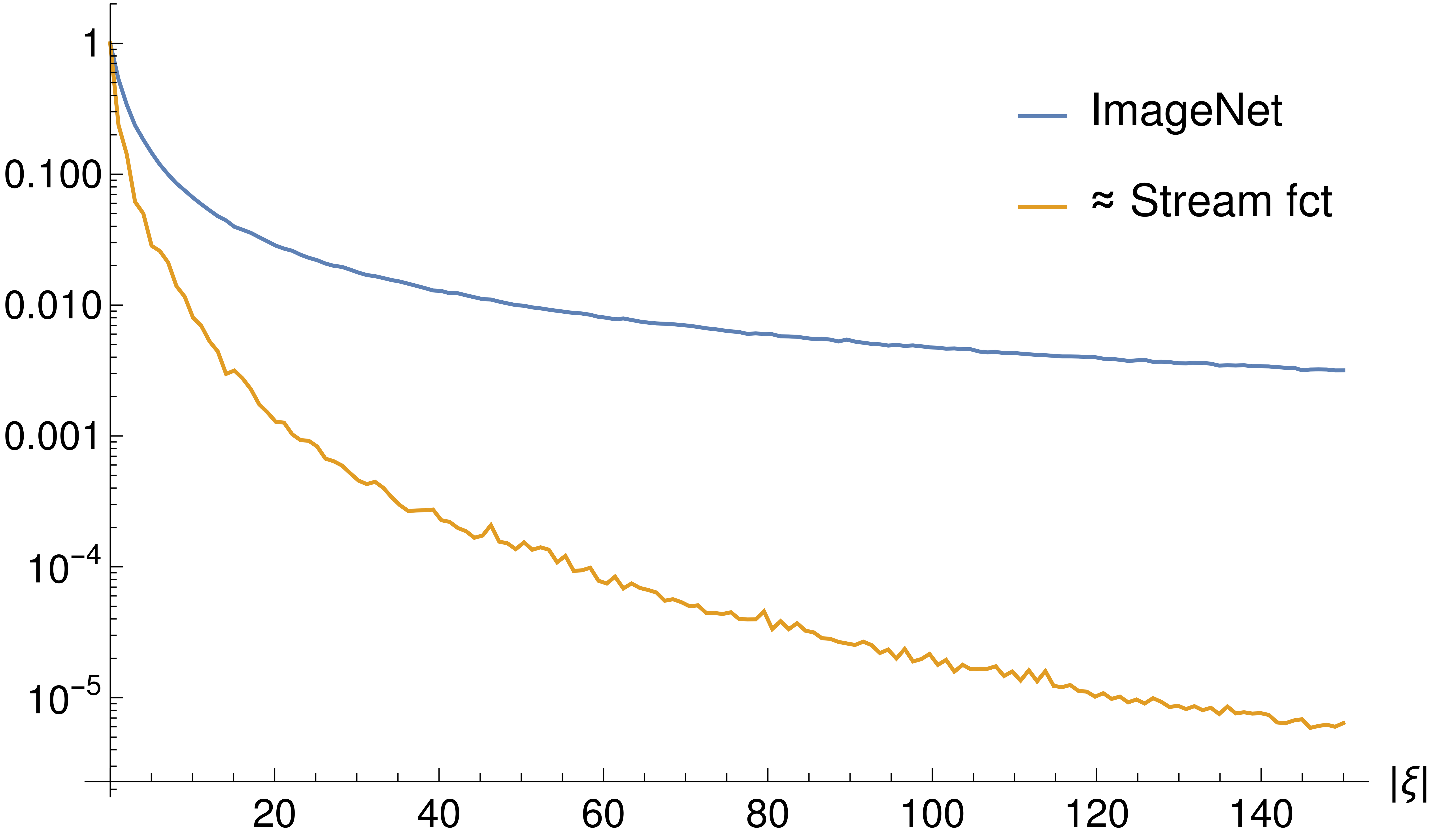
ERA5 versus ImageNet

stream function velocity potential	$\Leftrightarrow$	velocity vector field	$\Leftrightarrow$	vorticity divergence

ERA5 versus ImageNet



ERA5 versus ImageNet



ERA5 versus ImageNet

stream function
velocity potential $\Leftrightarrow$ velocity
vector field $\Leftrightarrow$ vorticity
divergence

Work with velocity with norm that emphasizes small scale detail

AtmoRep network architecture

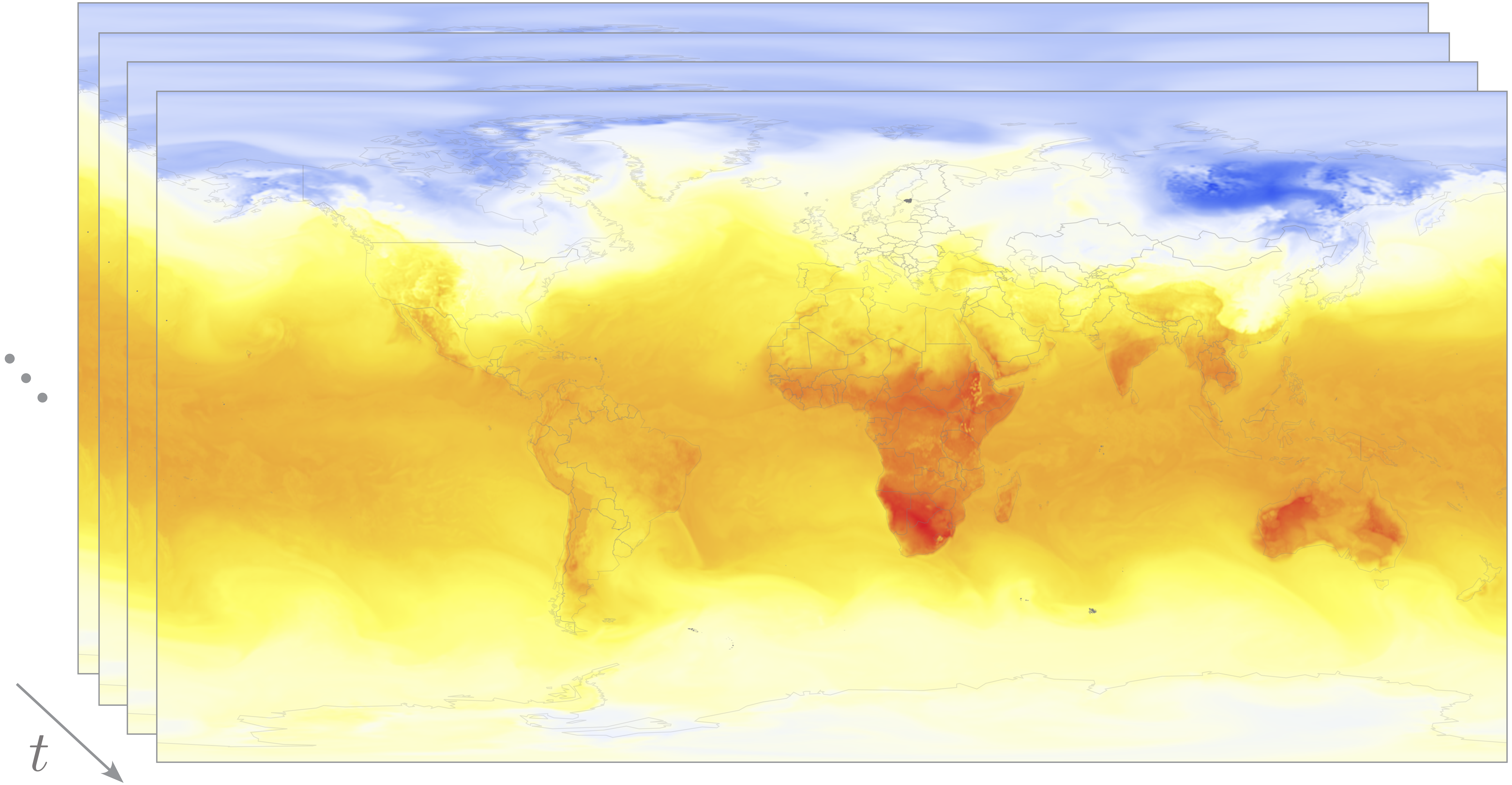
AtmoRep network architecture

- Transformer-based as network architecture
  - Scales well to very large data-sets
  - Generative model (with decoder)
  - Attention maps provide (physical) interpretability

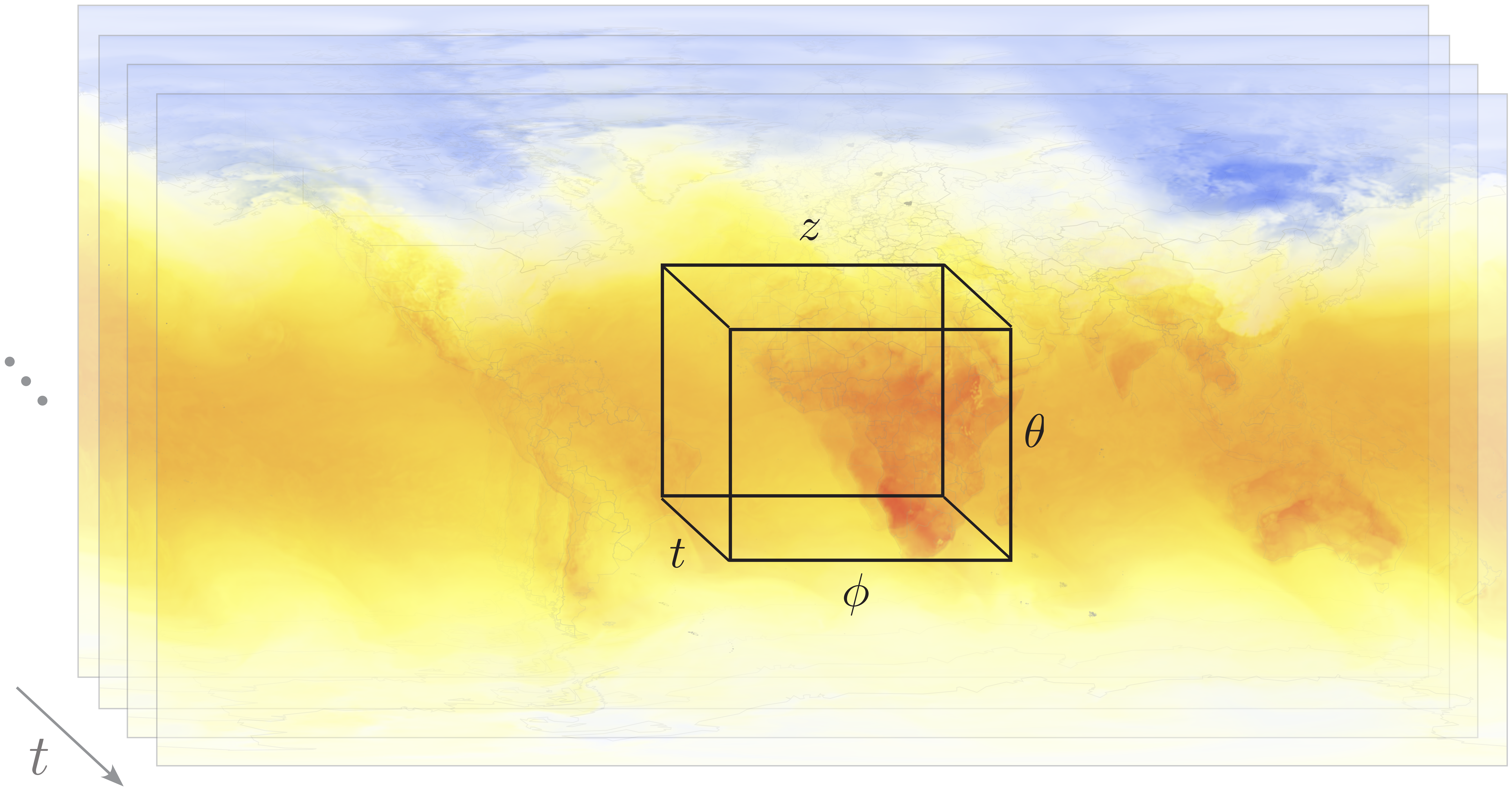
AtmoRep network architecture

- Transformer-based as network architecture
  - Scales well to very large data-sets
  - Generative model (with decoder)
  - Attention maps provide (physical) interpretability
- Network is local in space-time

AtmoRep network architecture



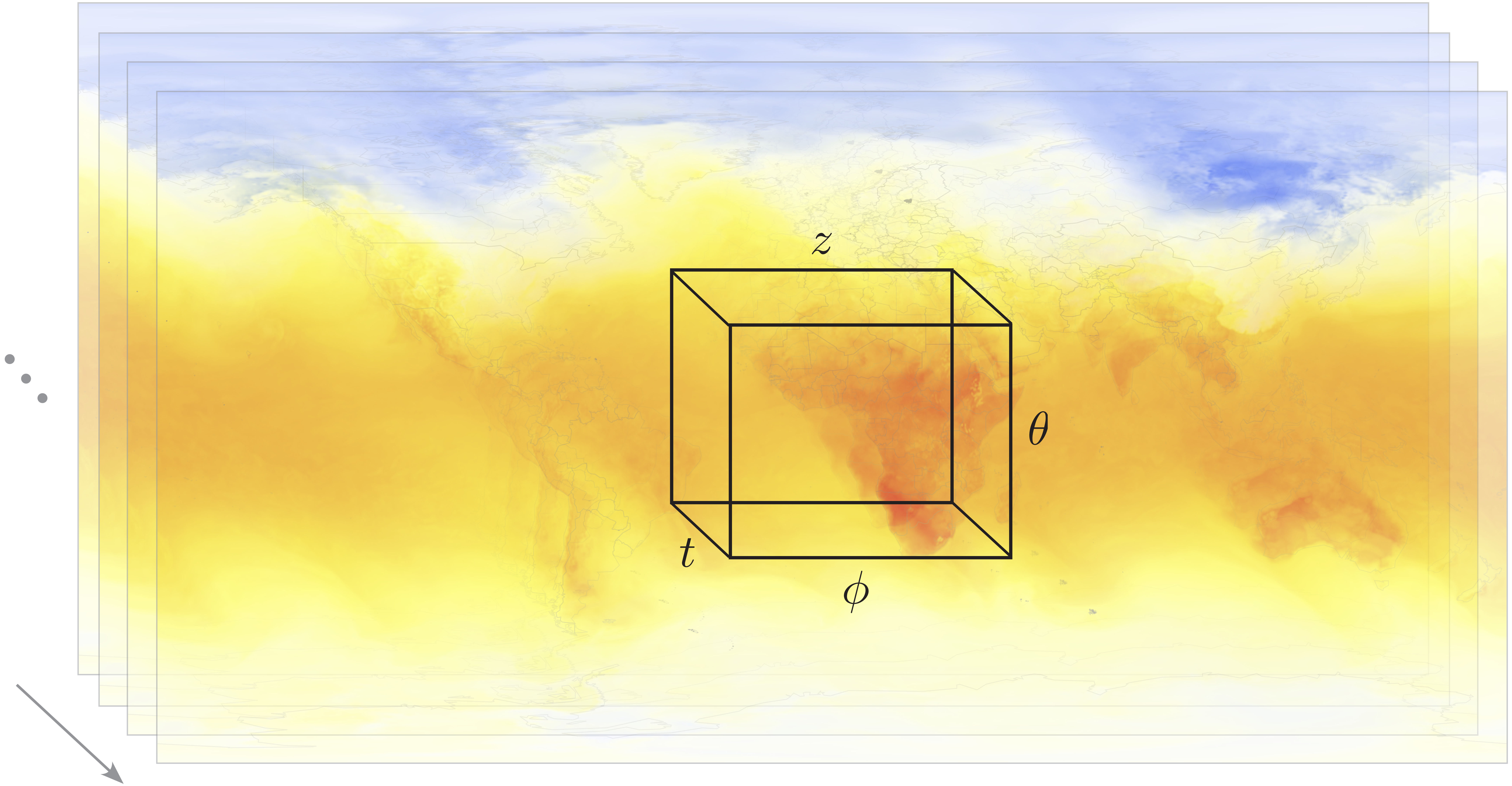
AtmoRep network architecture



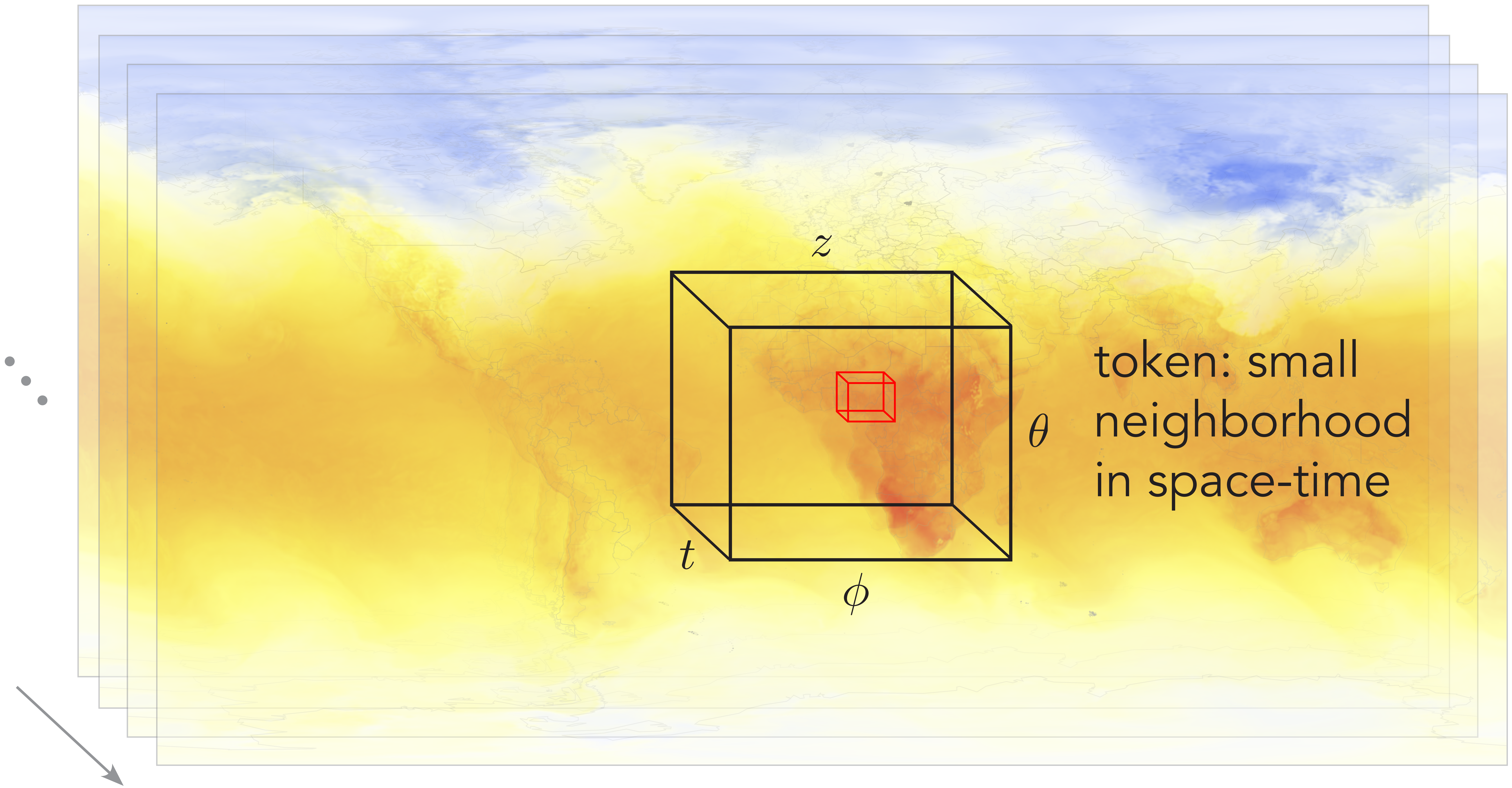
AtmoRep network architecture

- Transformer encoder-based network architecture
  - Scales well to very large data-sets
  - Generative model (with decoder)
  - Attention maps provide (physical) interpretability
- Network is local in space-time
  - Physics of dynamics are universally valid
  - Local particularities can be learned by providing time + space position as auxiliary information

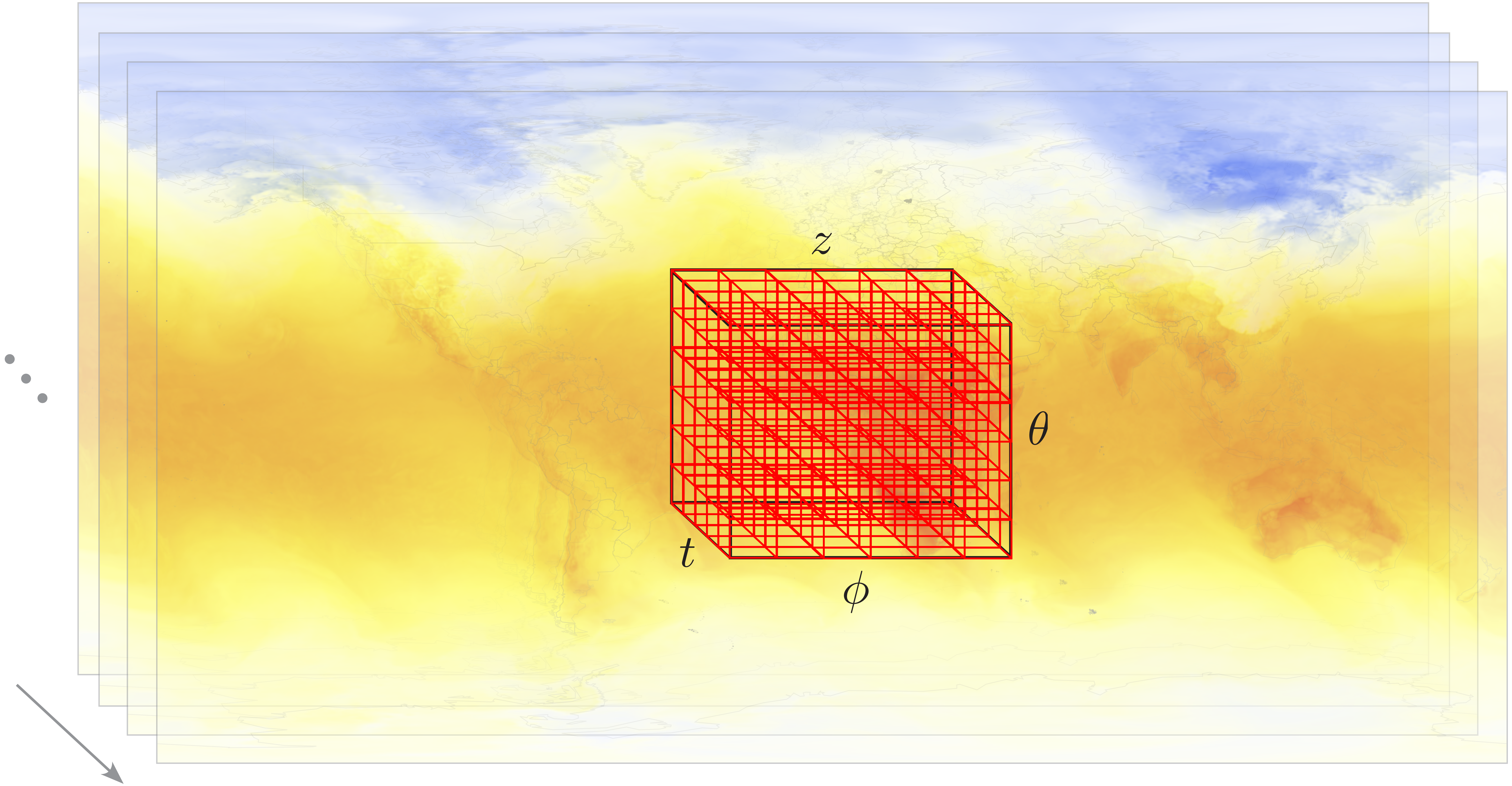
What is a token?



What is a token?

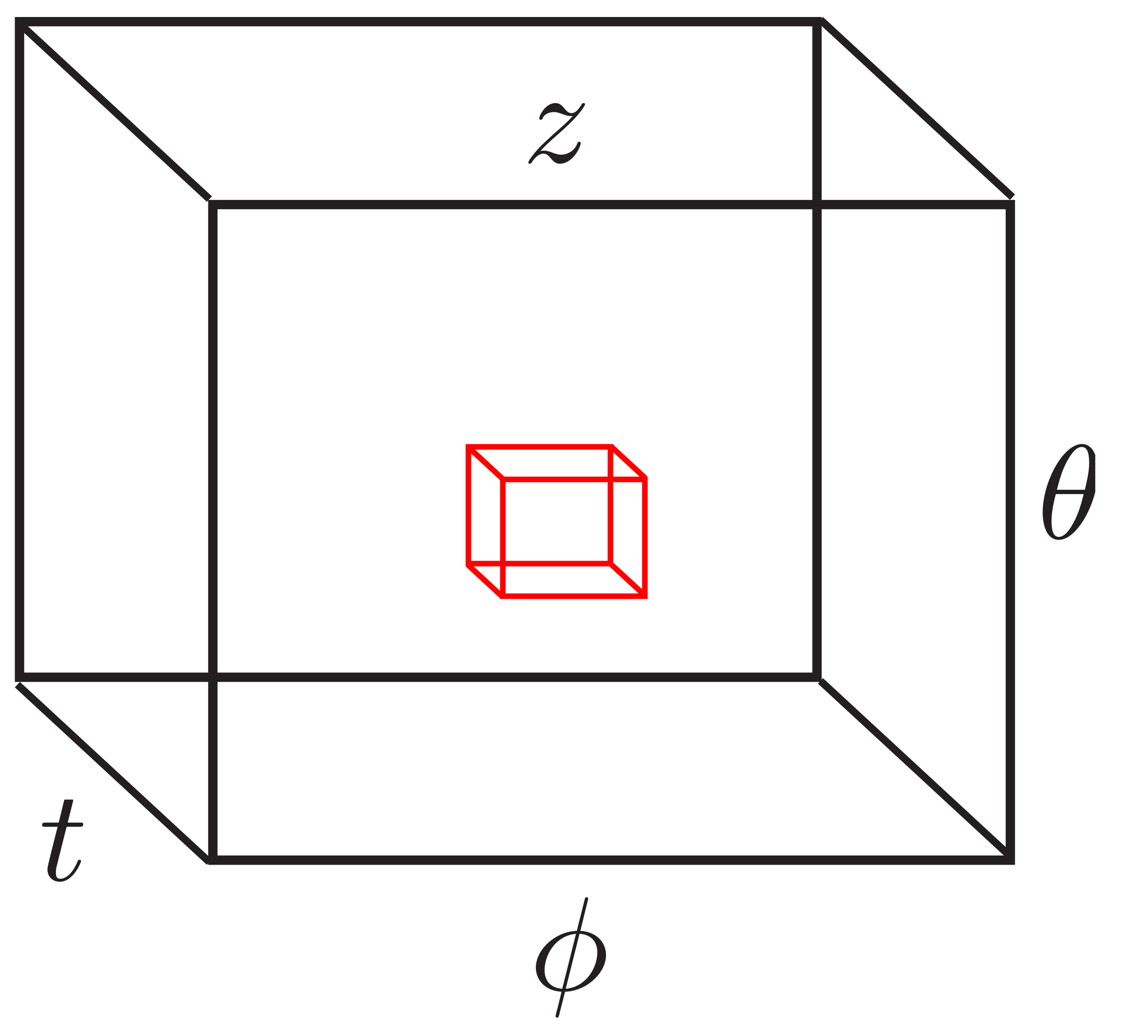


What is a token?



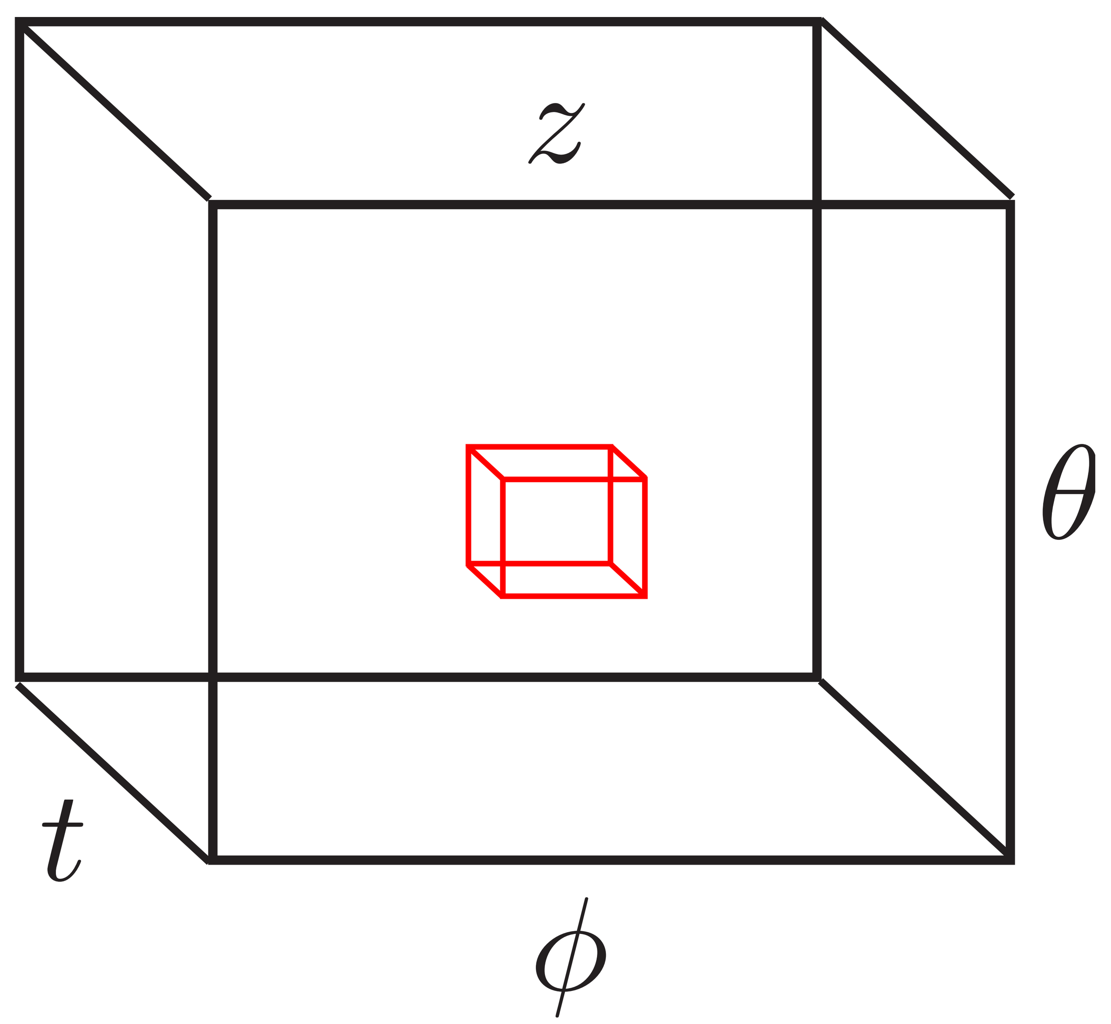
What is a token?

- Token is small neighborhood in space-time
  - Small for token attention / interaction to be informative
  - Big enough so token has rich internal strurcture



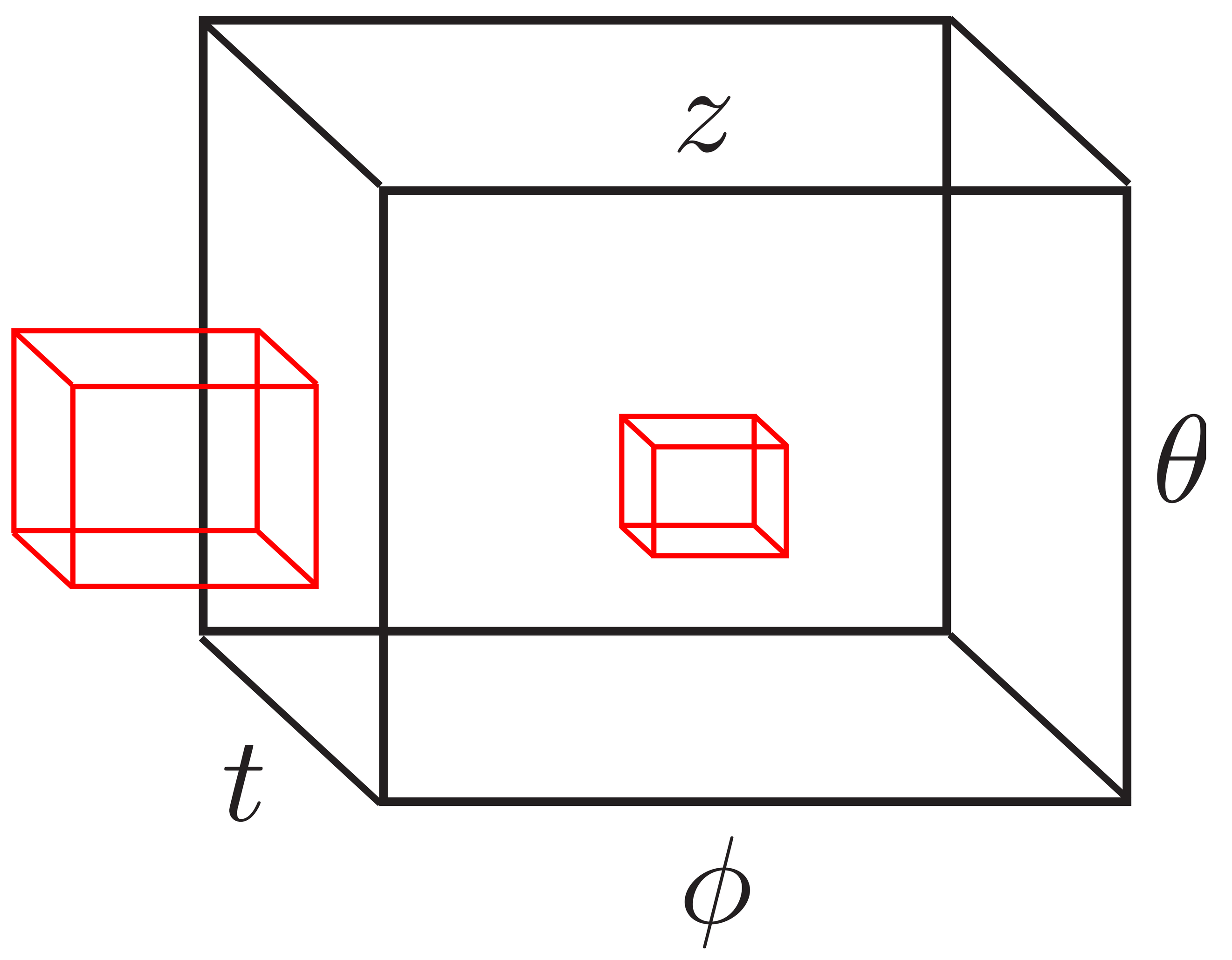
What is a token?

- Token is small neighborhood in space-time
  - Small for token attention / inter-action to be informative
  - Big enough so token has rich internal strurcture
- Token size is field-dependent

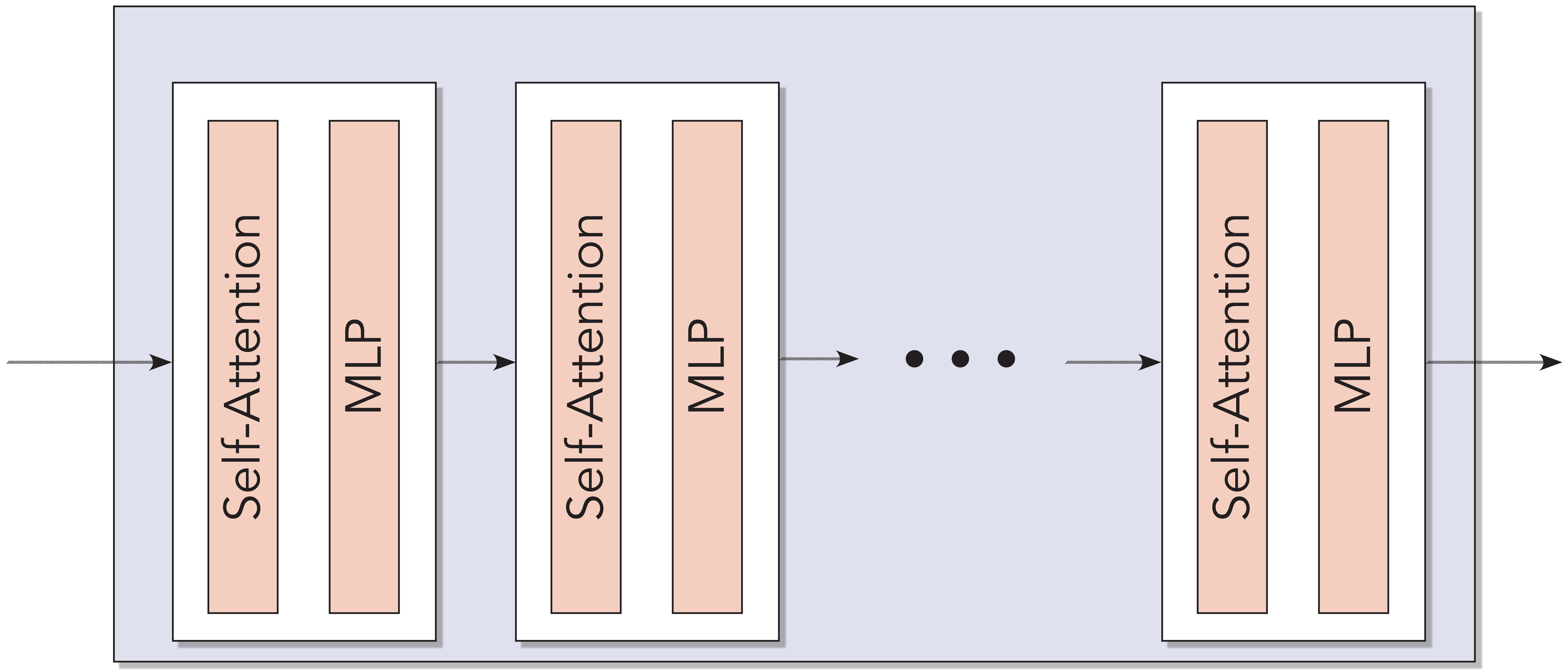


What is a token?

- Token is small neighborhood in space-time
  - Small for token attention / interaction to be informative
  - Big enough so token has rich internal strurcture
- Token size is field-dependent
- Multiple token sizes to provide multi-resolution structure and larger contexts



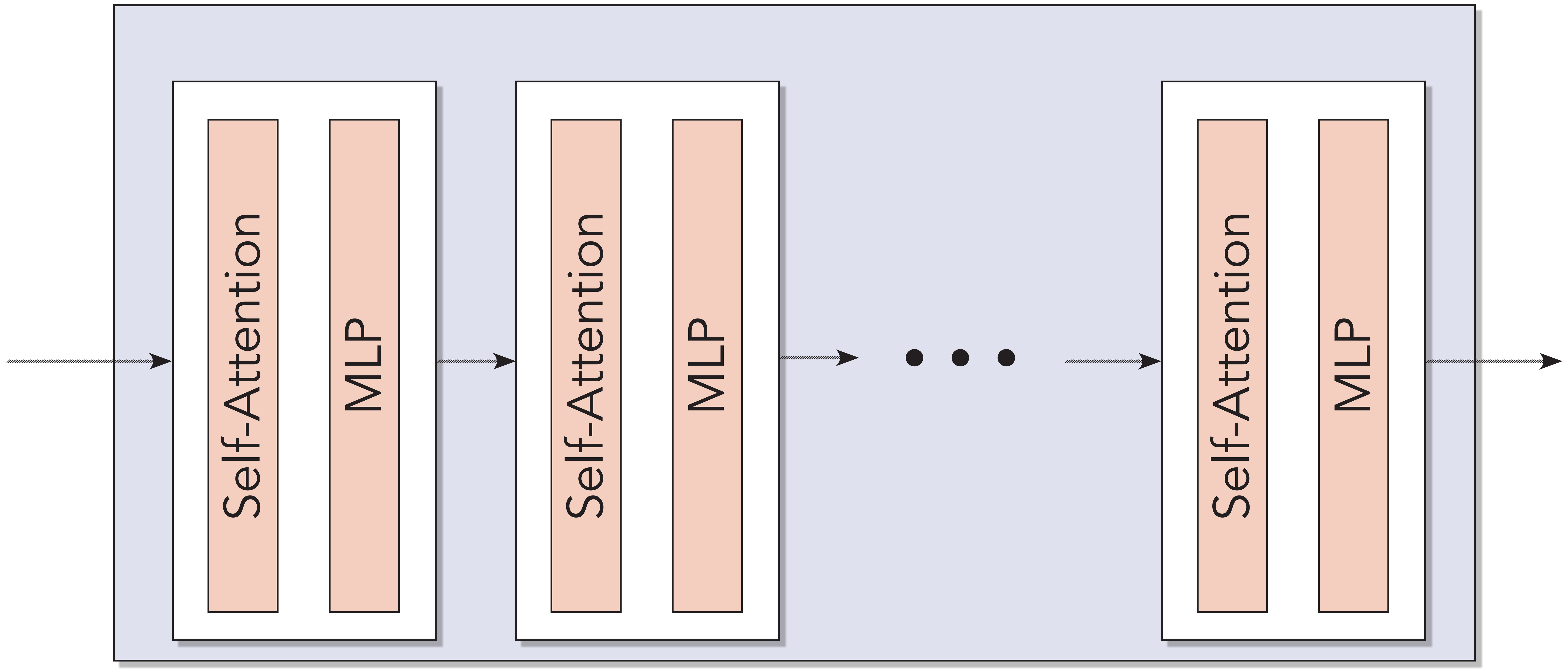
Multiformer



Multiformer

Self
attention

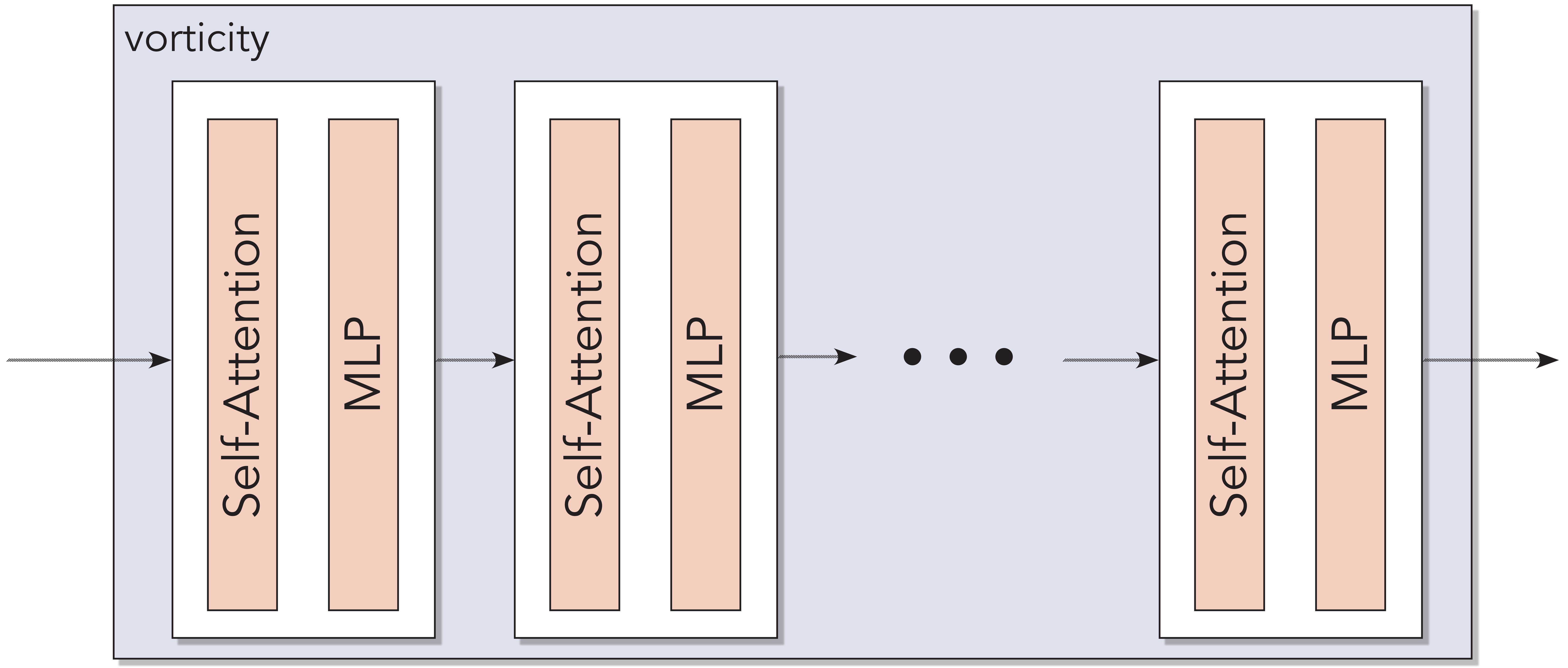
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

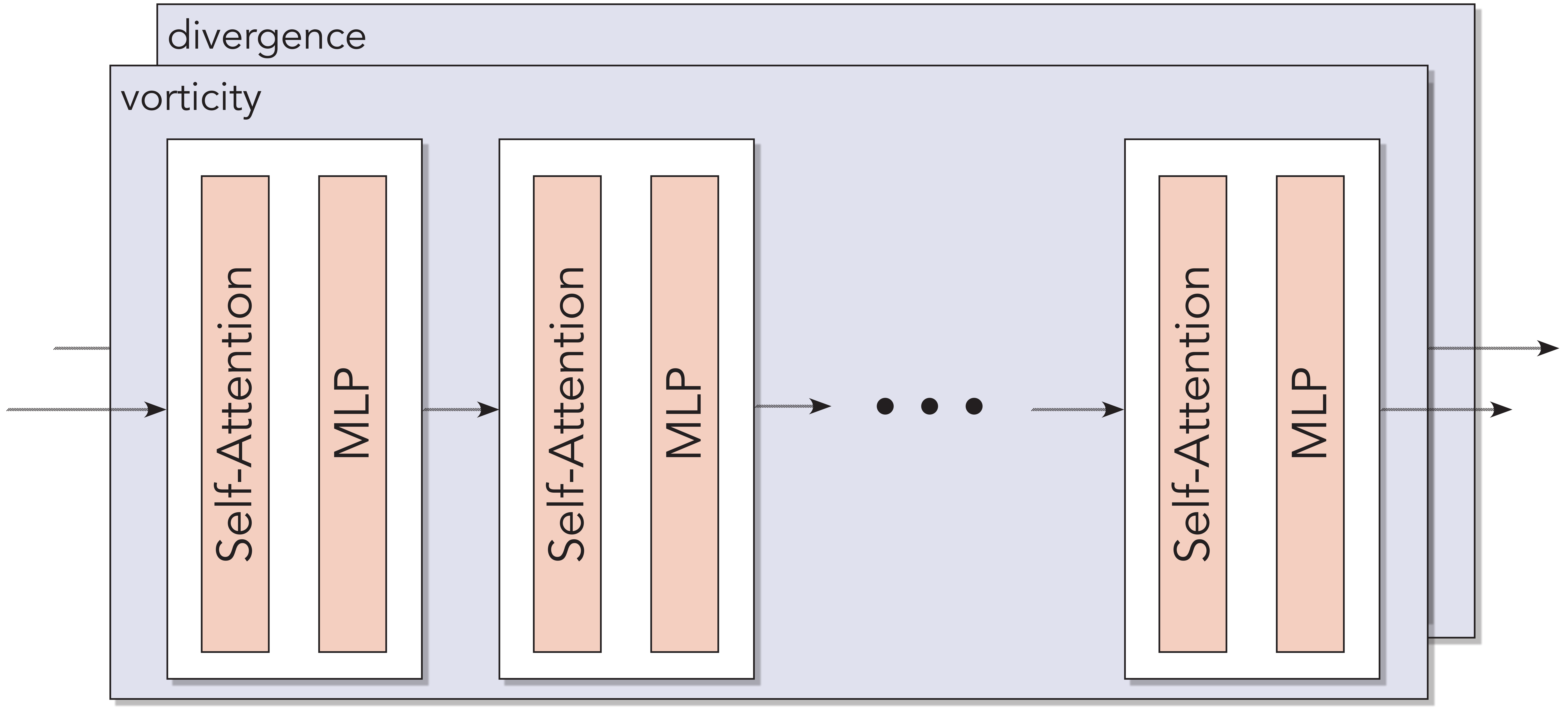
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

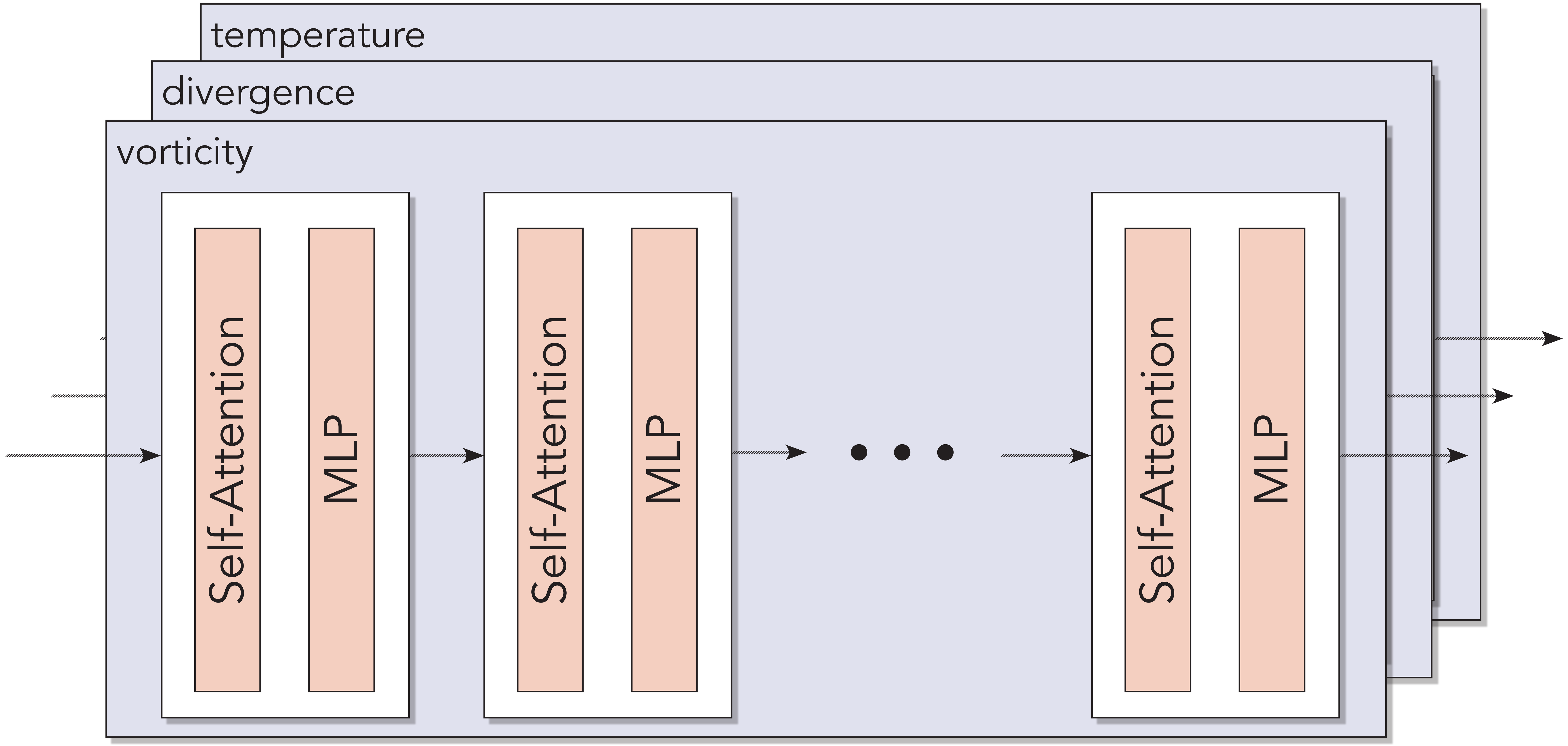
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

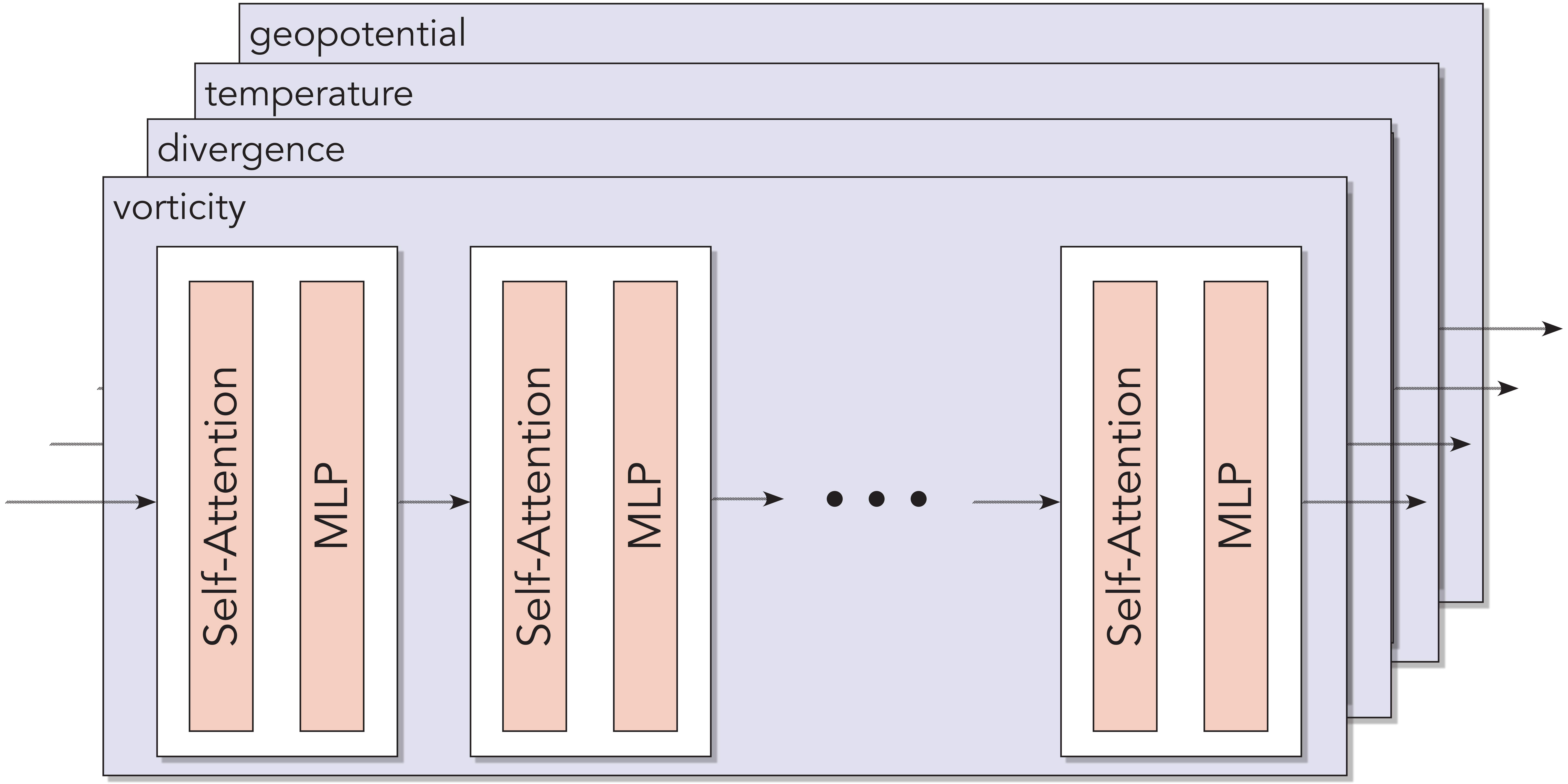
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

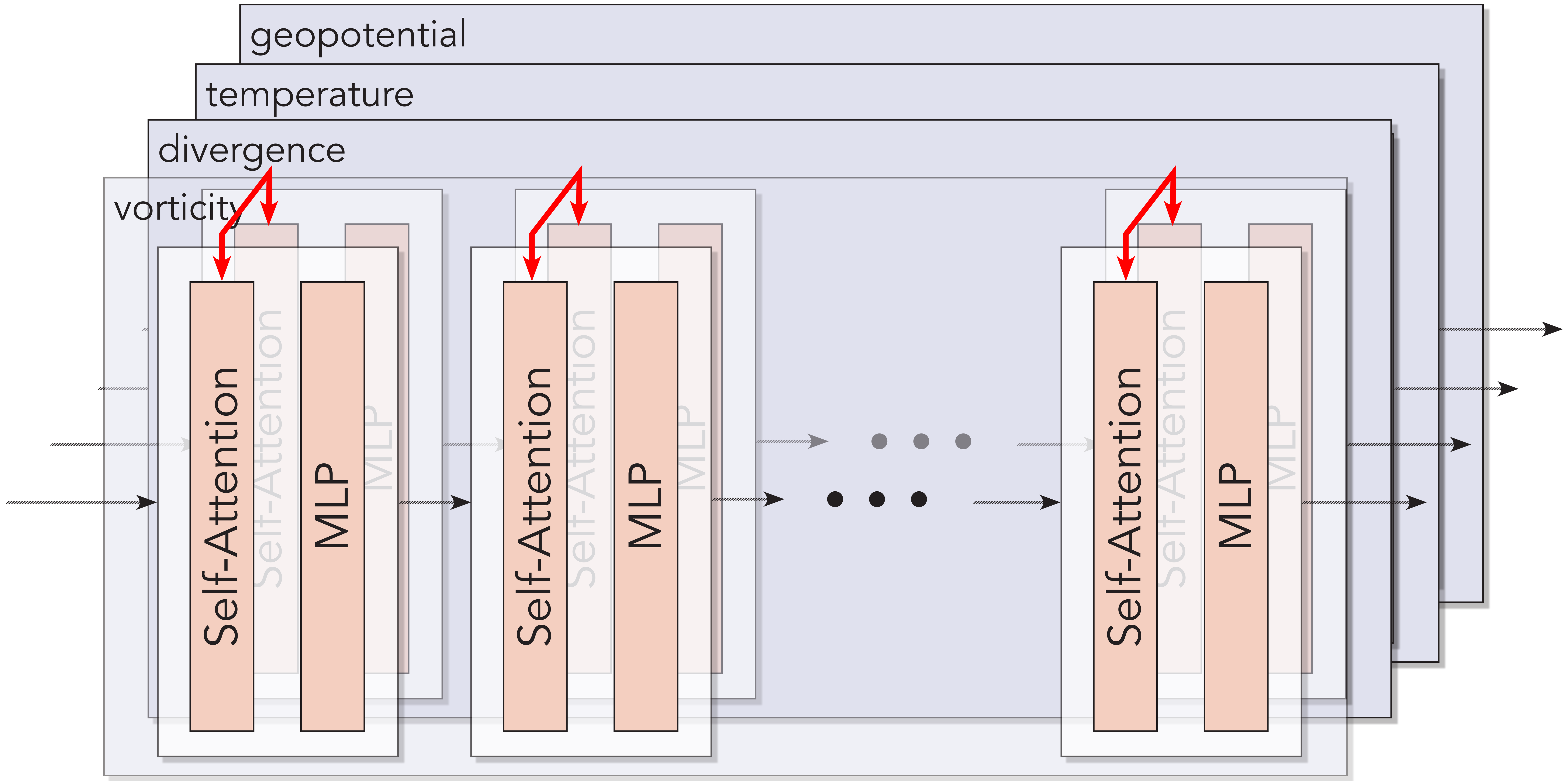
$$\sigma\left(Q\,K^{T}\right)V$$



Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)\,V$$



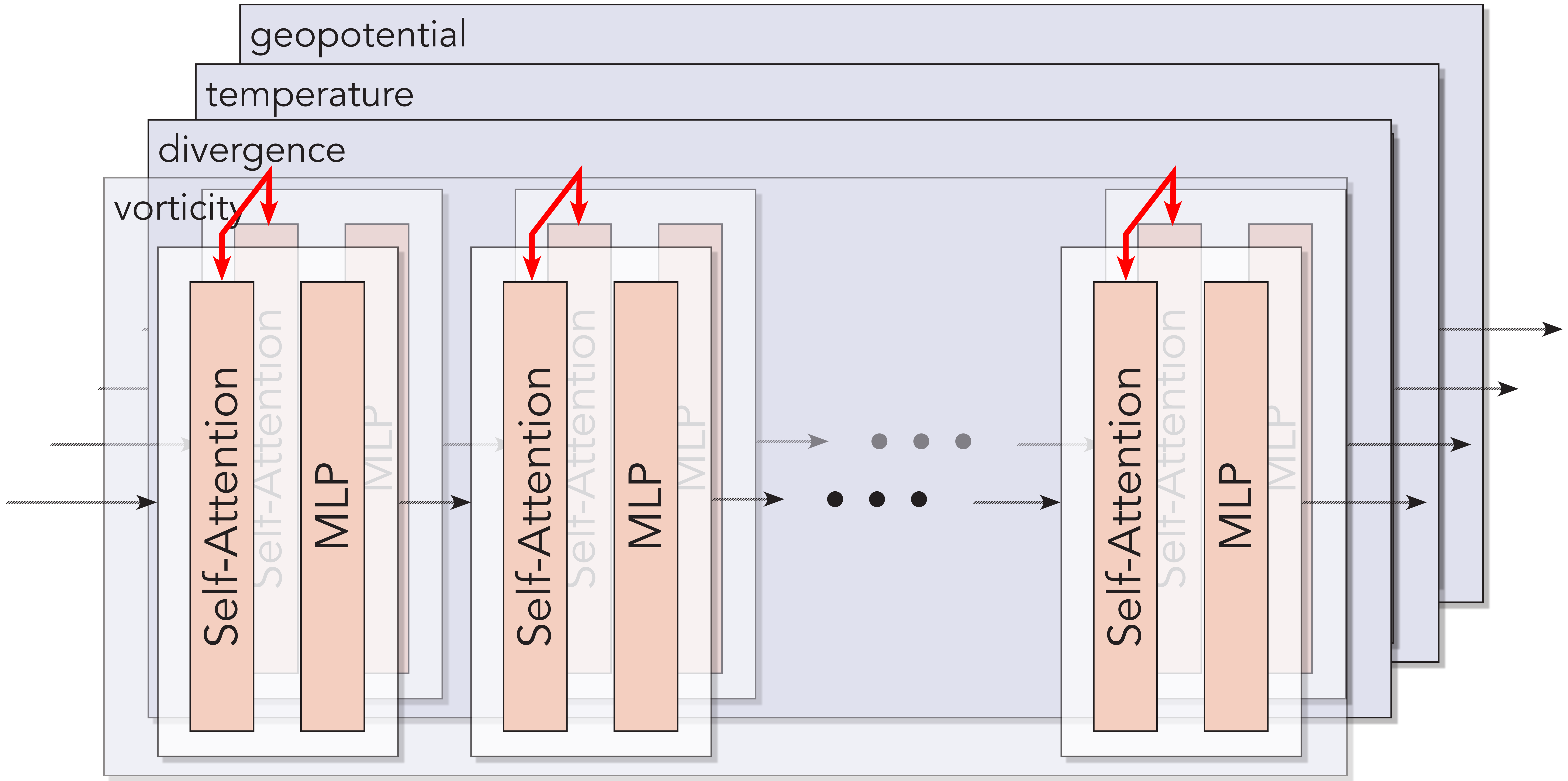
Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)\,V$$

Cross
attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)\,V_\mu$$



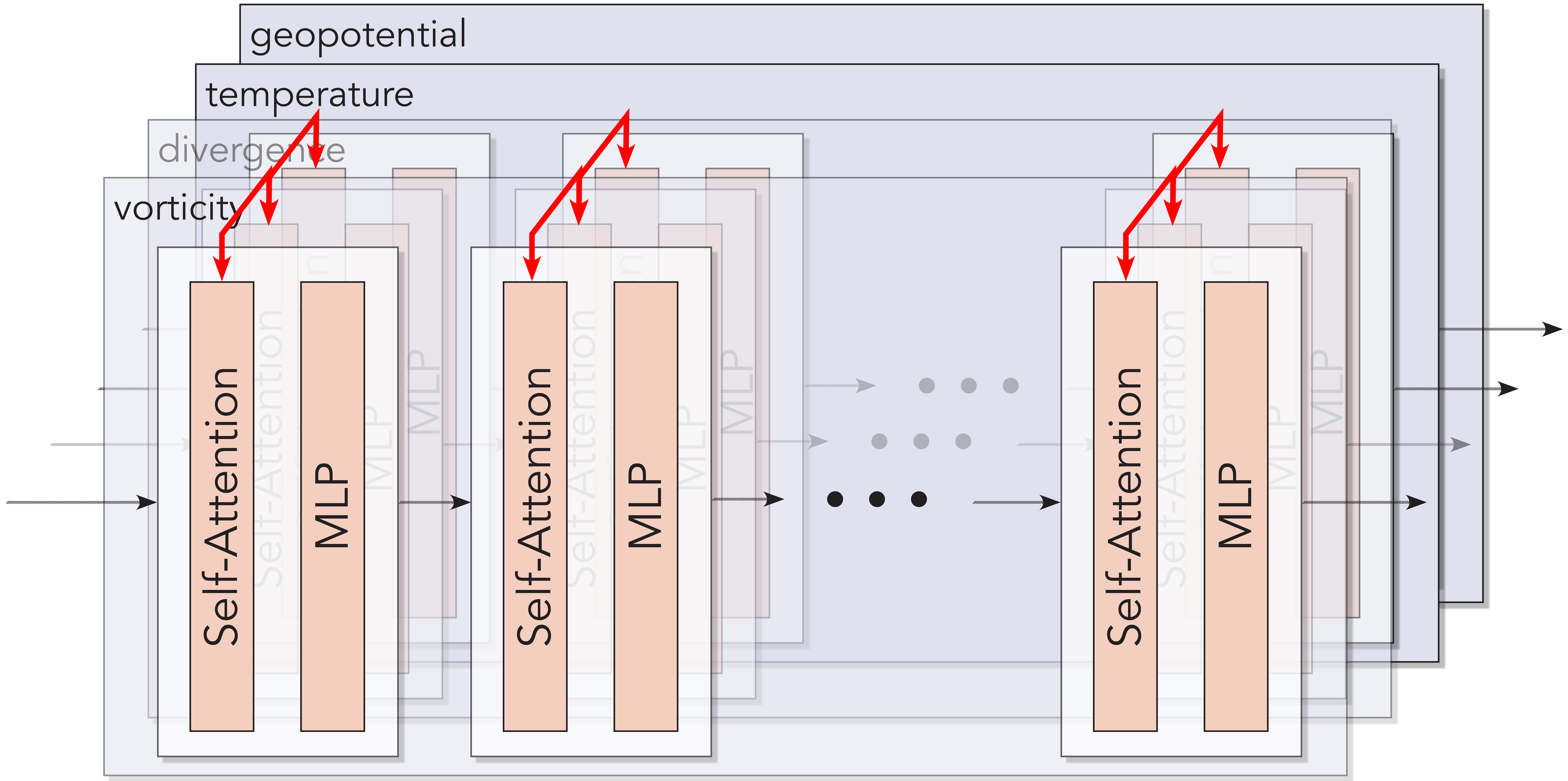
Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)\,V$$

Cross
attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)\,V_\mu$$



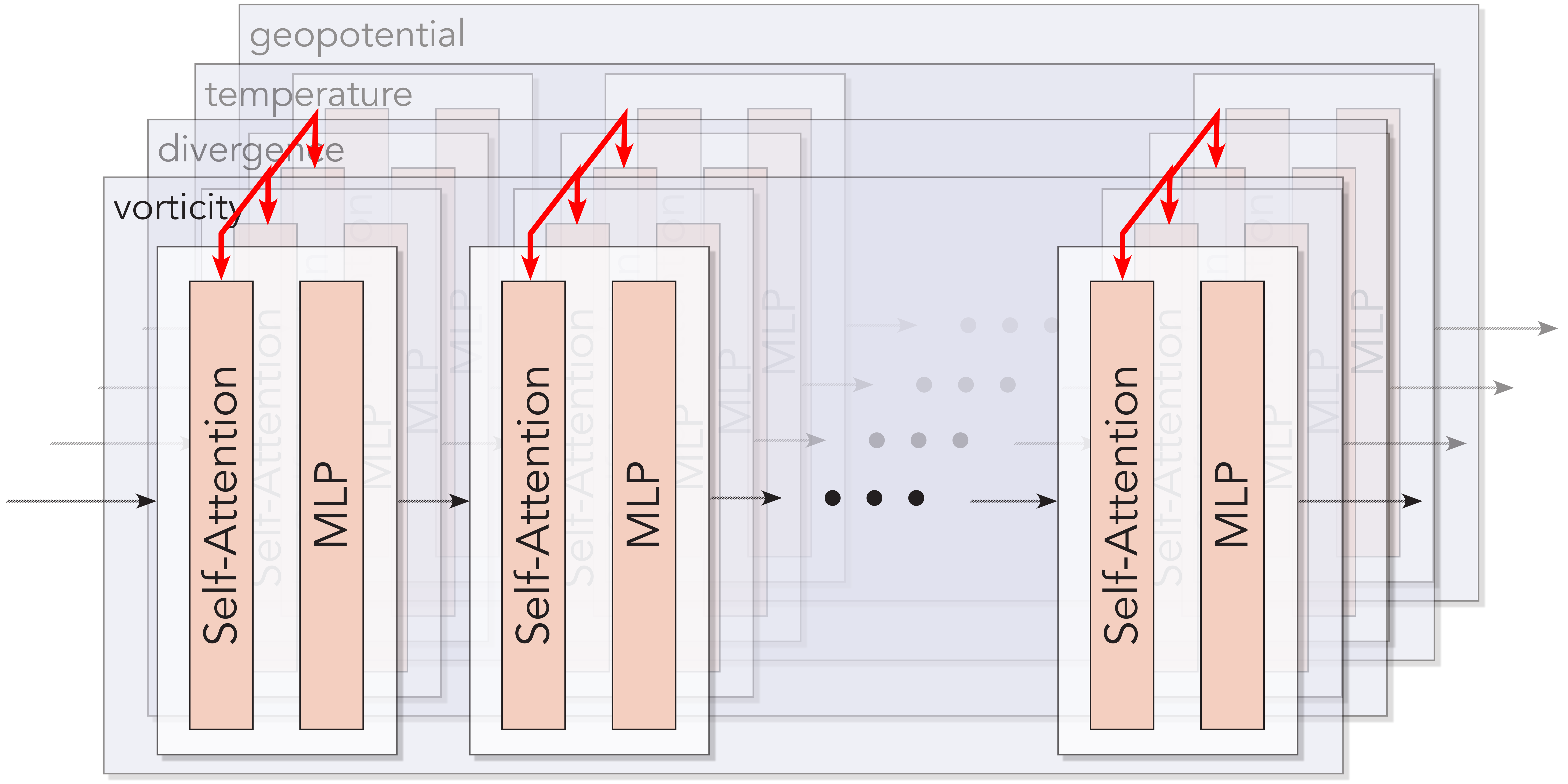
Multiformer

Self
attention

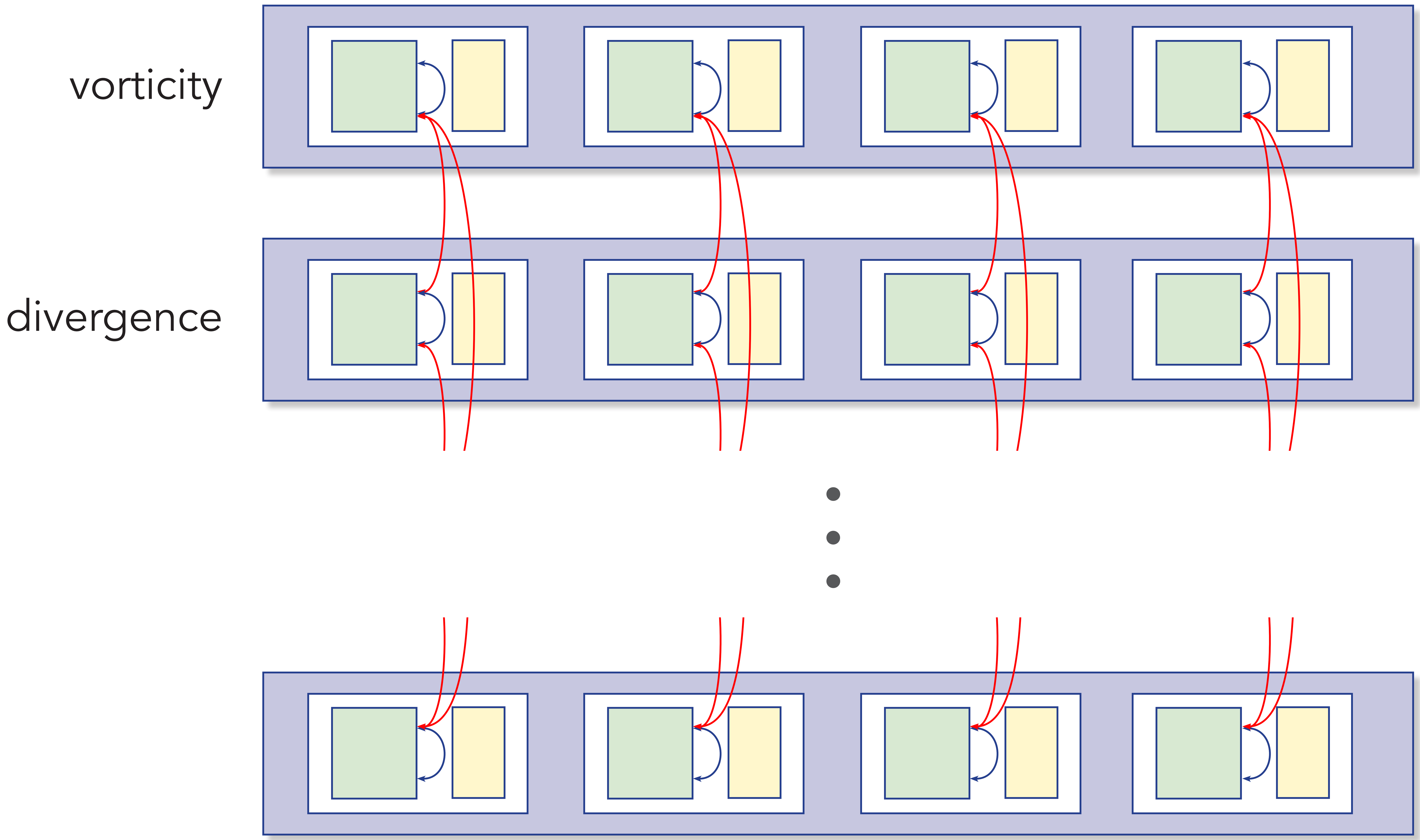
$$\sigma\left(Q\,K^T\right)\,V$$

Cross
attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)\,V_\mu$$



Multiformer

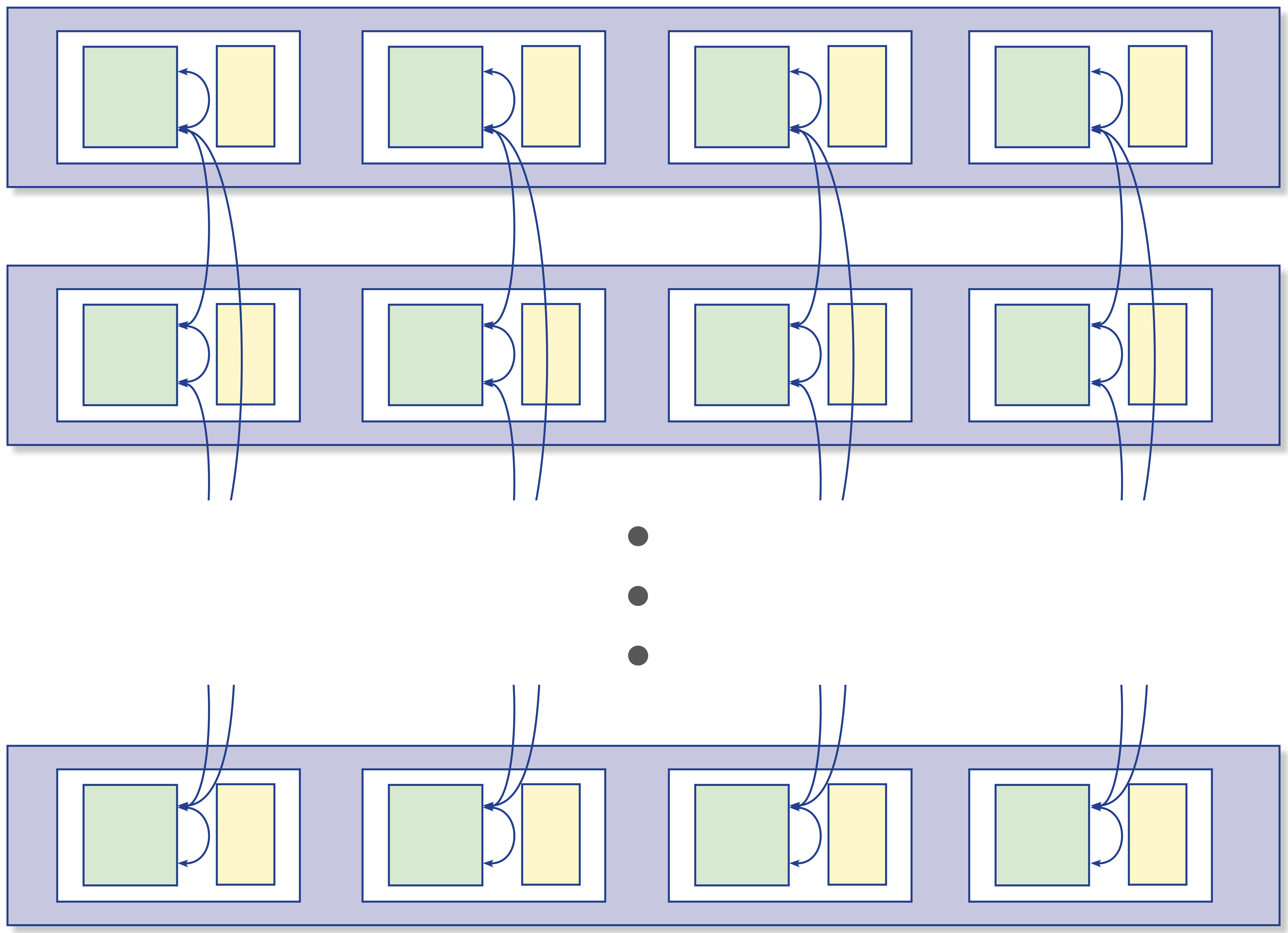


Multiformer

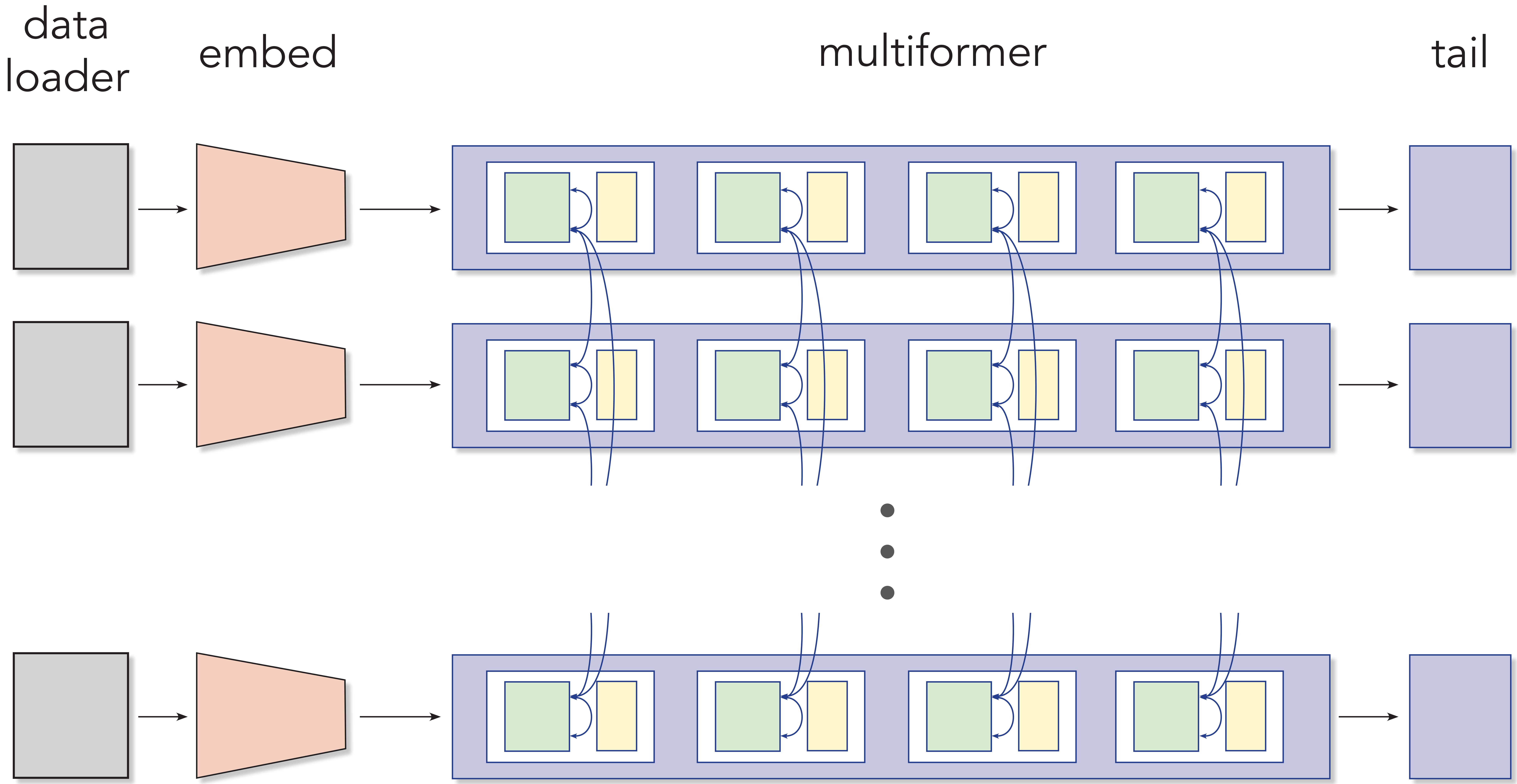
- Different physical fields with different properties have separate latent spaces (and transformations for these)
- Individual fields can be pre-trained independently
- Plug-and-play of fields
  - Fields can be added/removed with limited (or no) computational effort
- Cross-attention allows for explicit introspection of inter-action between fields

Embedding of tokens

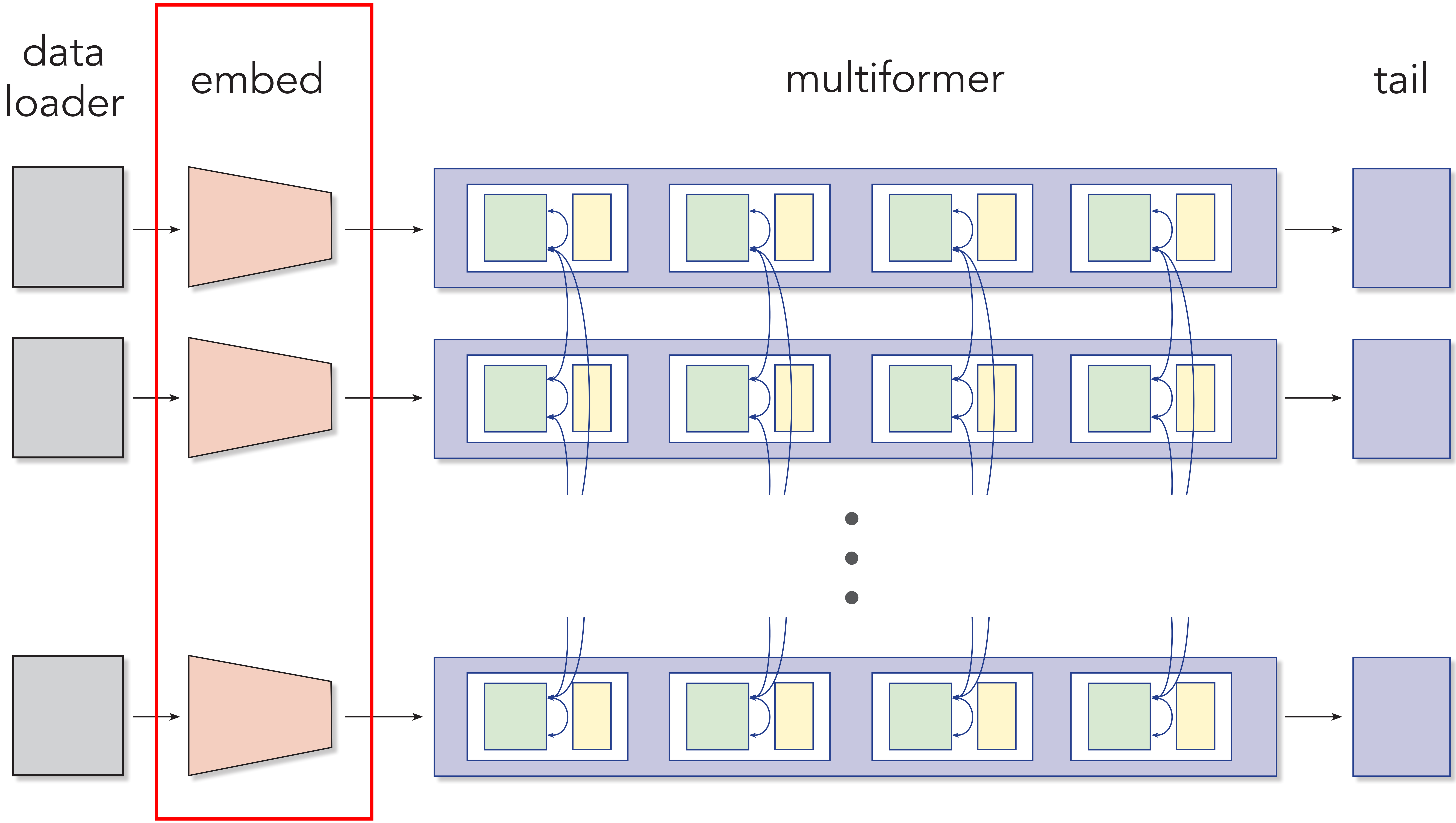
multiformer



Embedding of tokens



Embedding of tokens



Embedding network

- Multiformer models longer range effects and field inter-actions in a rich latent space

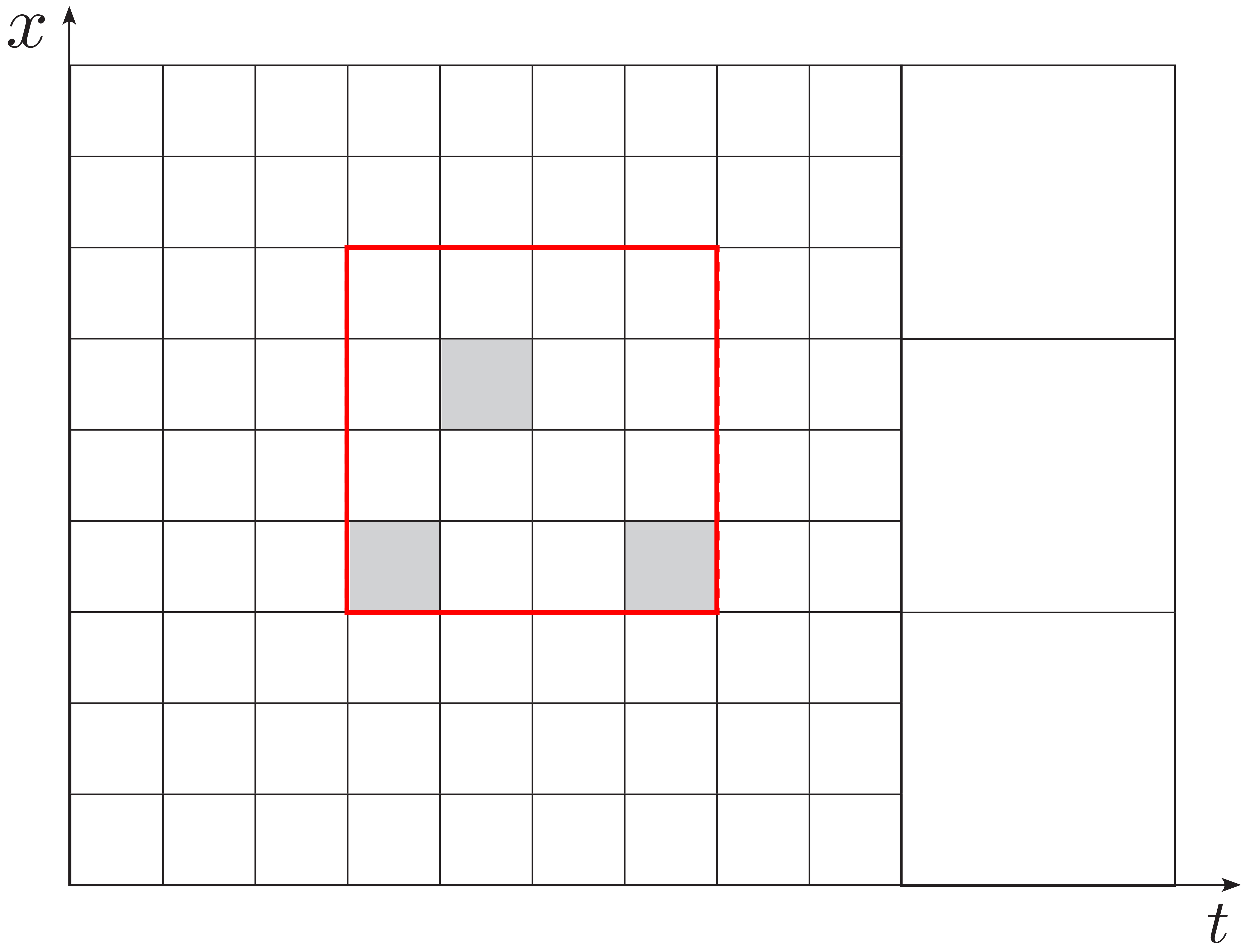
Embedding network

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field

Embedding network

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

Embedding of tokens



Embedding of tokens

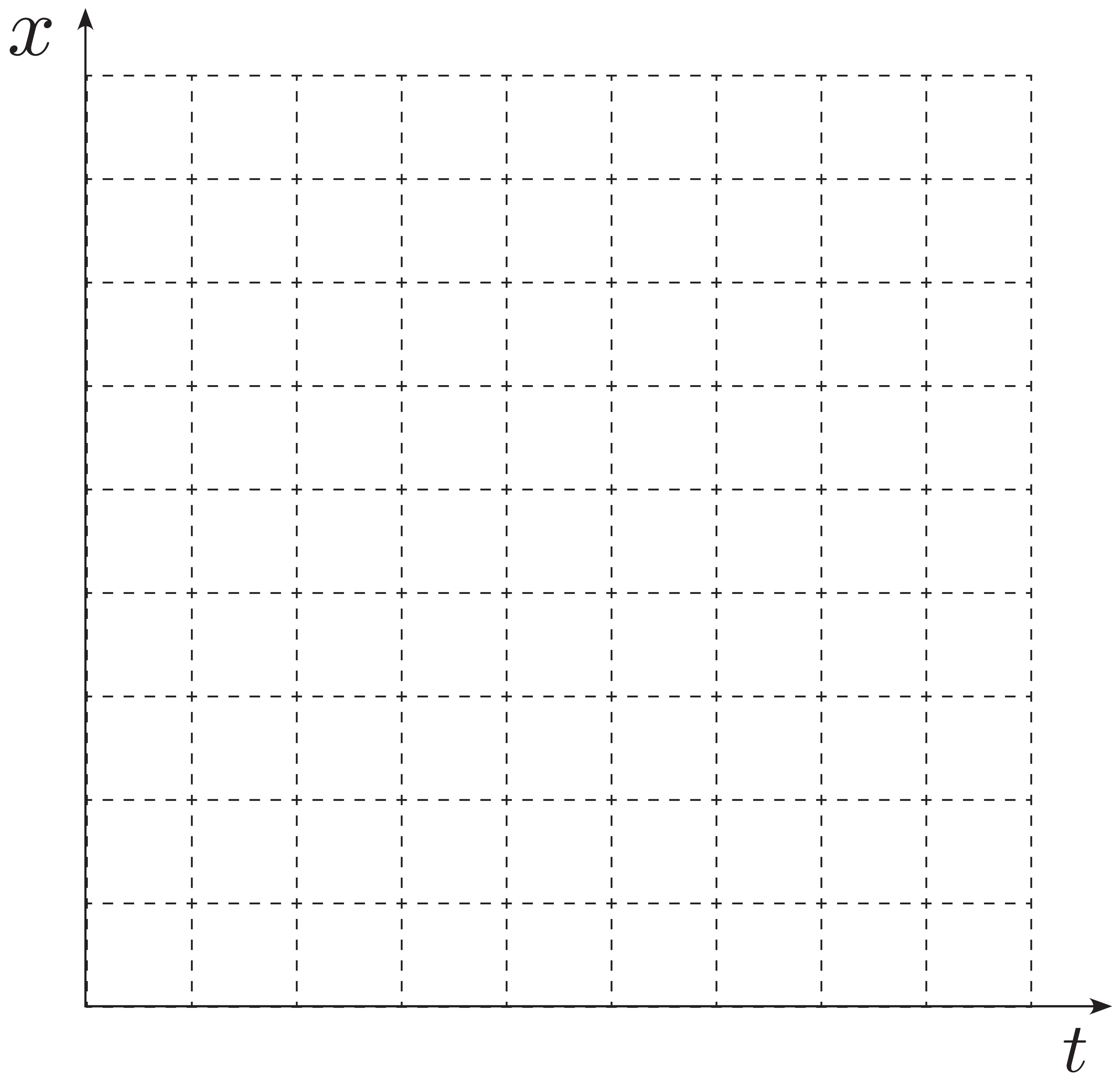
- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

Embedding of tokens

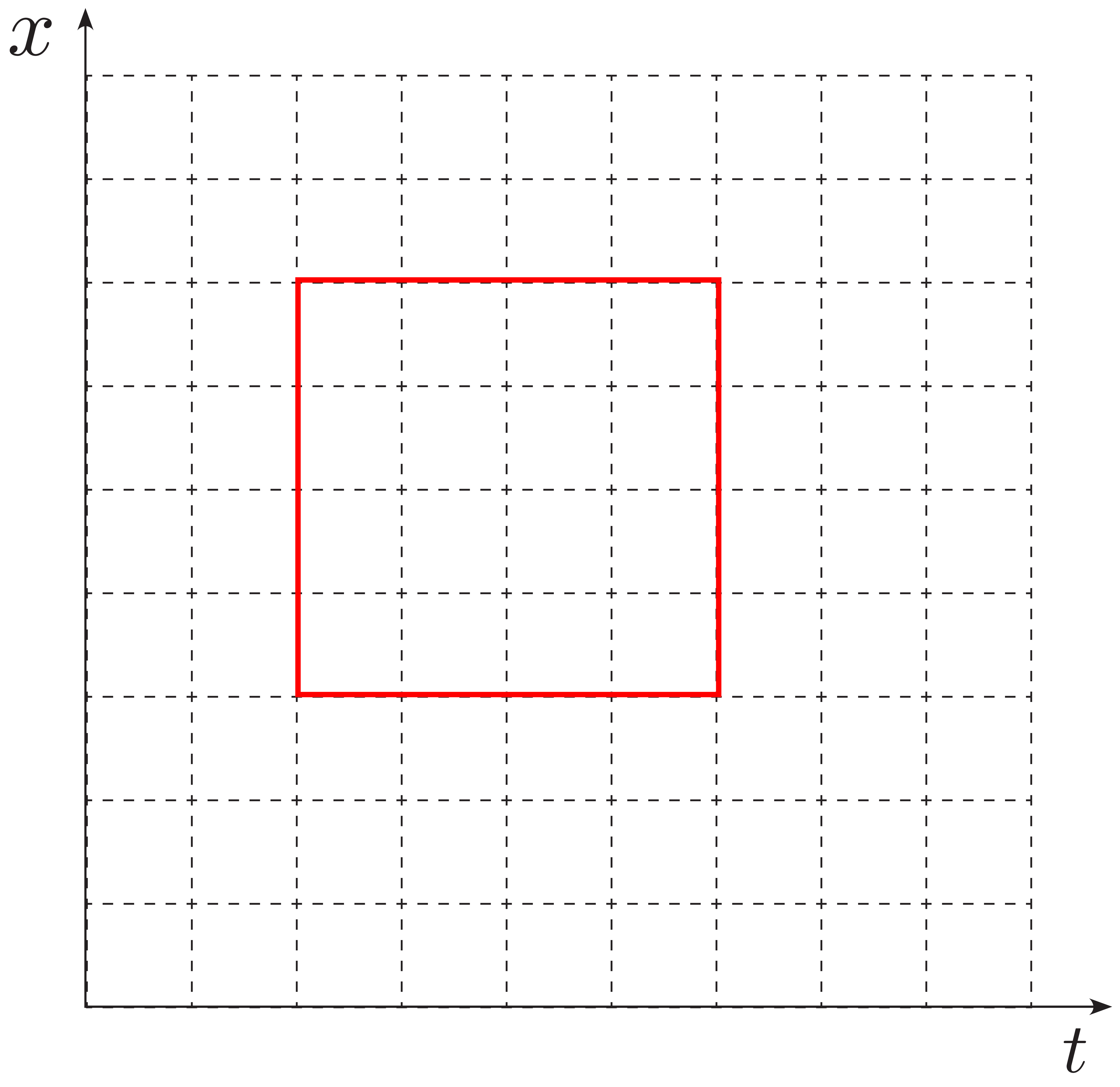
- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

$\Rightarrow$ Use transformer as embedding network

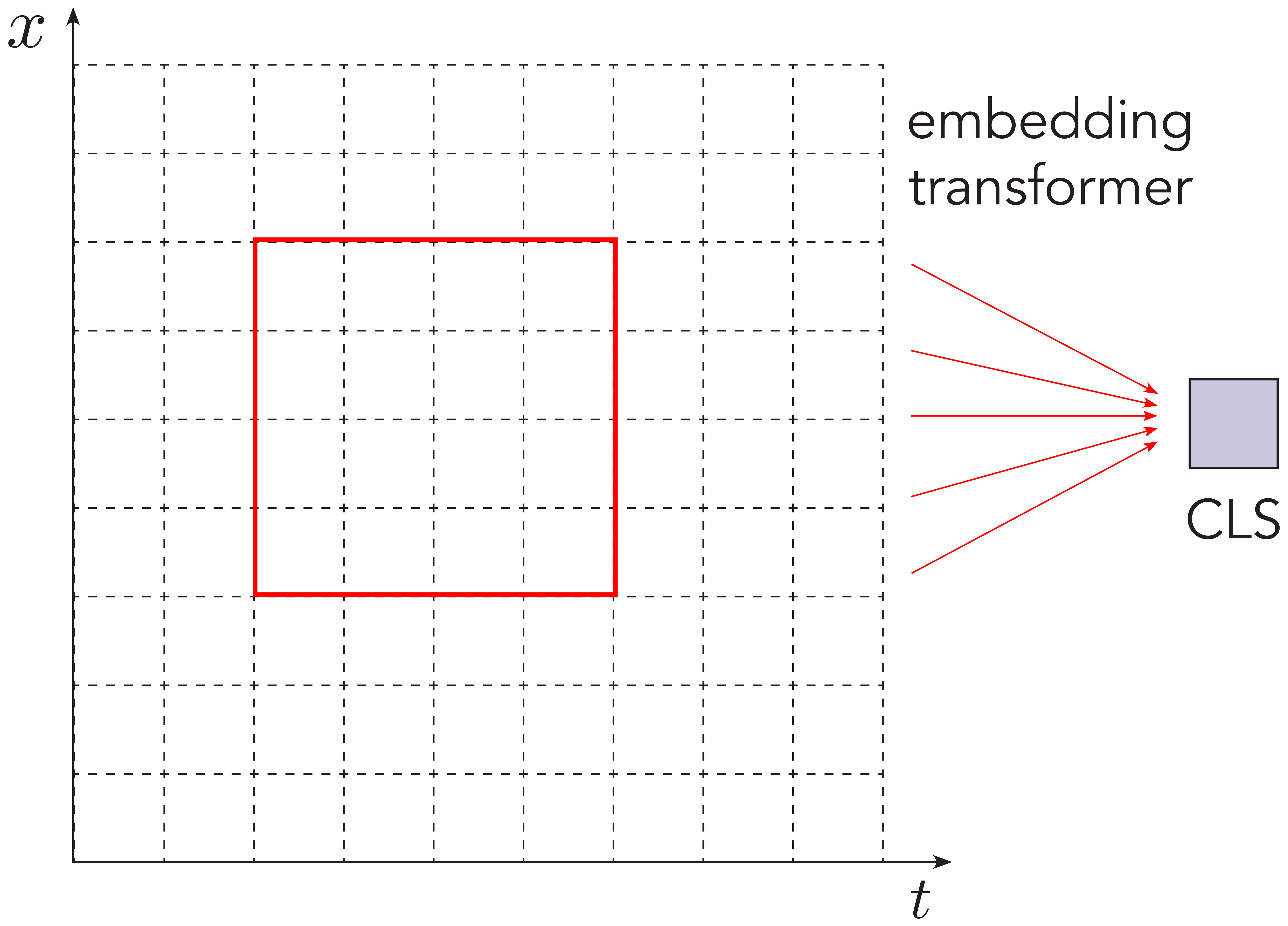
Embedding of tokens



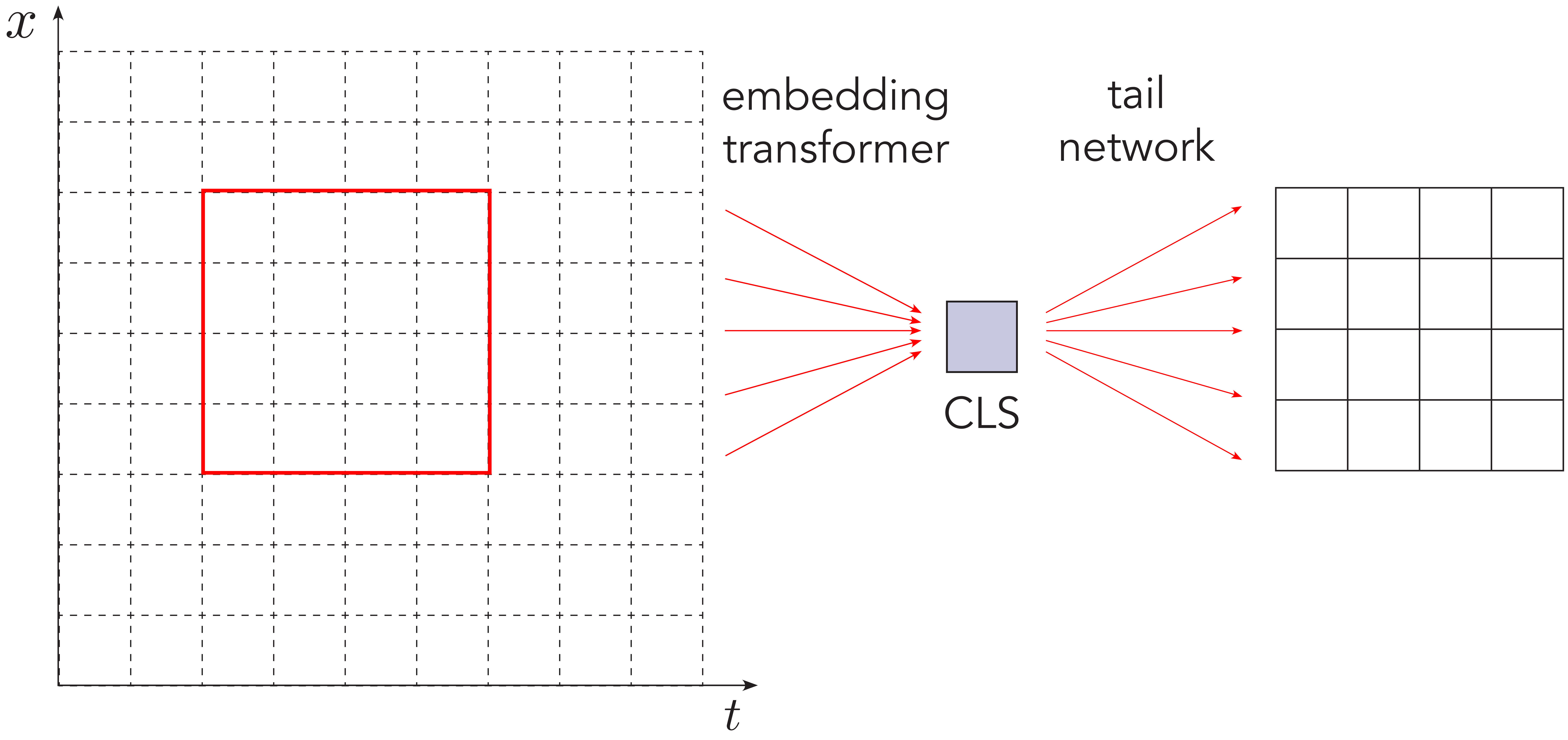
Embedding of tokens



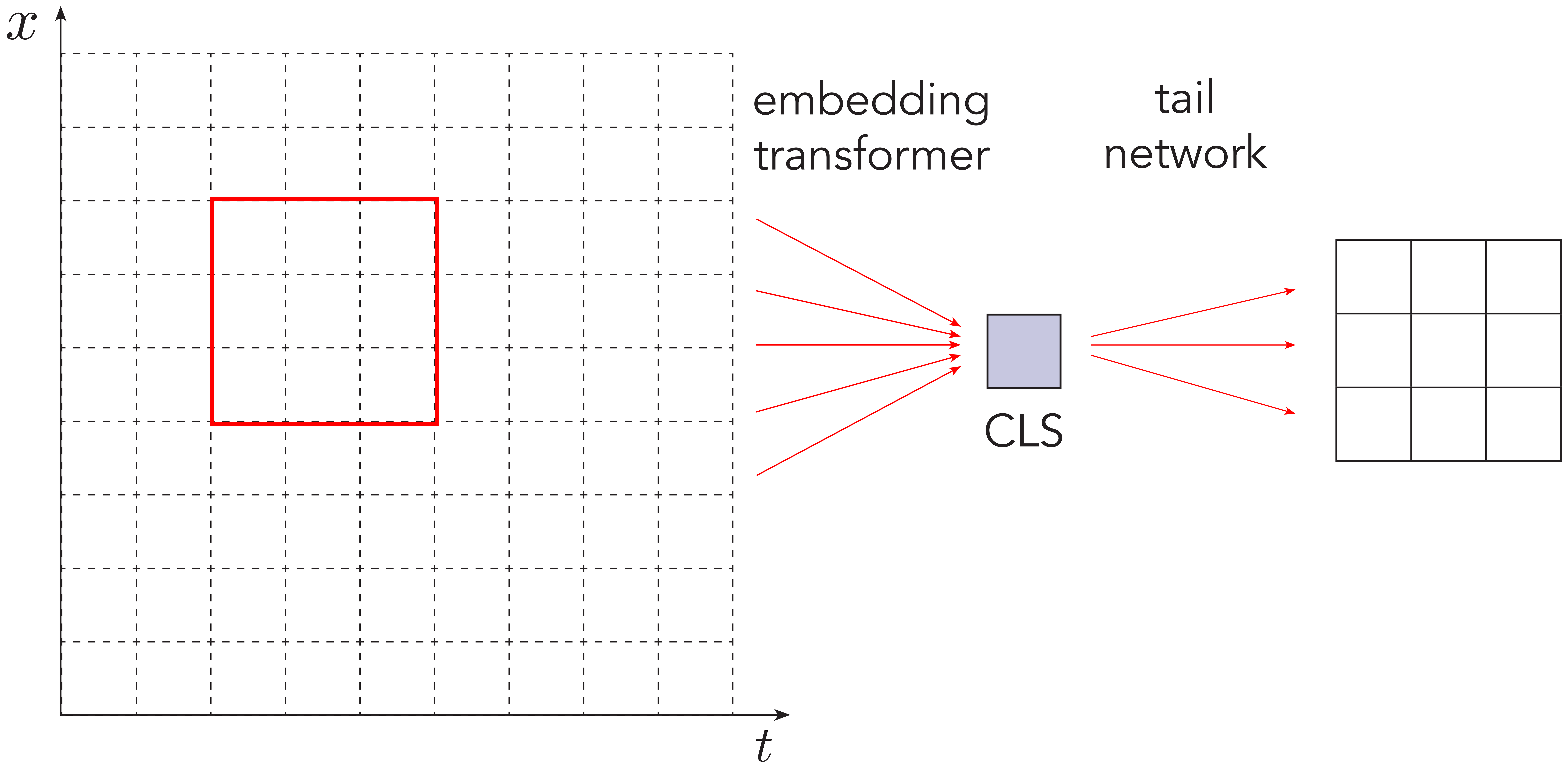
Embedding of tokens



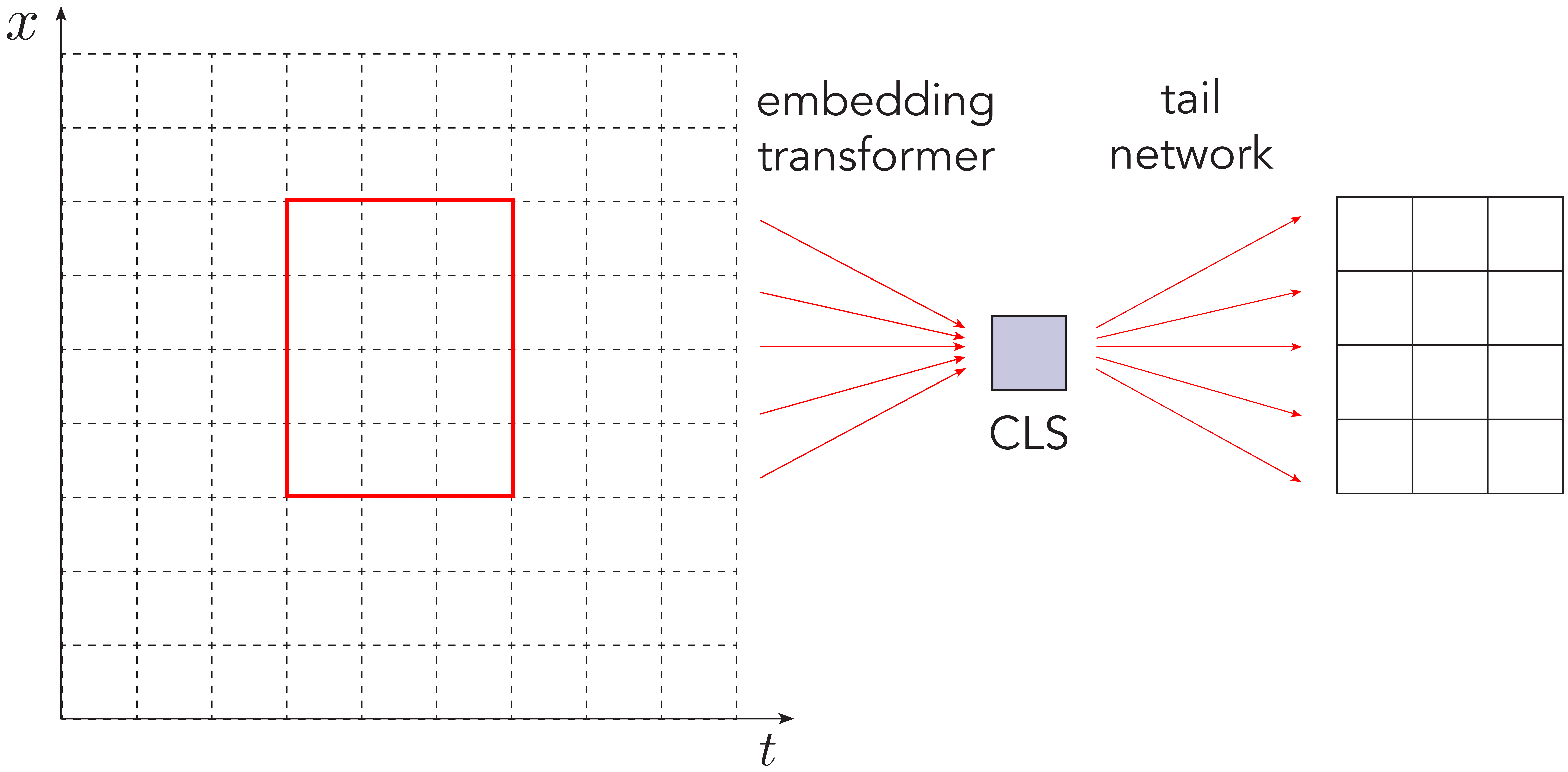
Embedding of tokens



Embedding of tokens



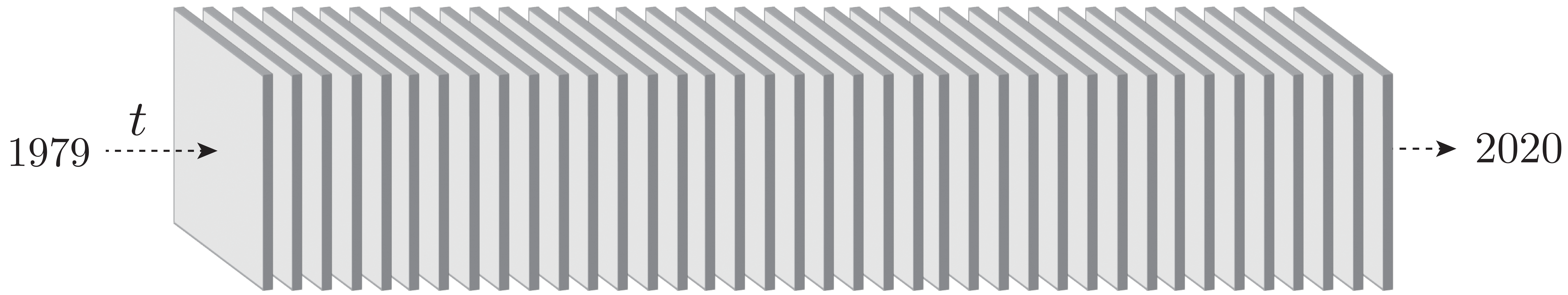
Embedding of tokens



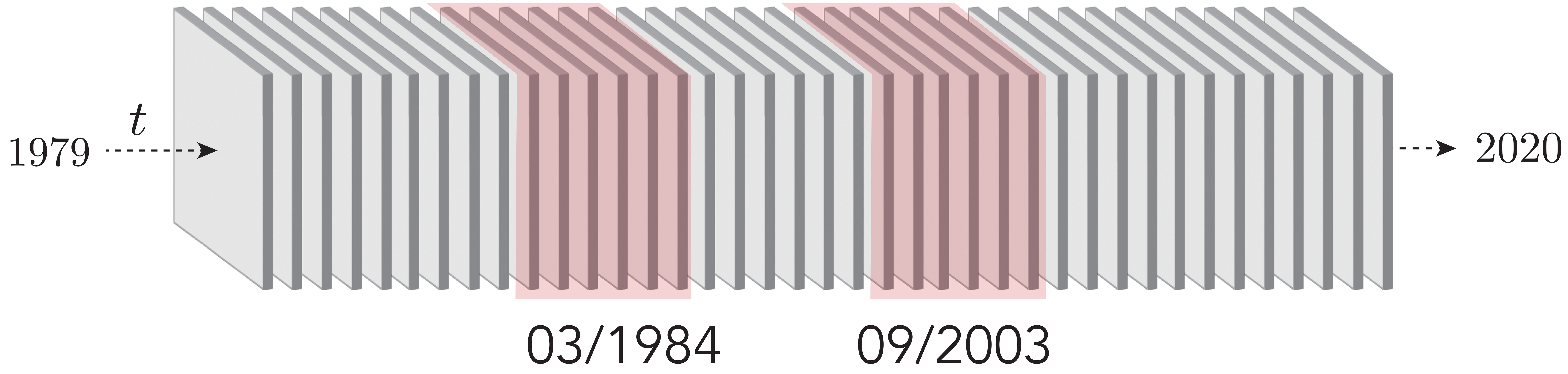
Training

- Unbiased hierarchical Monte Carlo sampling of all possible ERA5 space-time cubes

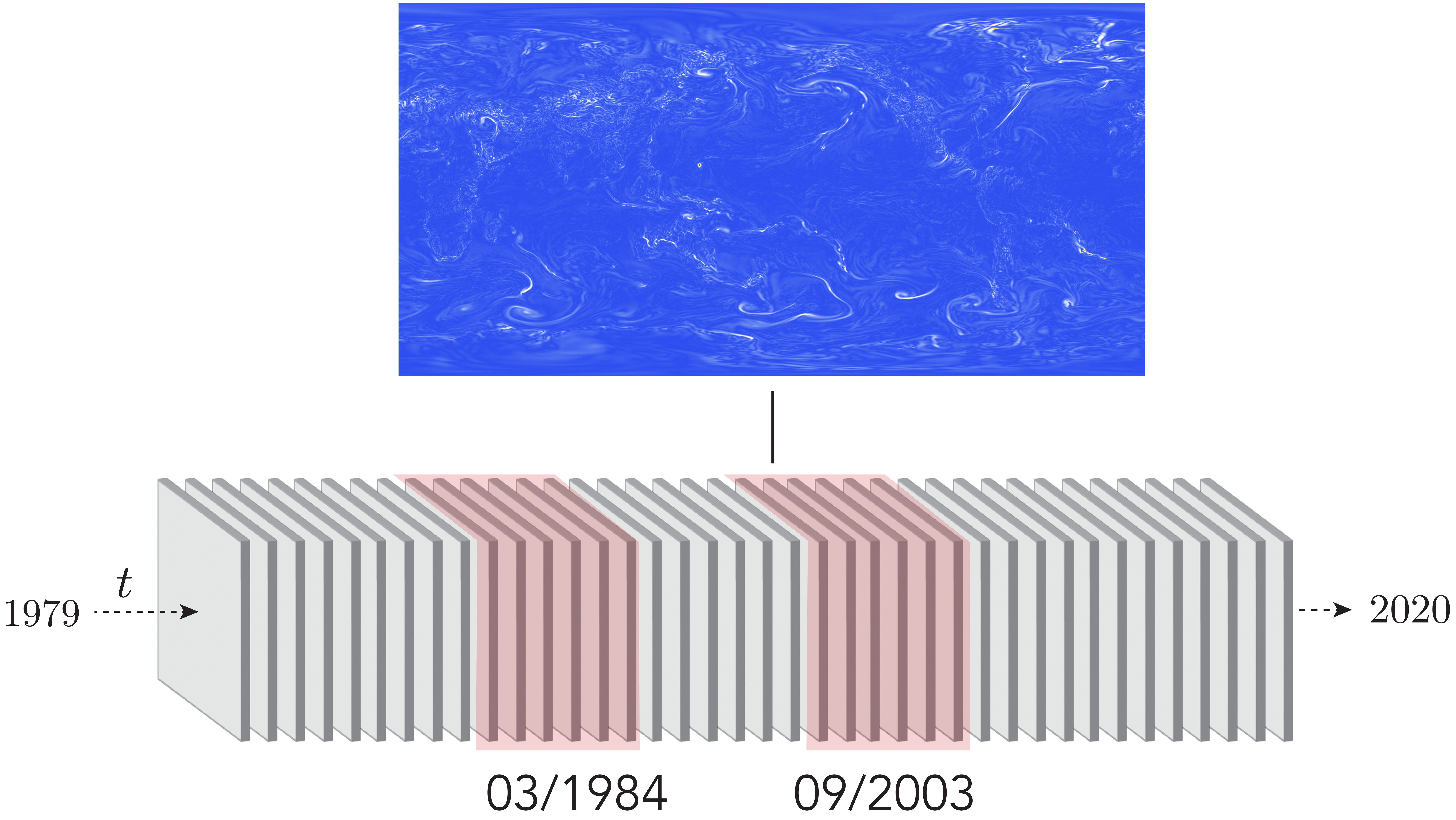
Training



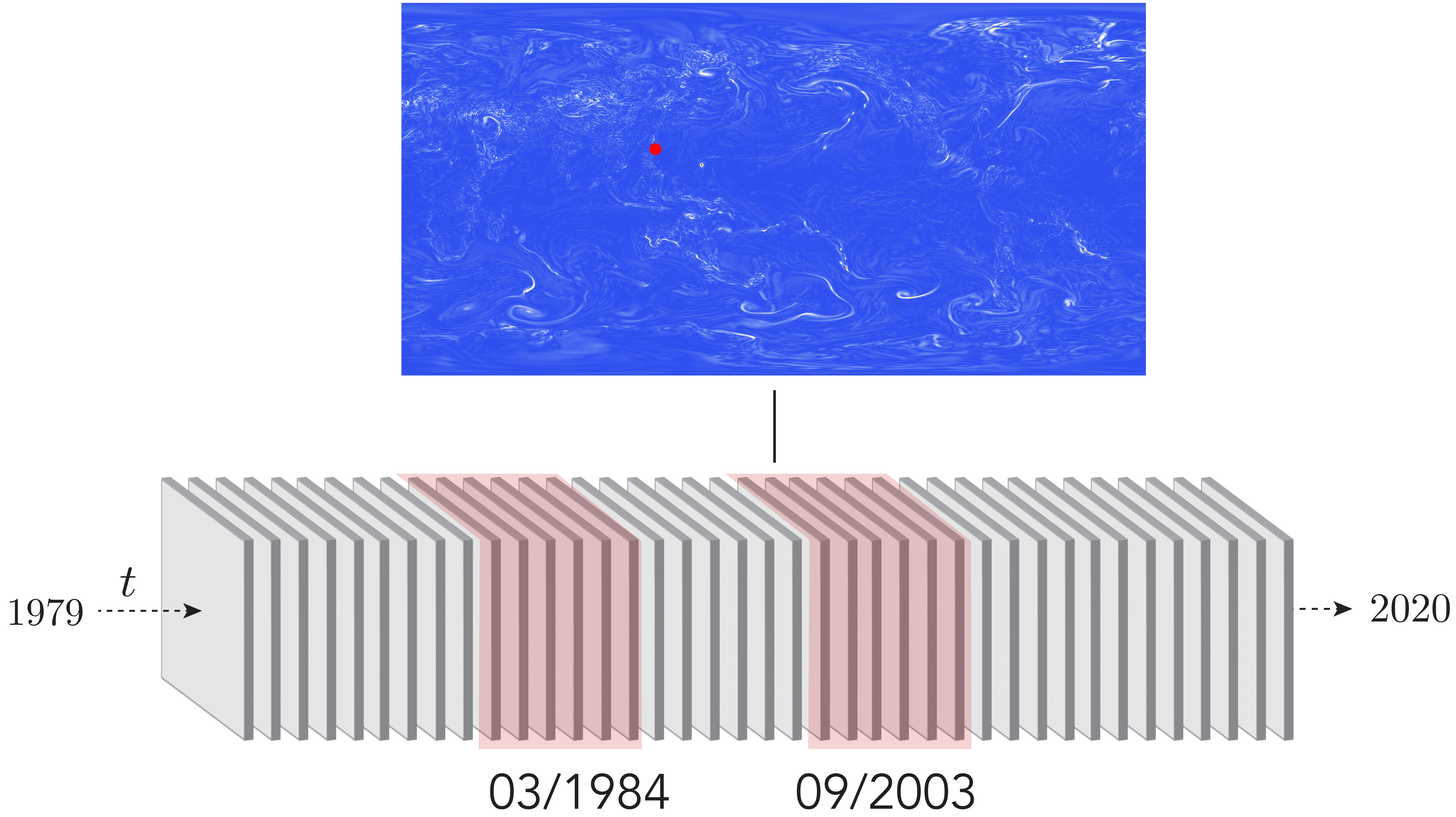
Training



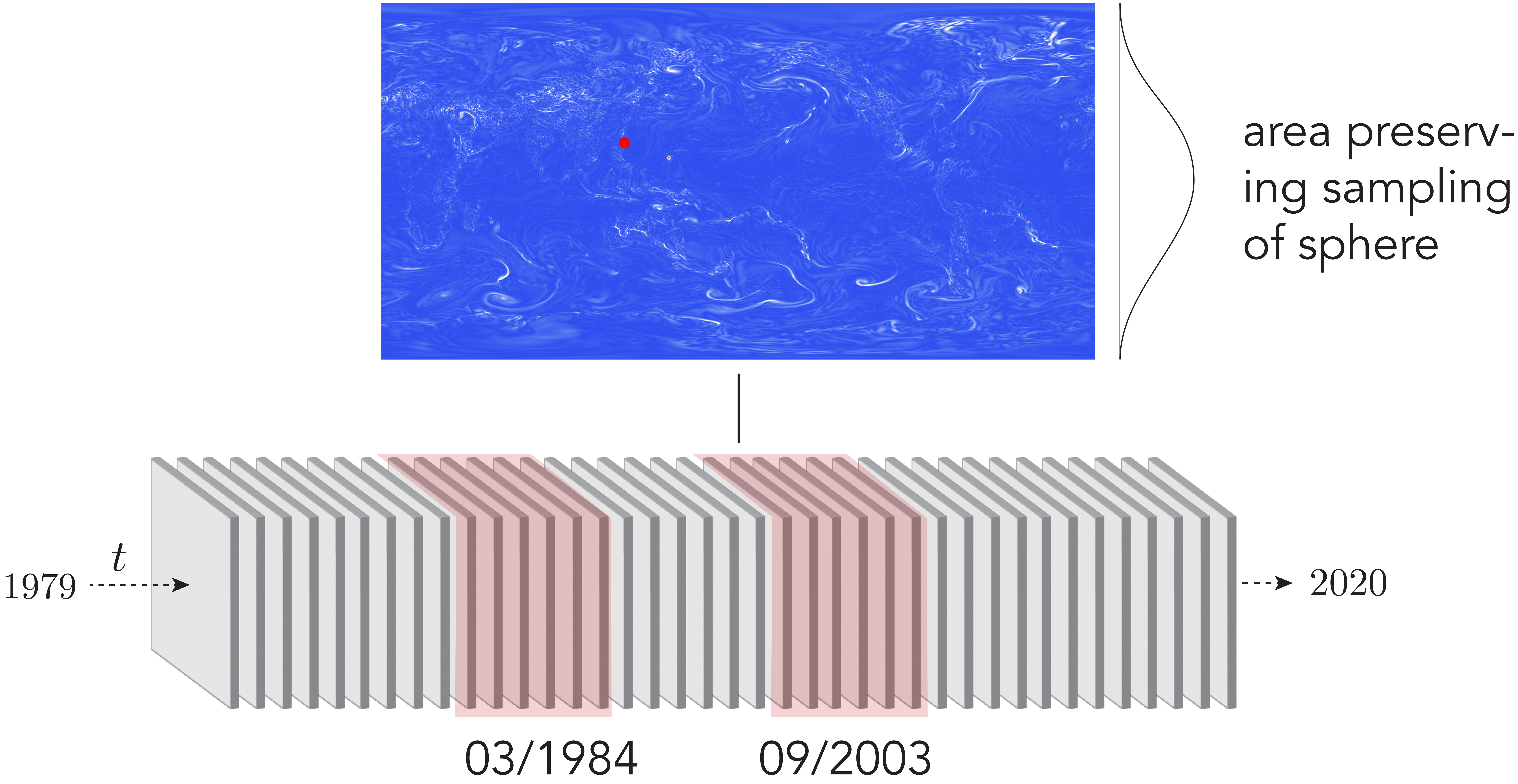
Training



Training

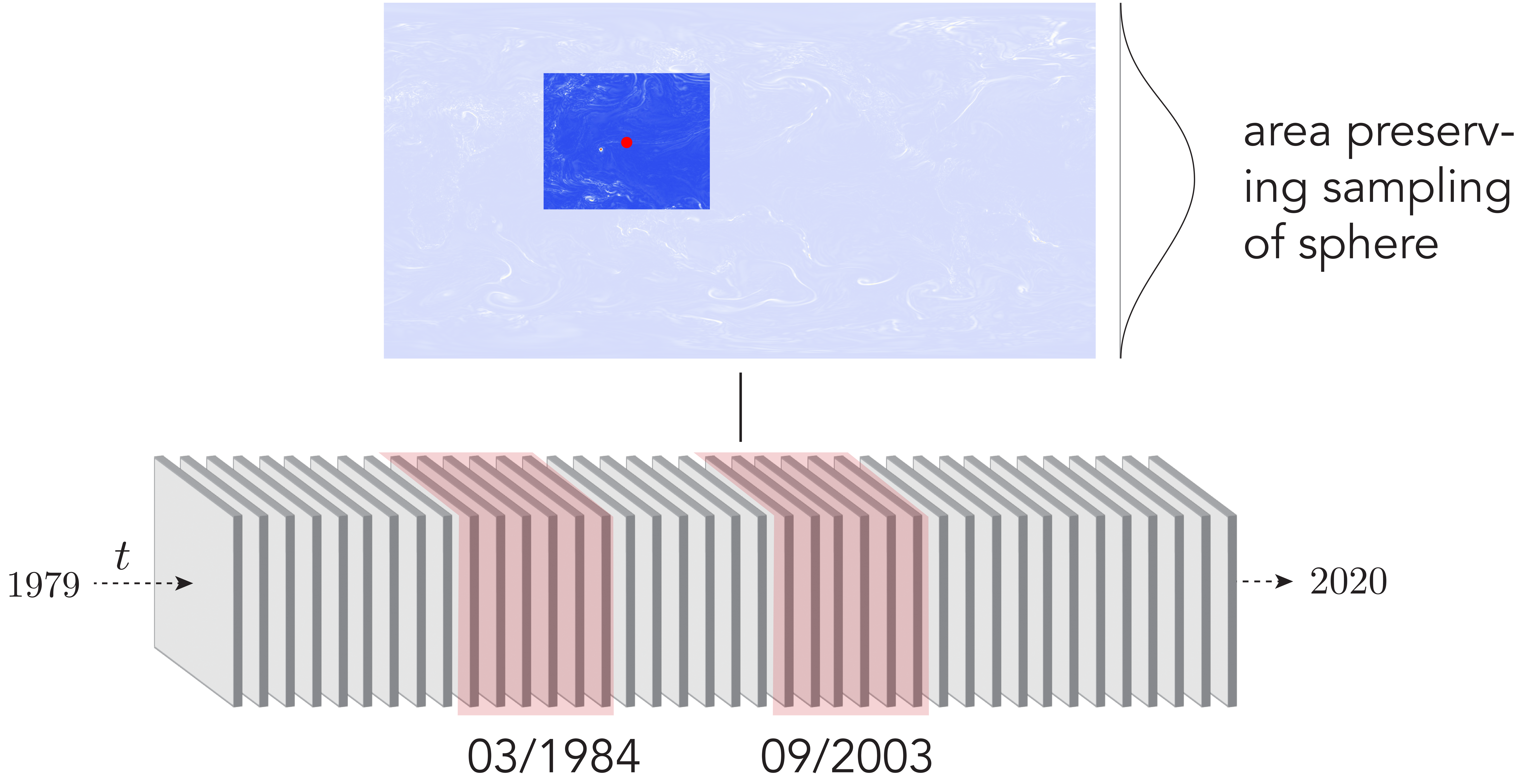


Training



area preserv-
ing sampling
of sphere
1979
t
2020
03/1984
09/2003

Training


area preserv-
ing sampling
of sphere
1979
t
2020
03/1984
09/2003

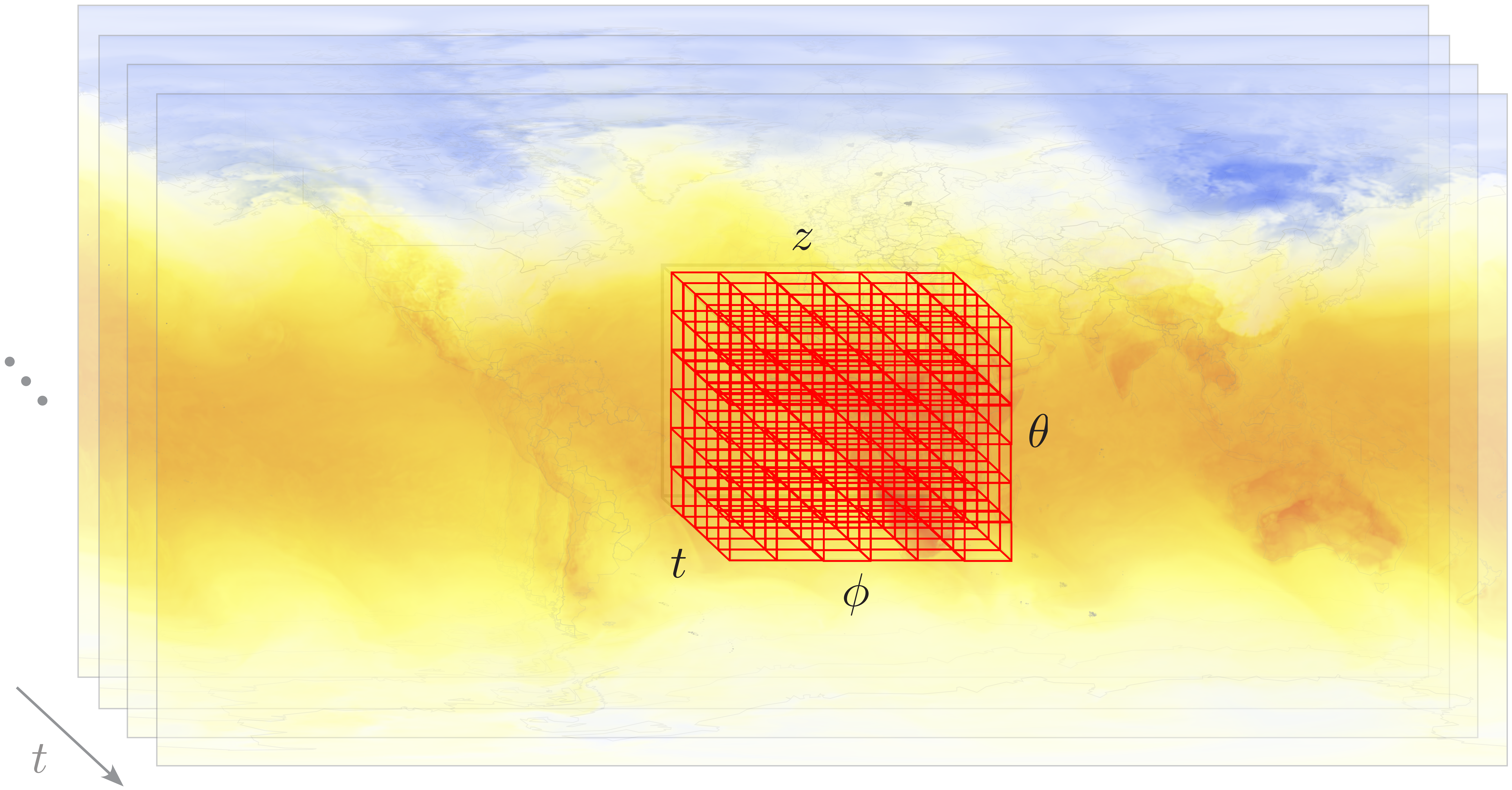
Training

- Unbiased hierarchical Monte Carlo sampling of all possible ERA5 space-time cubes
  - Random sampling of (year,month) tuples corresponding to individual files
  - Random sampling of space-time cubes in tuples
  - Trivially parallelizable with one data loader per field
- Area preserving sampling for sphere/Earth to compensate for distortion of regular grid

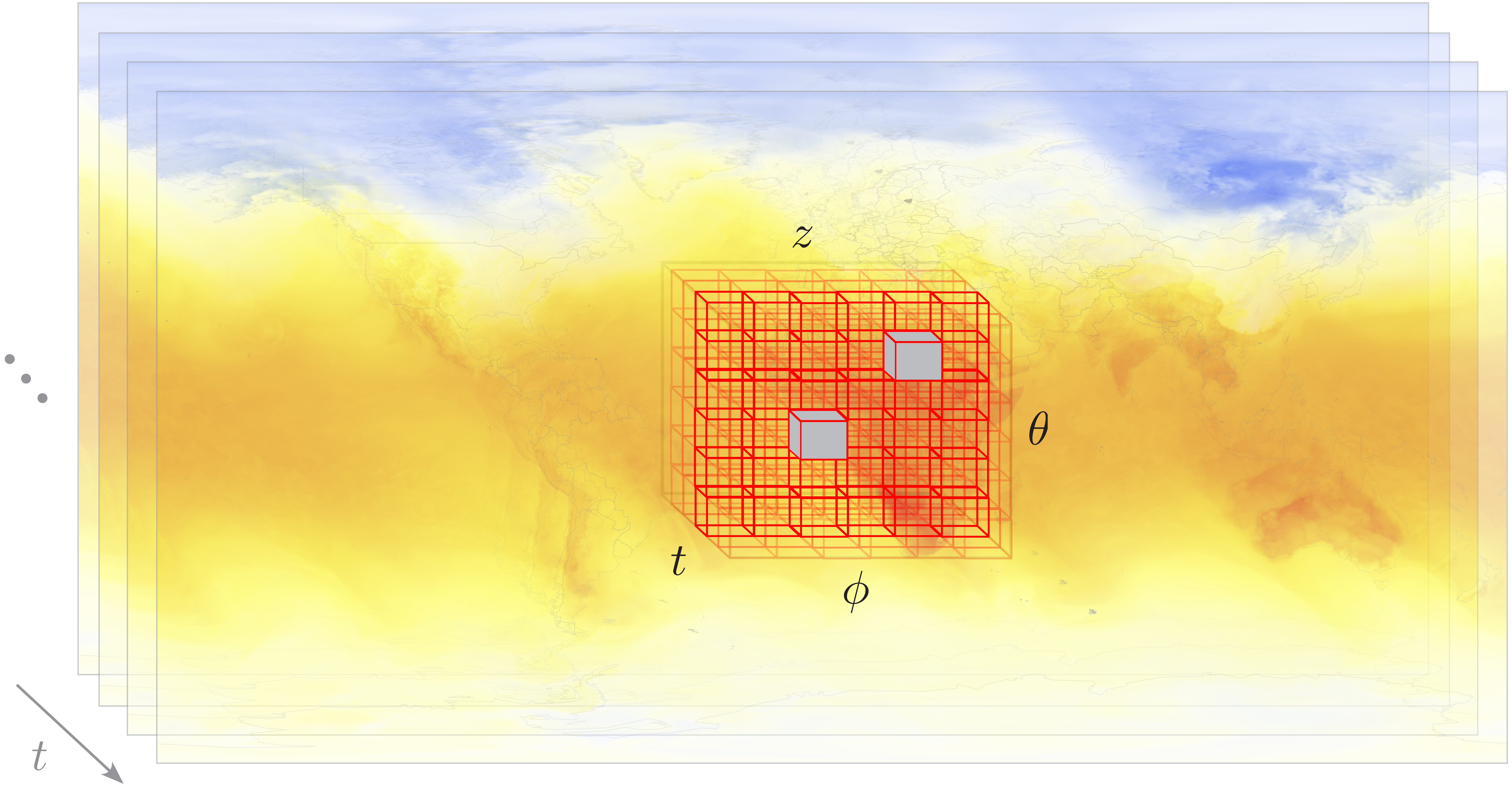
Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language (or token) model

Spatio-temporal BERT

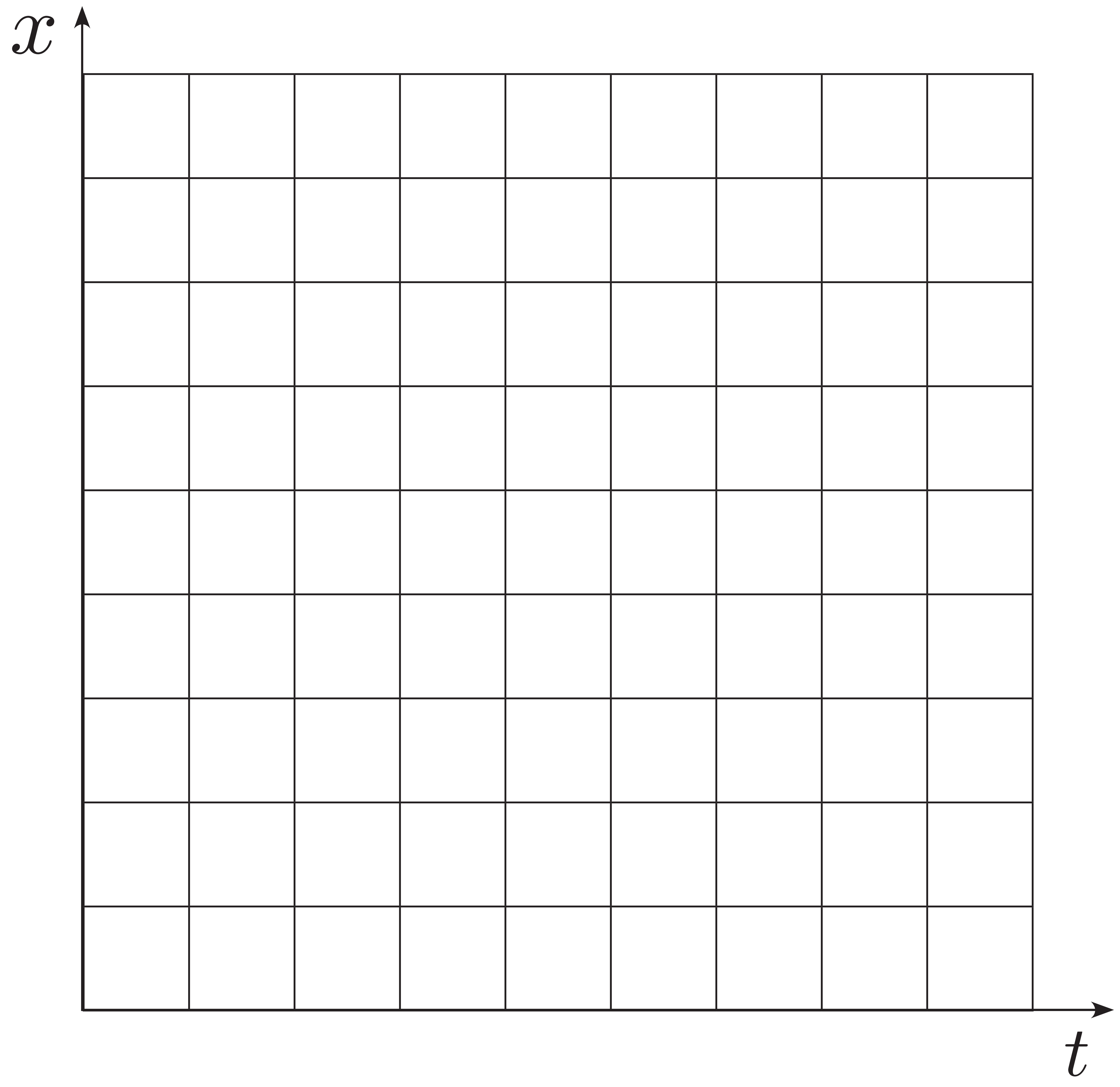


Spatio-temporal BERT



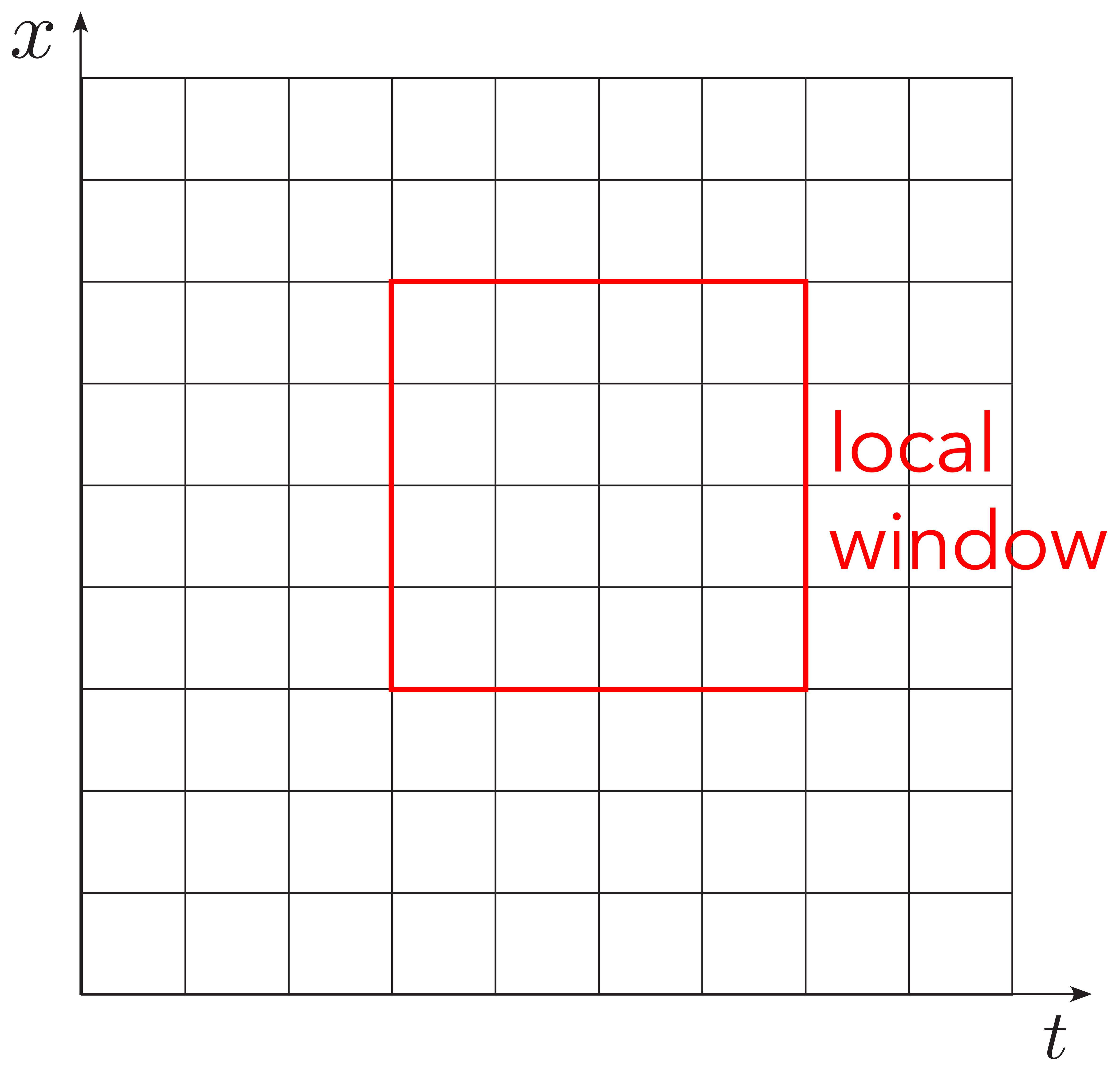
Spatio-temporal BERT

Flatland view



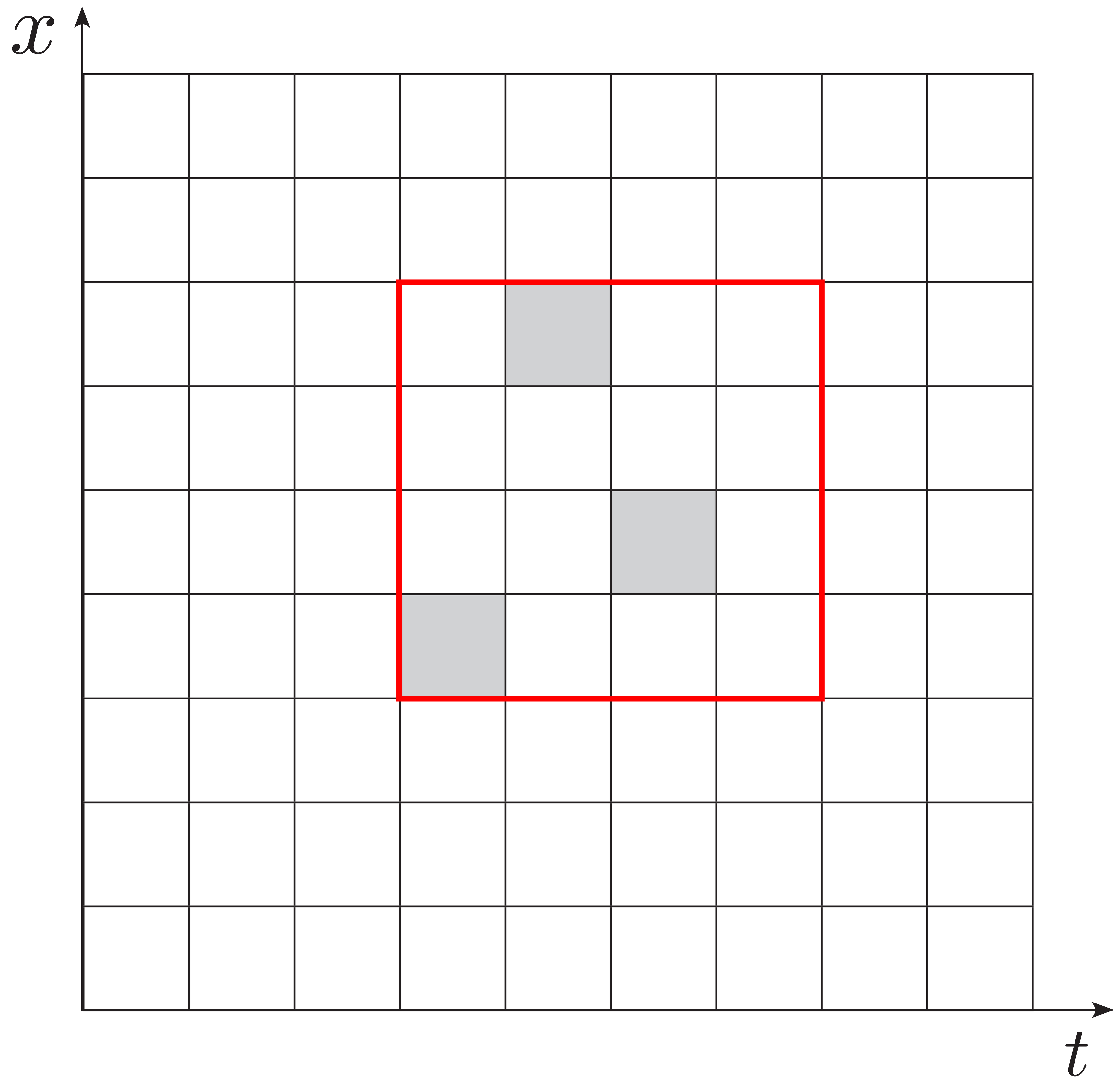
Spatio-temporal BERT

Flatland view



Spatio-temporal BERT

Flatland view of BERT



Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language language model
  - Natural interpretation as forecasting / hindcasting / interpolation

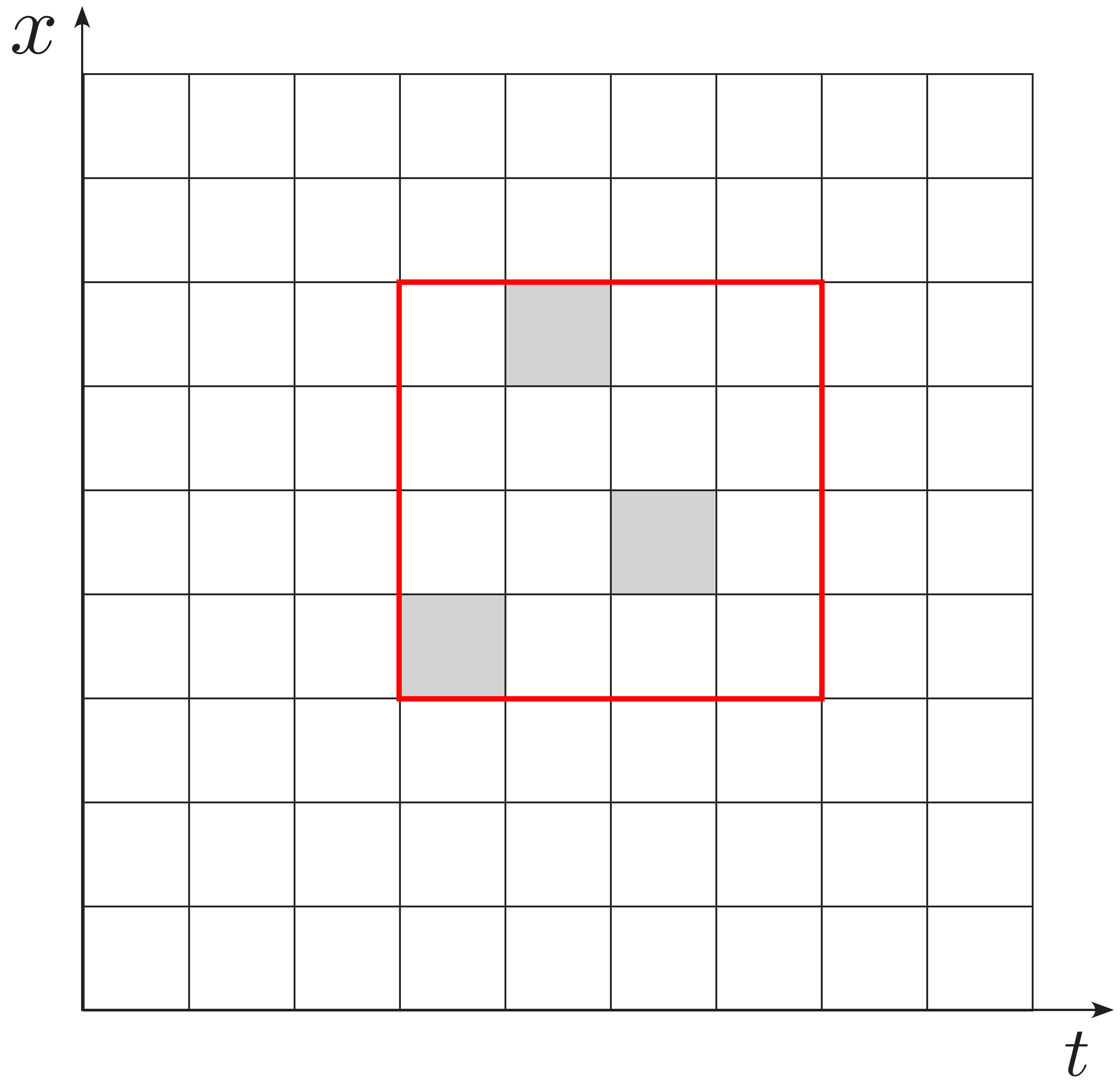
Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language language model
  - Natural interpretation as forecasting / hindcasting / interpolation
  - Bi-directional training poses interesting physical questions

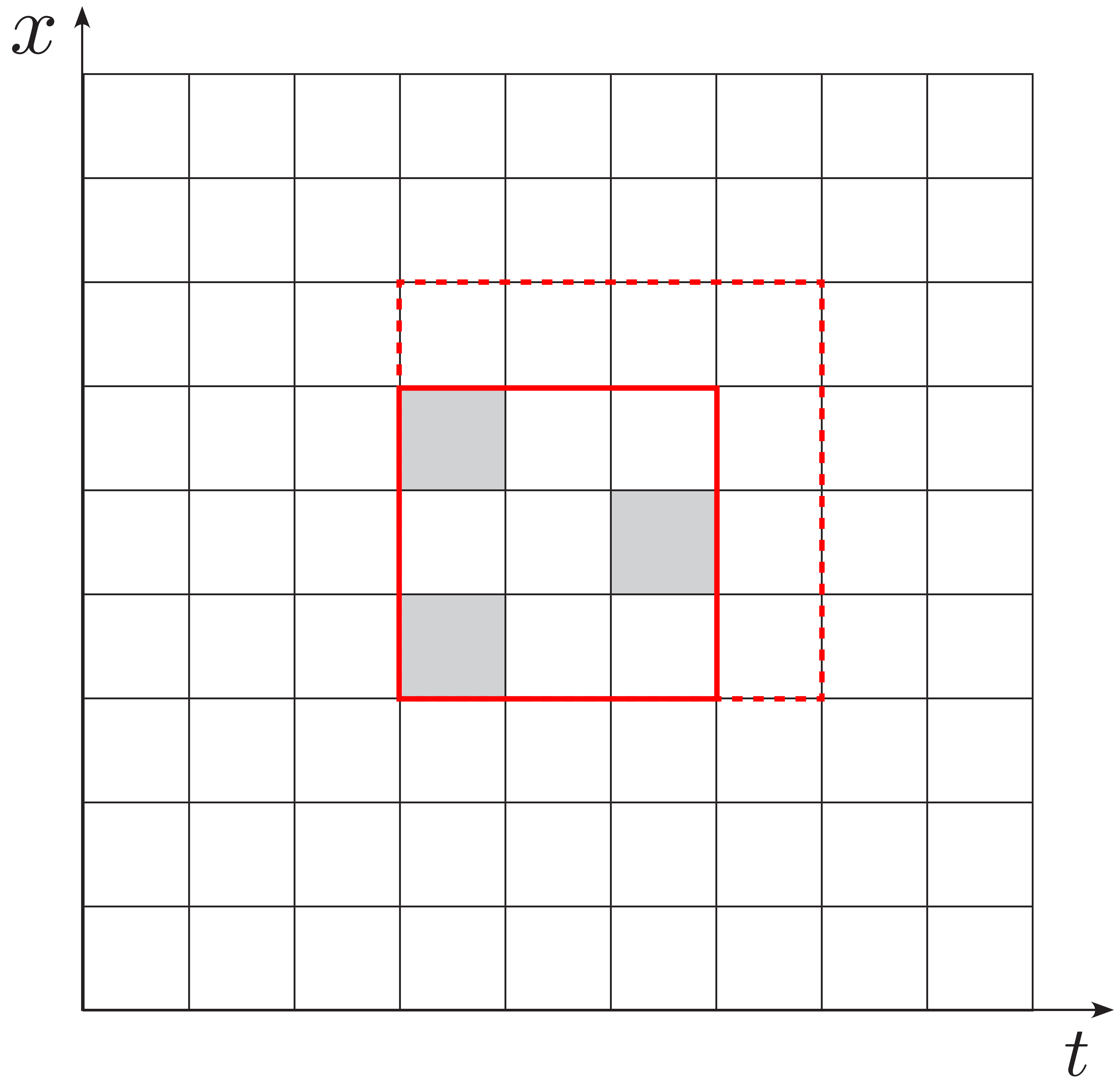
Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language language model
  - Natural interpretation as forecasting / hindcasting / interpolation
  - Bi-directional training poses interesting physical questions
  - Performed on randomly cropped subset to be invariant to specific token numbers

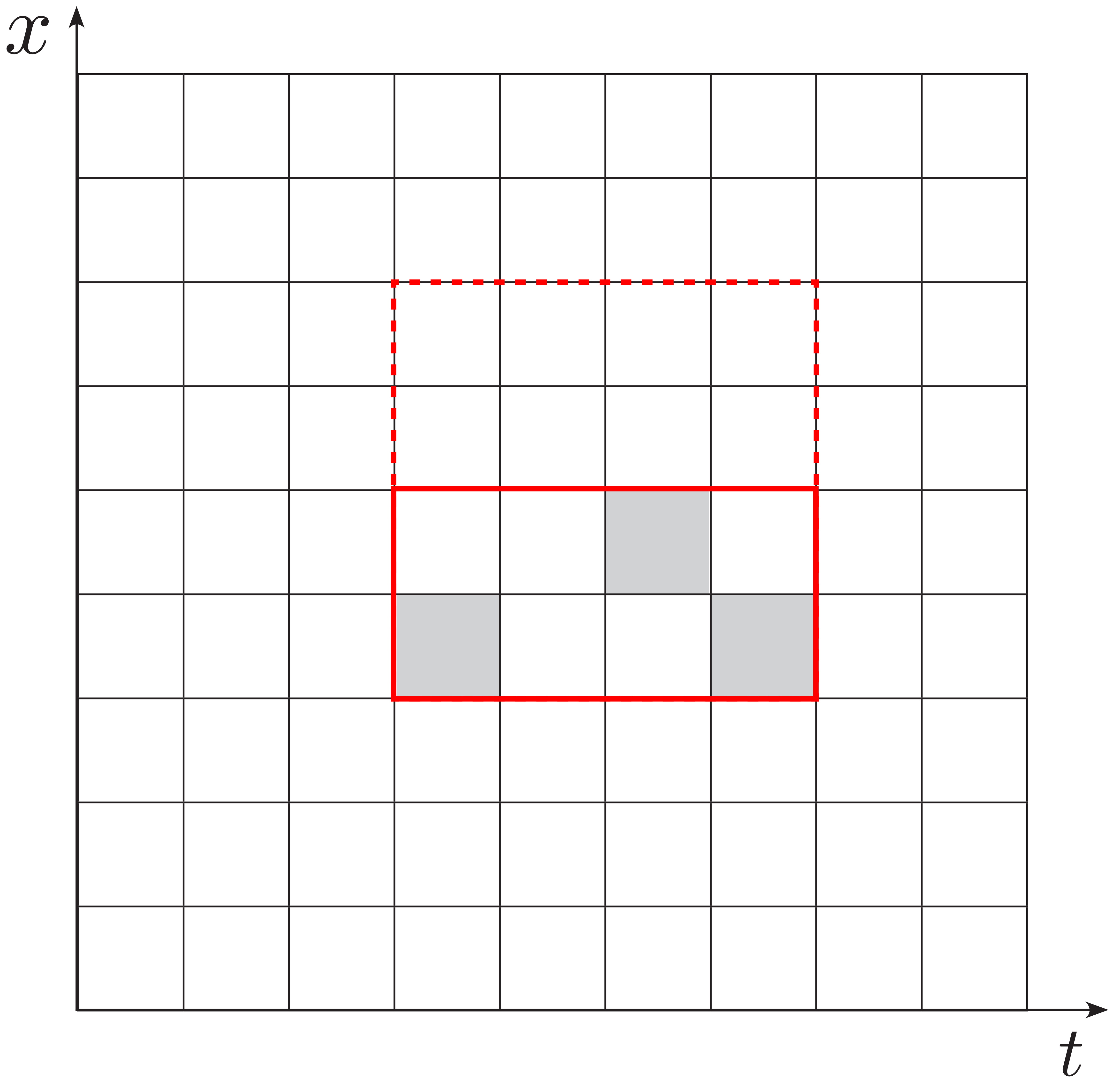
Spatio-temporal BERT



Spatio-temporal BERT



Spatio-temporal BERT



Statistical loss

- ML: Training on MSE loss is problematic in terms of training dynamics

Statistical loss

- ML: Training on MSE loss is problematic in terms of training dynamics
- Training on just the mean is sub-optimal to learn a probabilitic/statistical representation of the dynamics and the system

Statistical loss

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *ICLR*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendix for "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

The advantage of this procedure is that the Transformer encoder does not know which words it will be asked to predict or which have been replaced by random words, so it is forced to keep a distributional contextual representation of *every* input token. Additionally, because random replacement only occurs for 1.5% of all tokens (i.e., 10% of 15%), this does not seem to harm the model's language understanding capability. In Section C.2, we evaluate the impact this procedure.

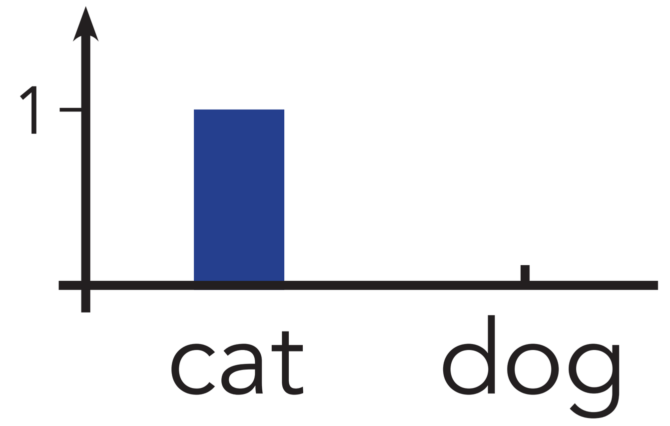
J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Statistical loss

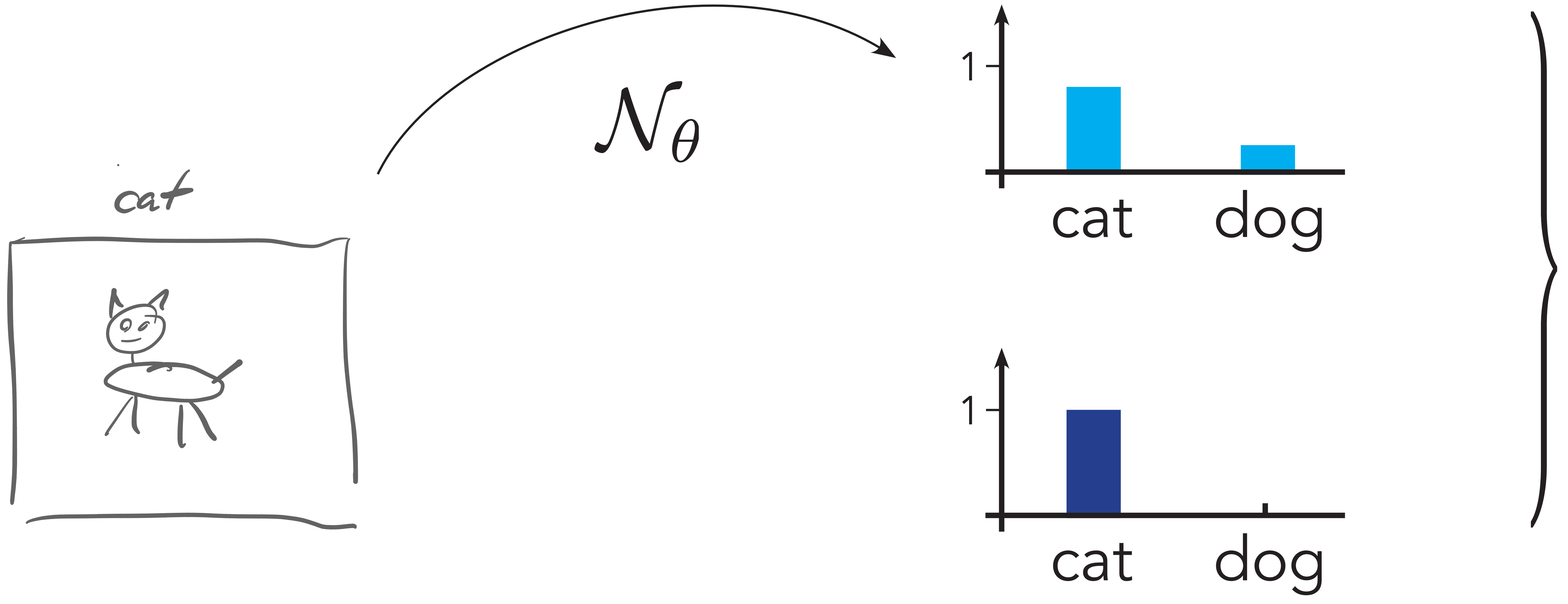
How to obtain improved training dynamics and ensure probabilistic/statistical representation in network?

Statistical loss

Statistical loss

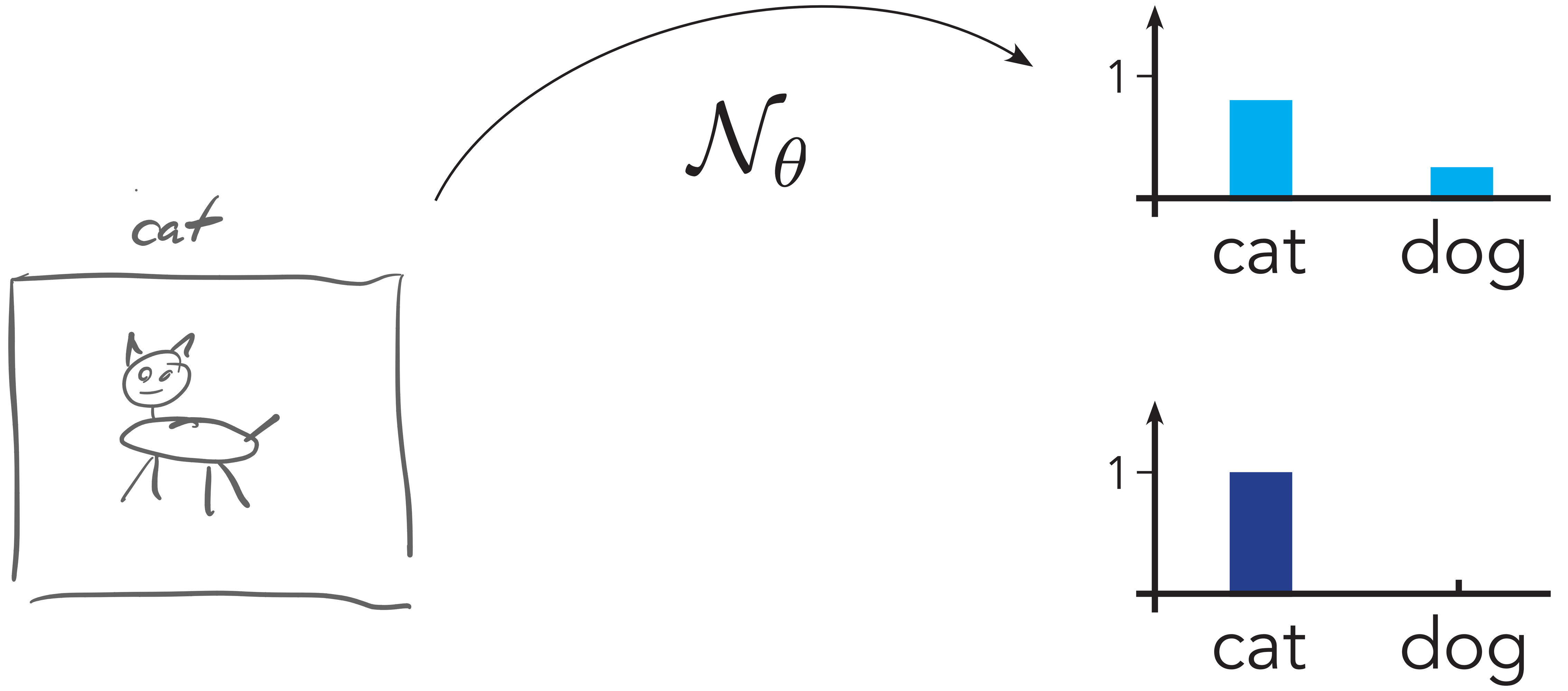


Statistical loss

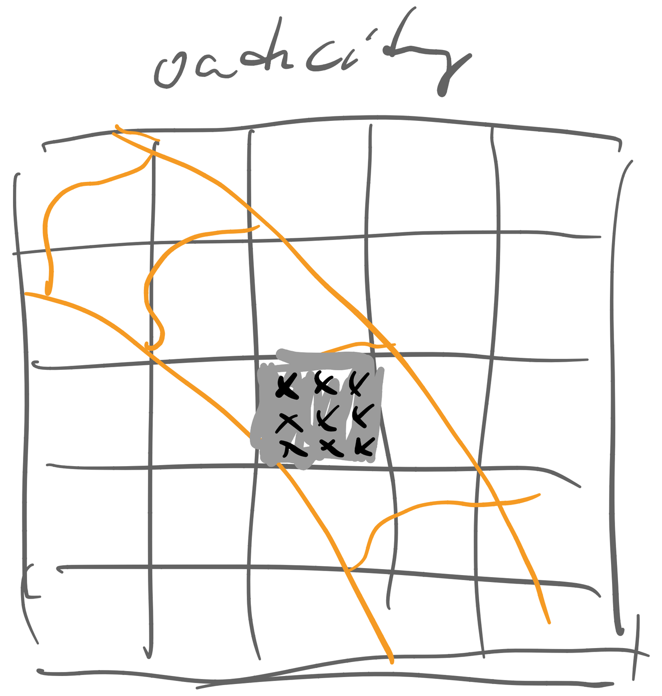


cross
entropy
loss

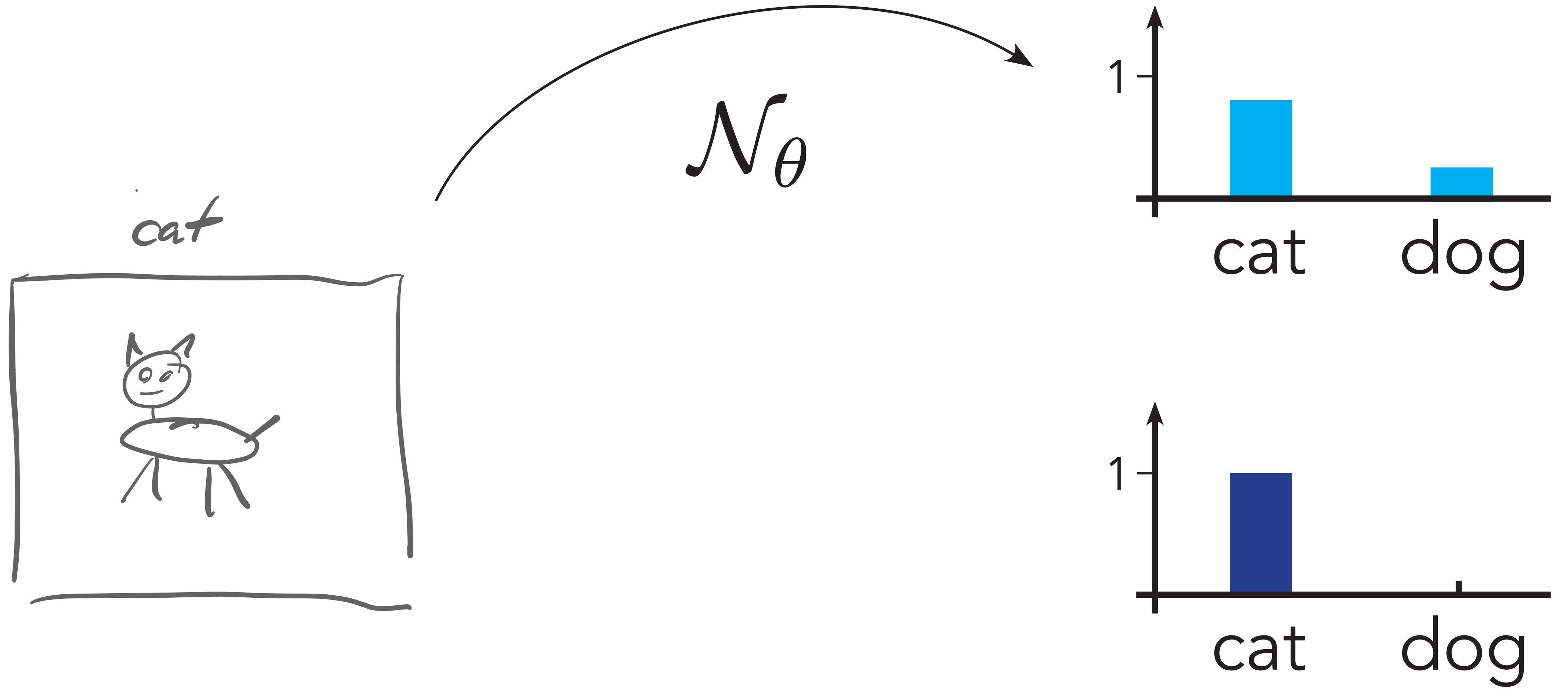
Statistical loss



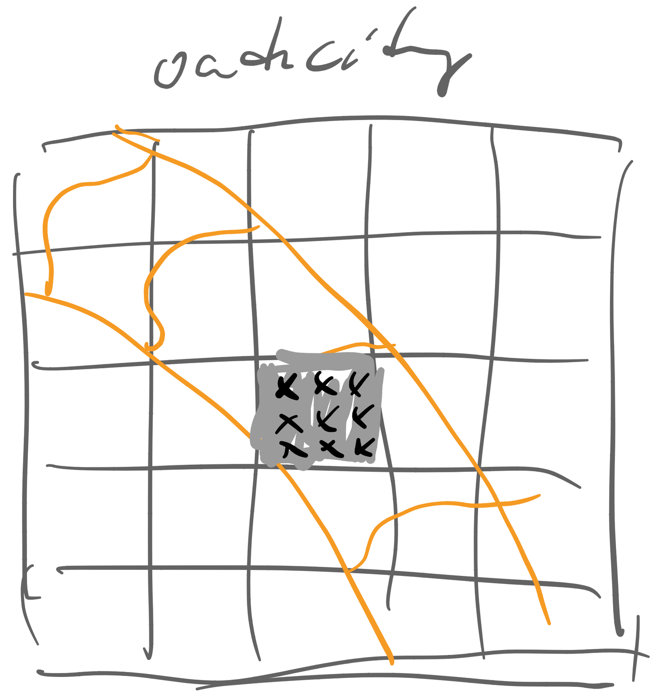
cross entropy loss



Statistical loss

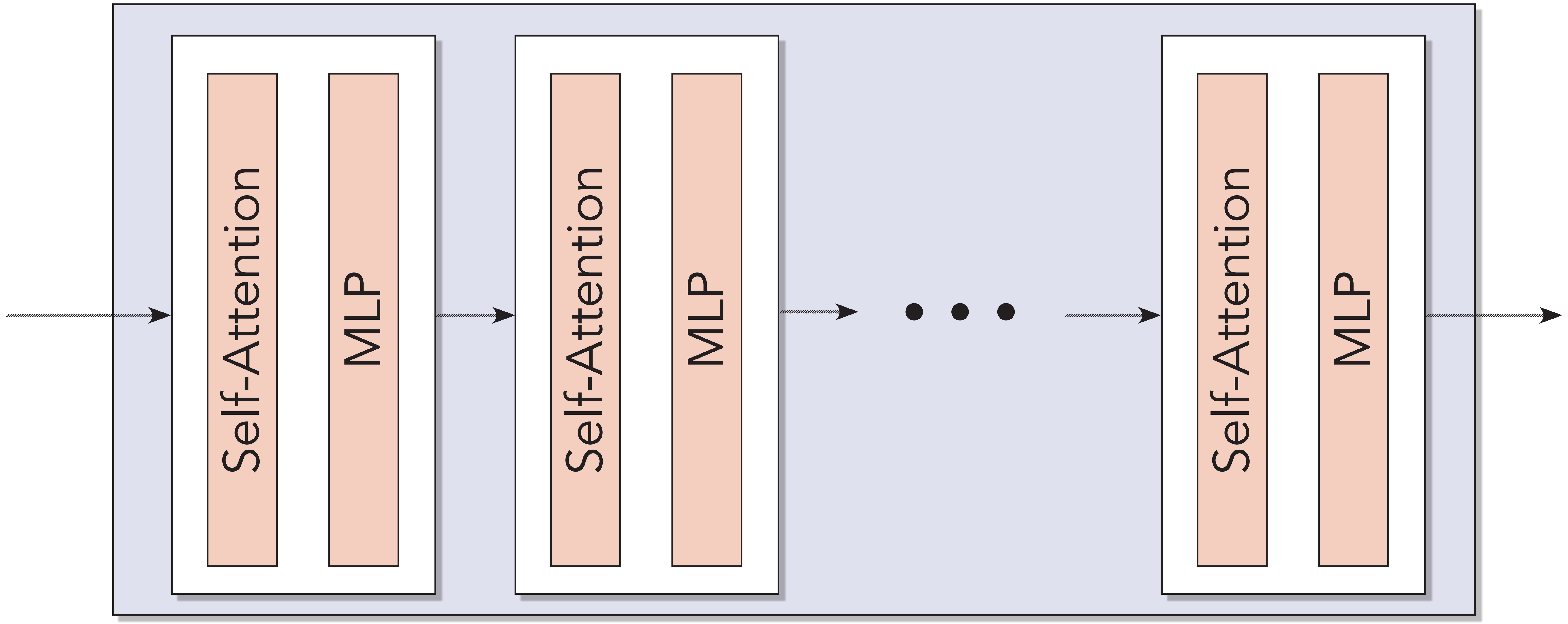


cross
entropy
loss

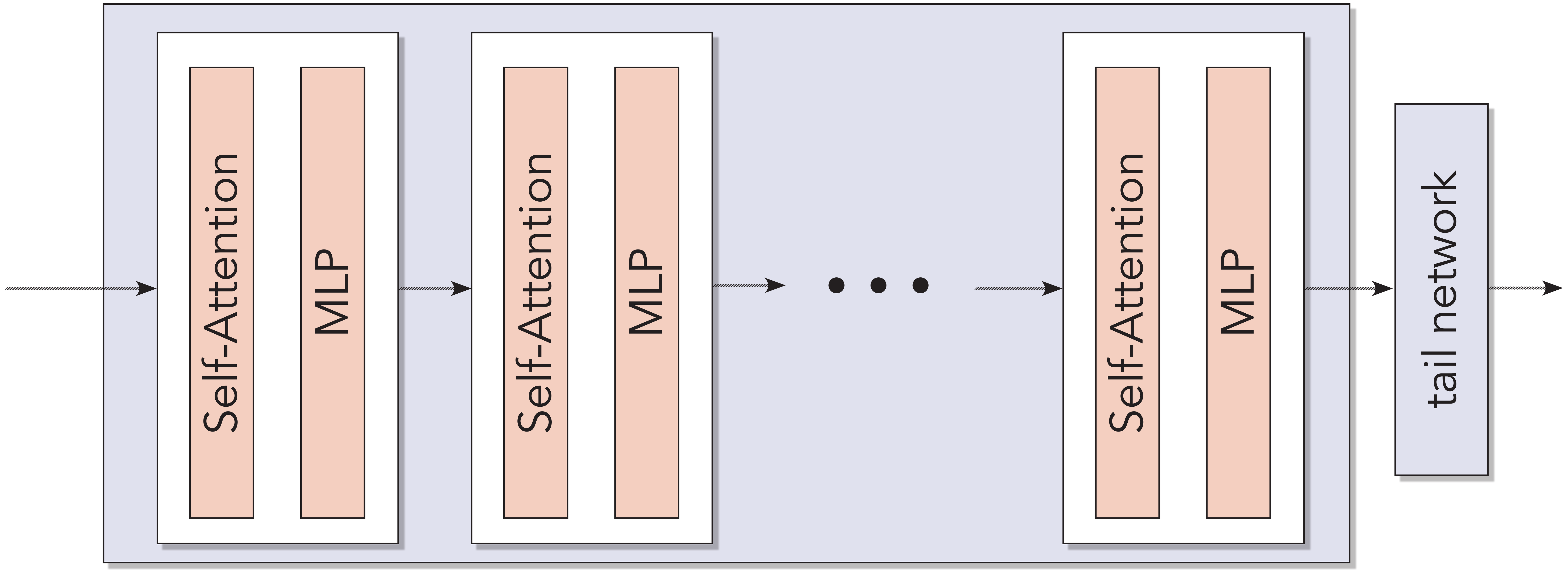


?

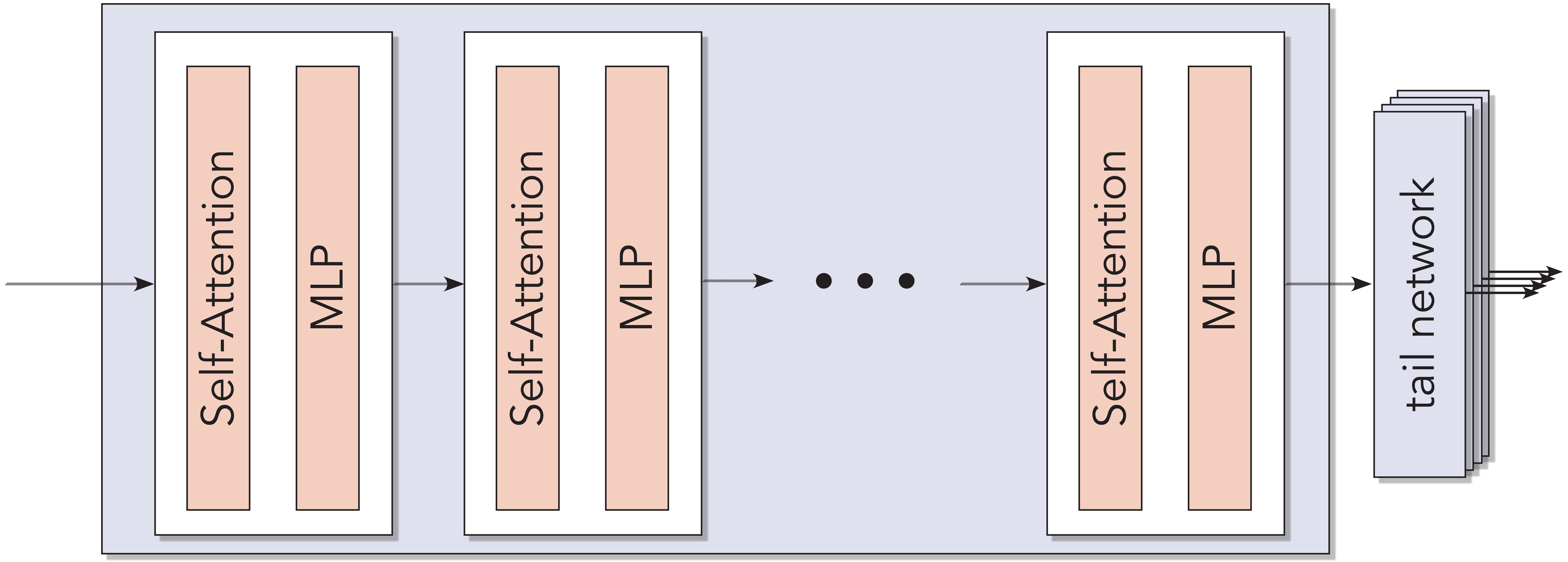
Statistical loss



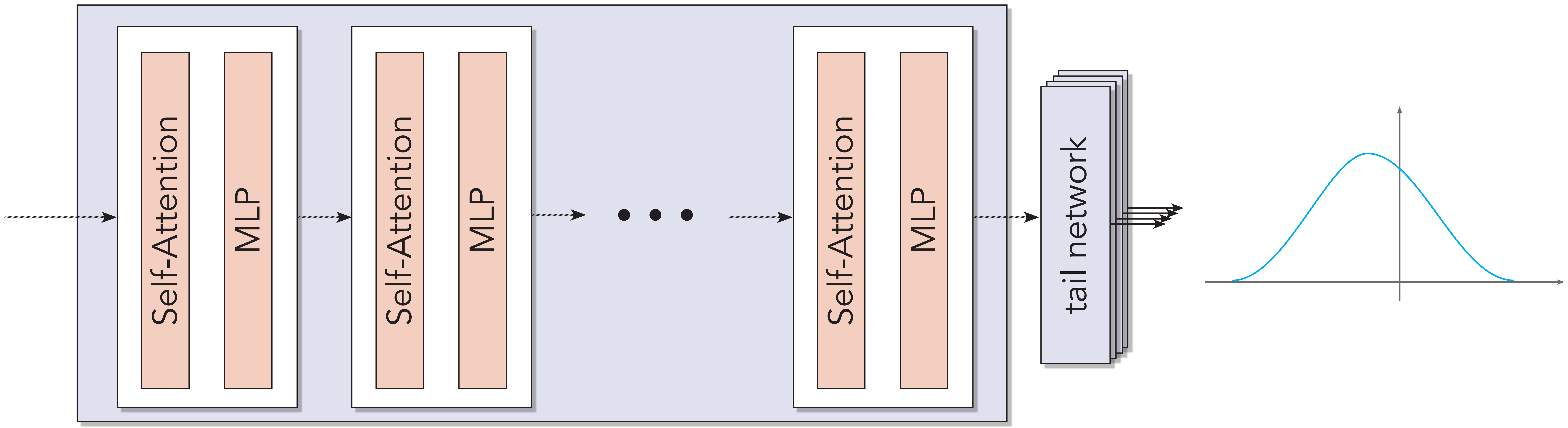
Statistical loss



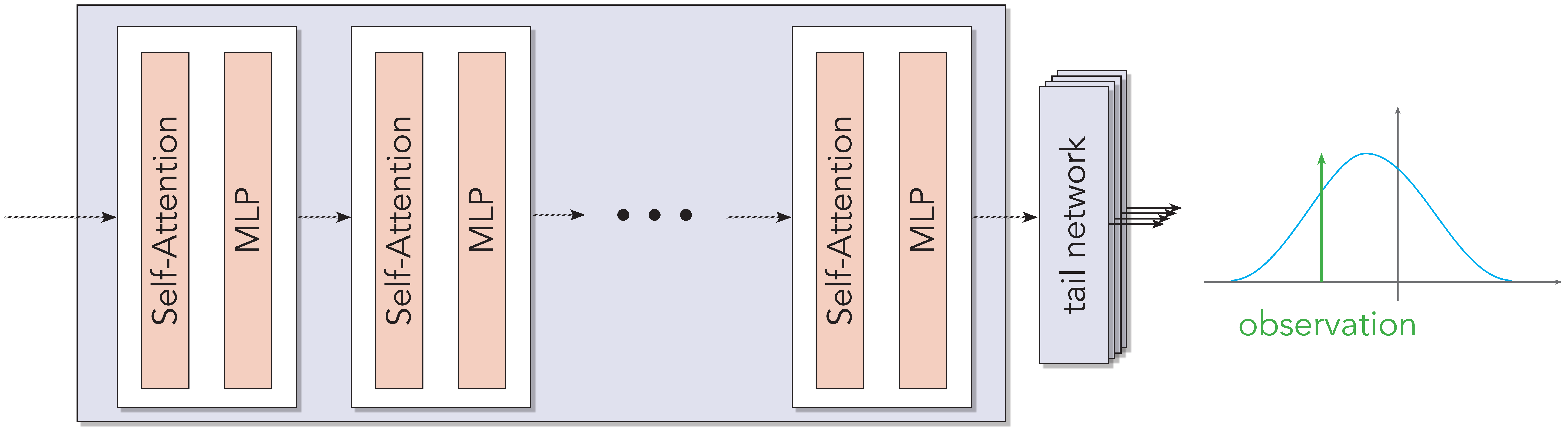
Statistical loss



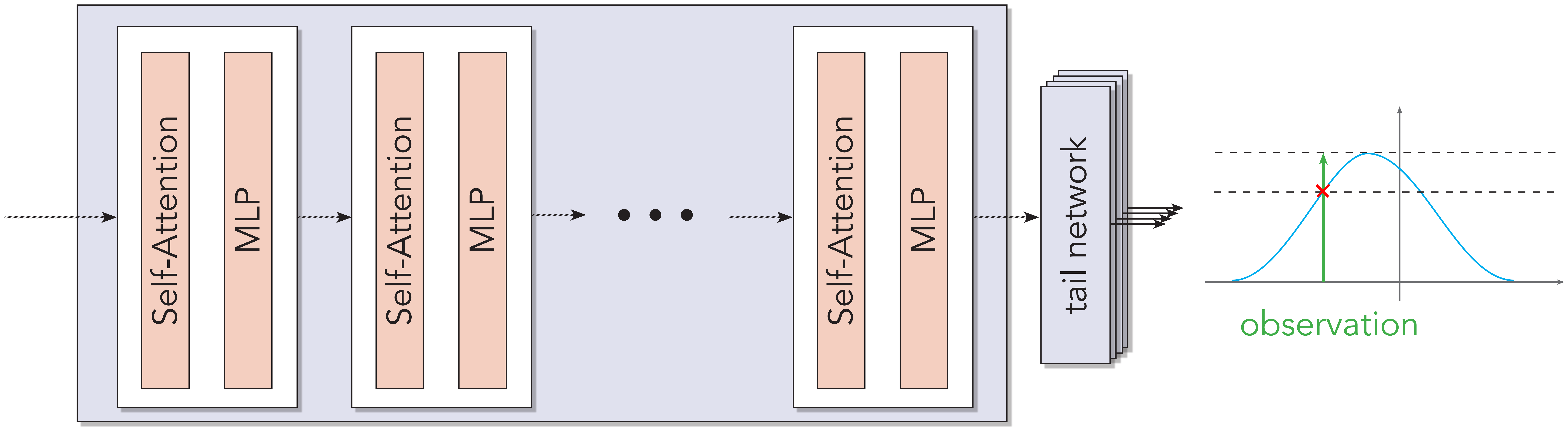
Statistical loss



Statistical loss

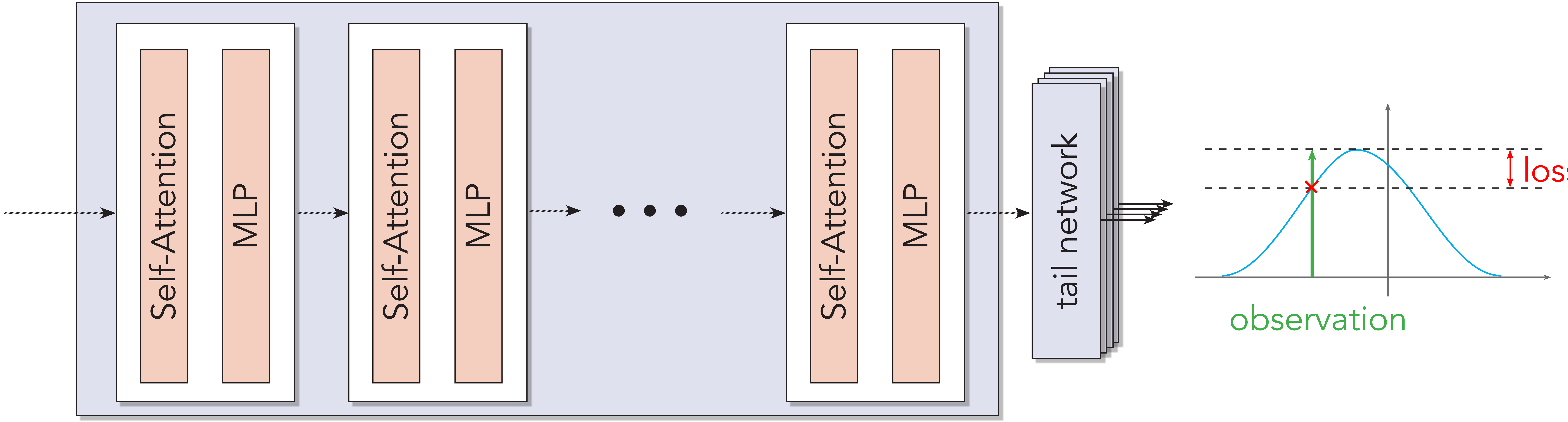


Statistical loss



Statistical loss

Statistical loss

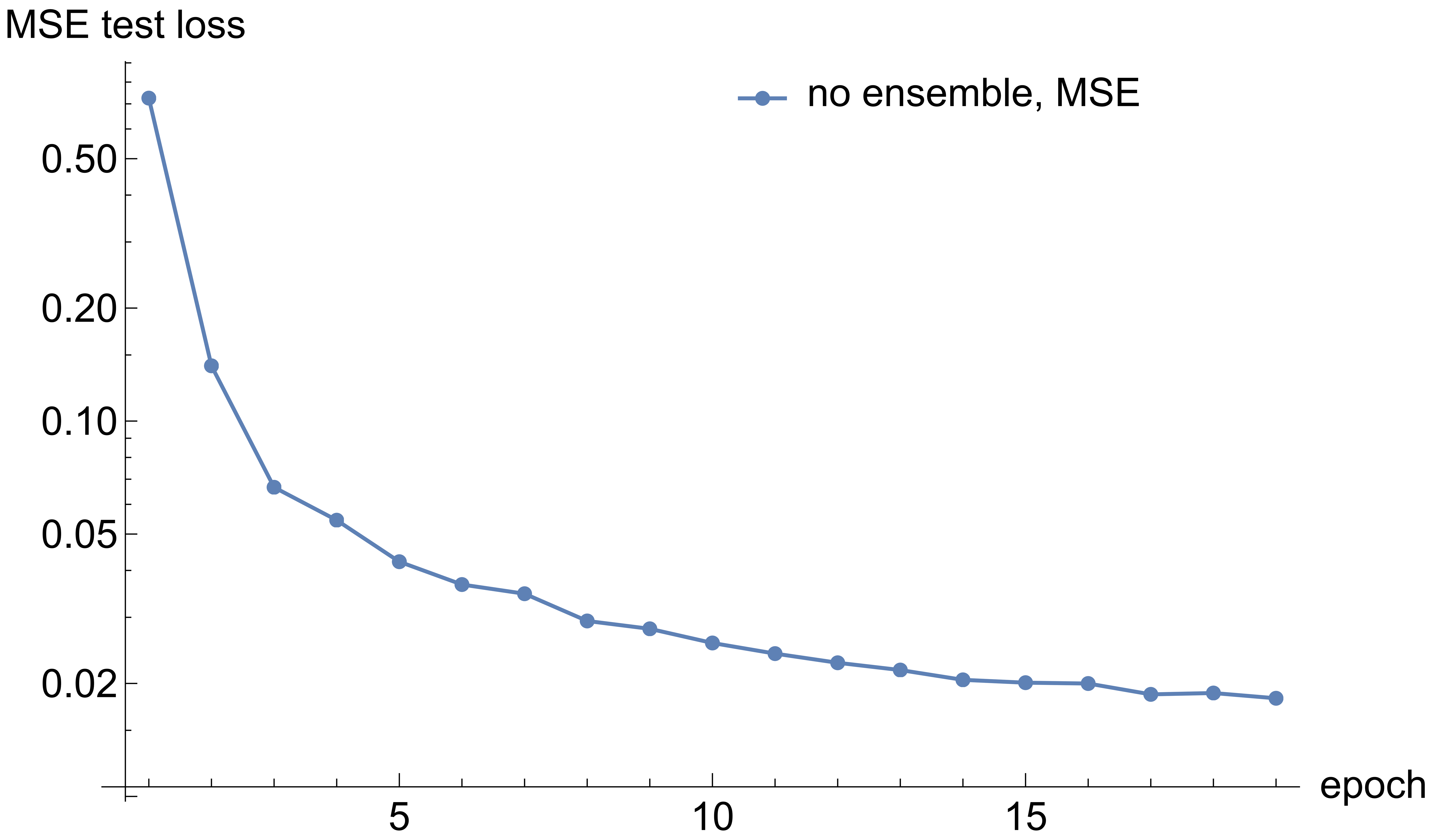


end-to-end training encourages statistical representation

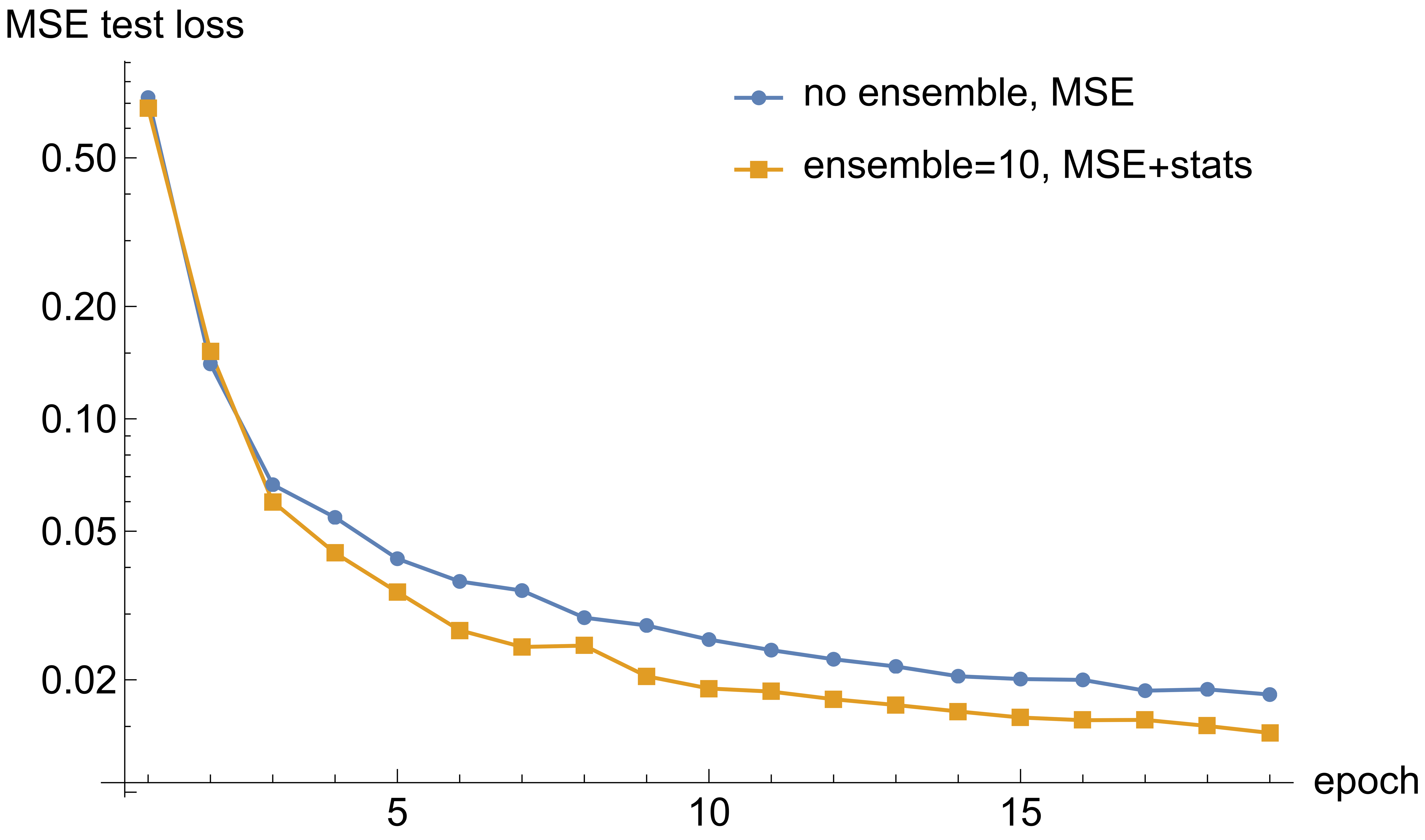
Statistical loss: experiments

- BERT with conditional masking
- 975 hPa (high frequency) vorticity
- 40 years of training data

Statistical loss



Statistical loss



Statistical loss

- Statistical loss:

$$\mathcal{L}_{\mathrm{stats}} = \left| 1 - \int_{\mathrm{R}} \delta_y(x)\, G_{\tilde{\mu},\tilde{\sigma}}(x)\, \mathrm{d}x \right|^2$$

Statistical loss

- Statistical loss:

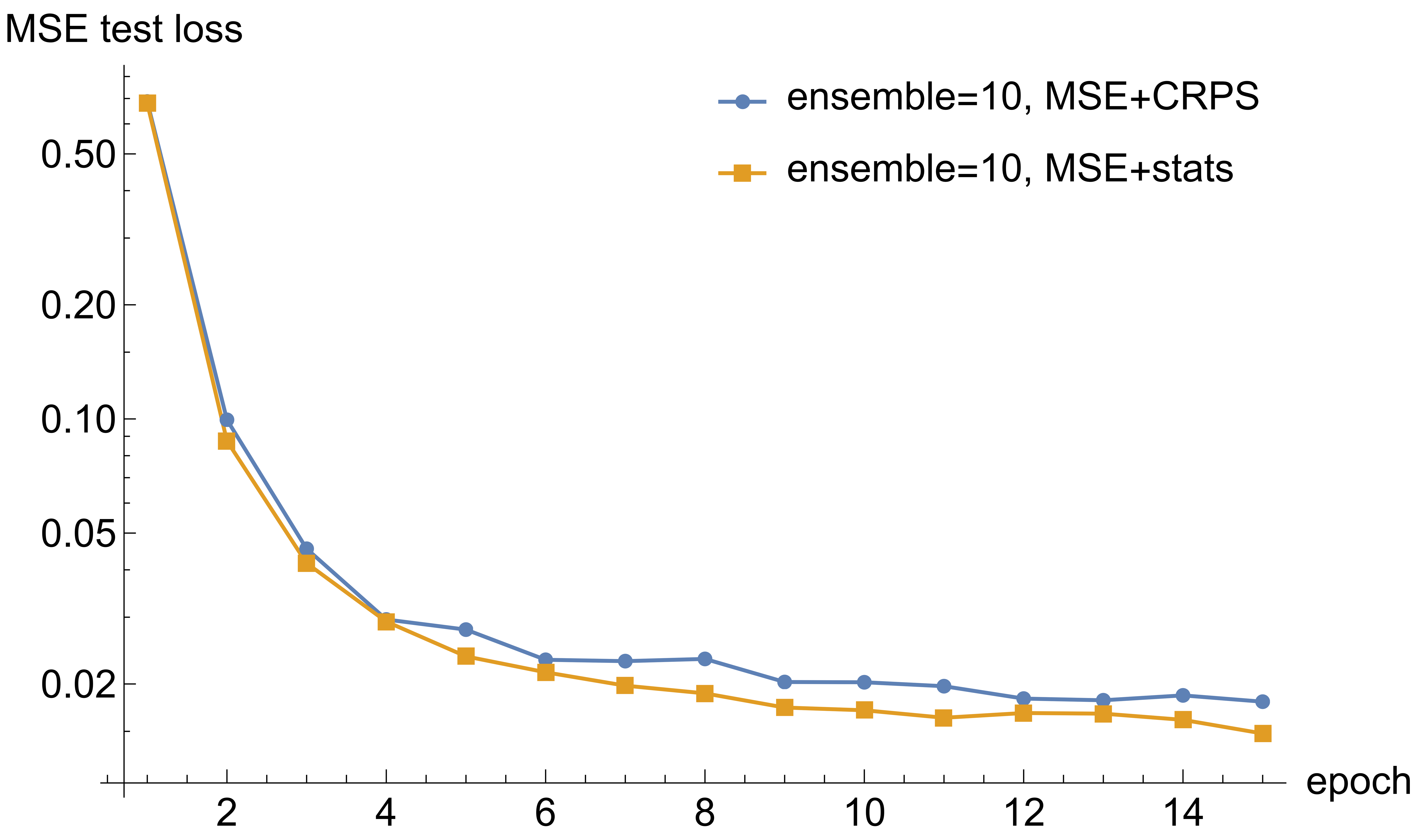
$$\mathcal{L}_{\mathrm{stats}} = \left| 1 - \int_{\mathrm{R}} \delta_y(x)\, G_{\tilde{\mu},\tilde{\sigma}}(x)\, \mathrm{d}x \right|^2$$

- CRPS:[1]

$$\mathcal{L}_{\mathrm{CRPS}} = \int_{\mathrm{R}} \left| H_y(x)\, \mathrm{erf}_{\tilde{\mu},\tilde{\sigma}}(x) \right|^2 \mathrm{d}x$$

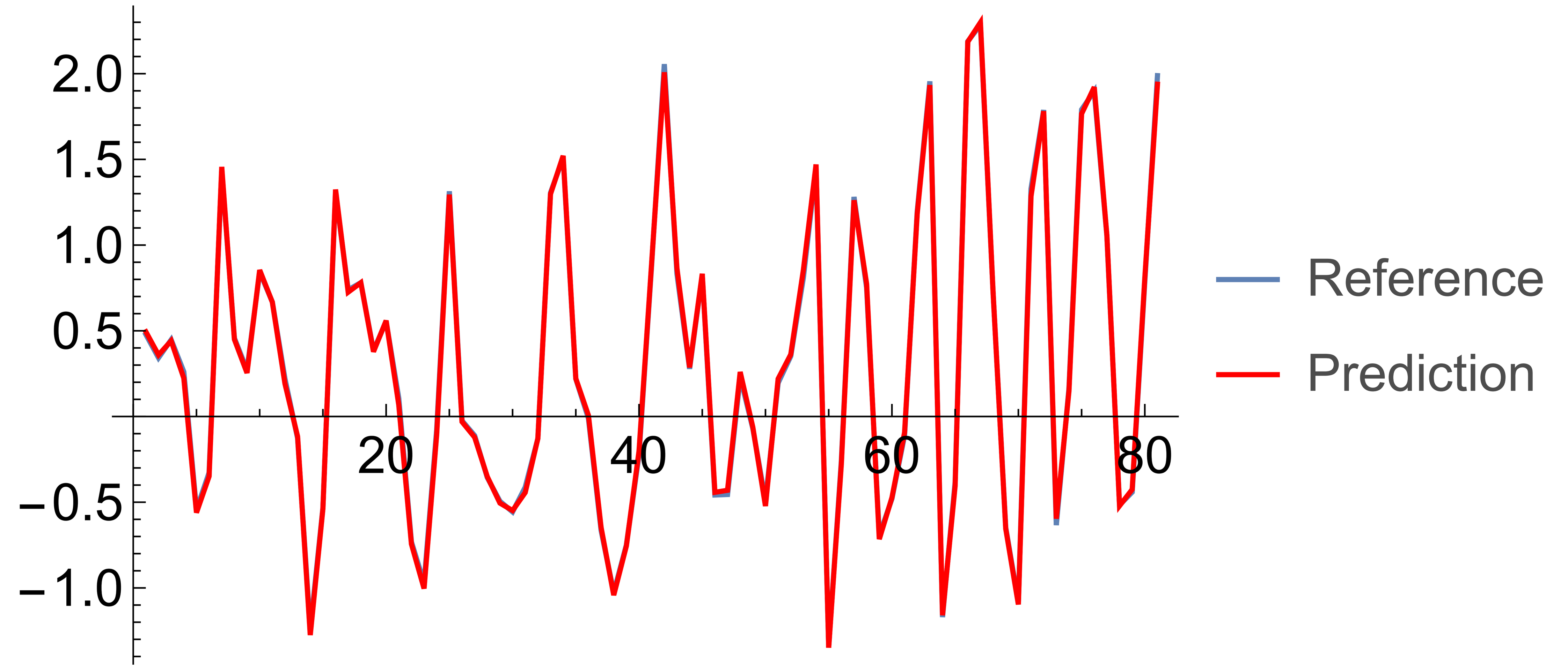
[1] S. Rasp and S. Lerch. Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11):3885 – 3900, 2018.

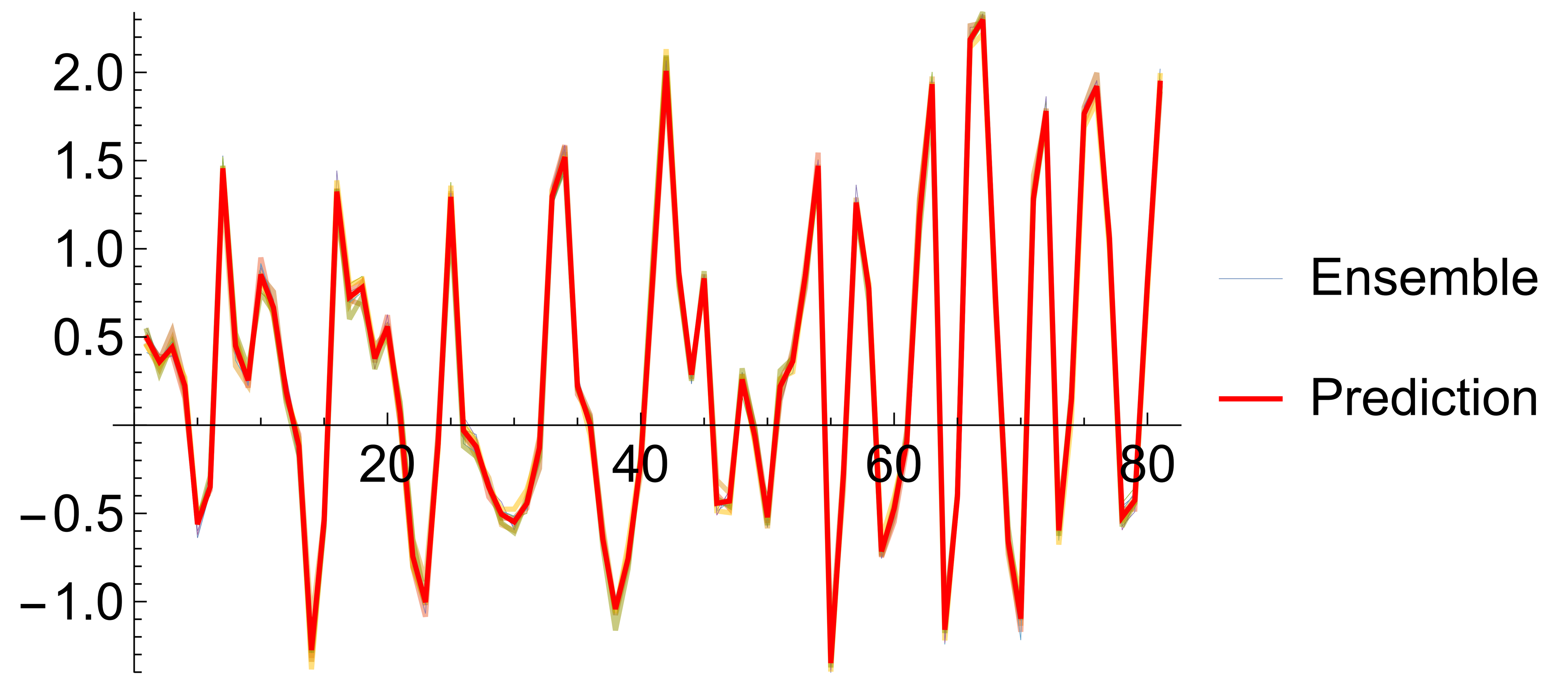
Statistical loss



Statistical loss

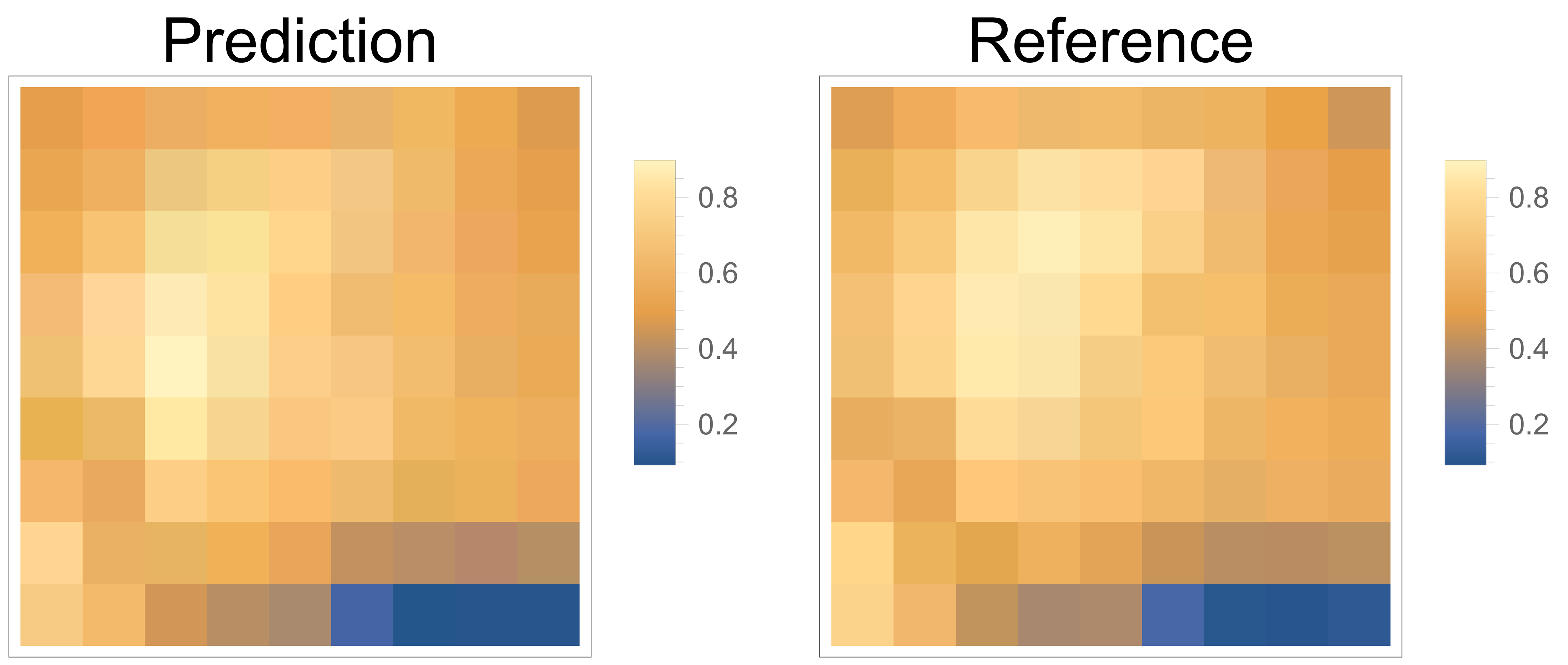
- Predictions:

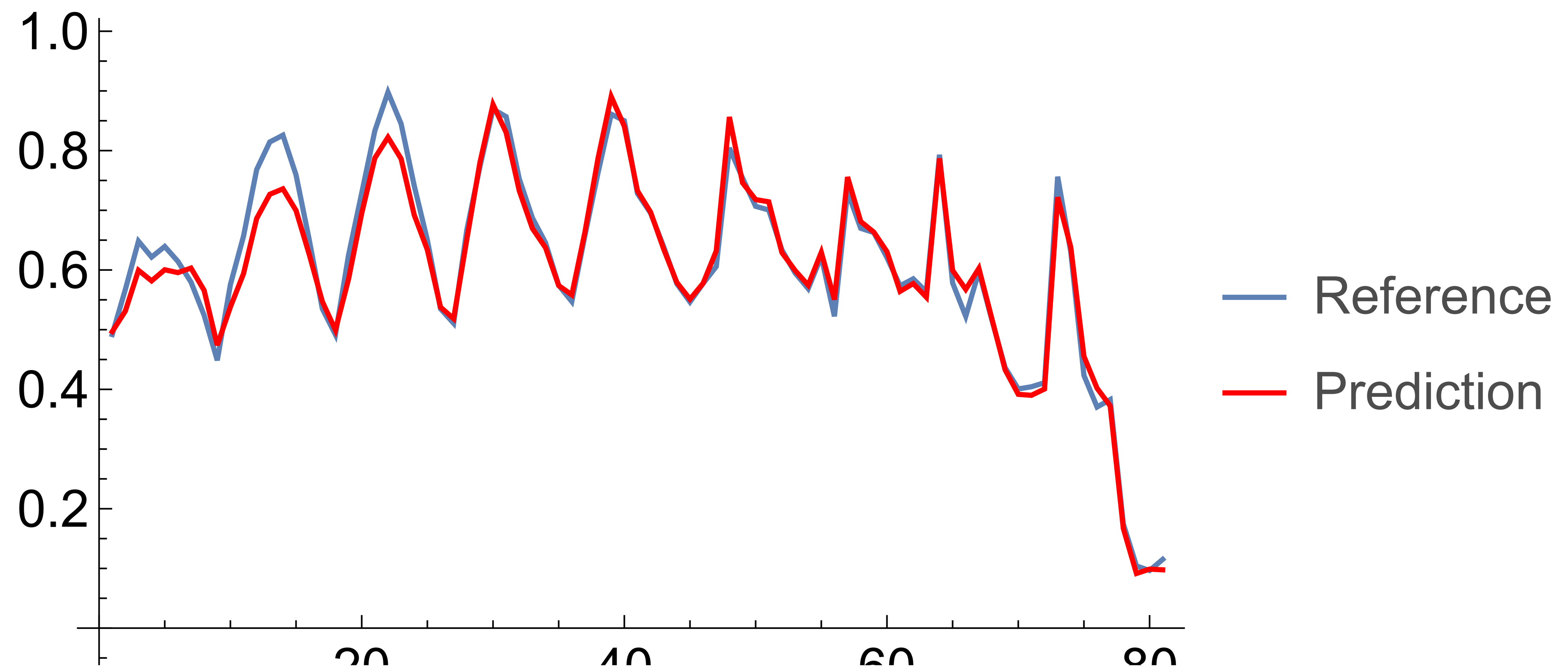


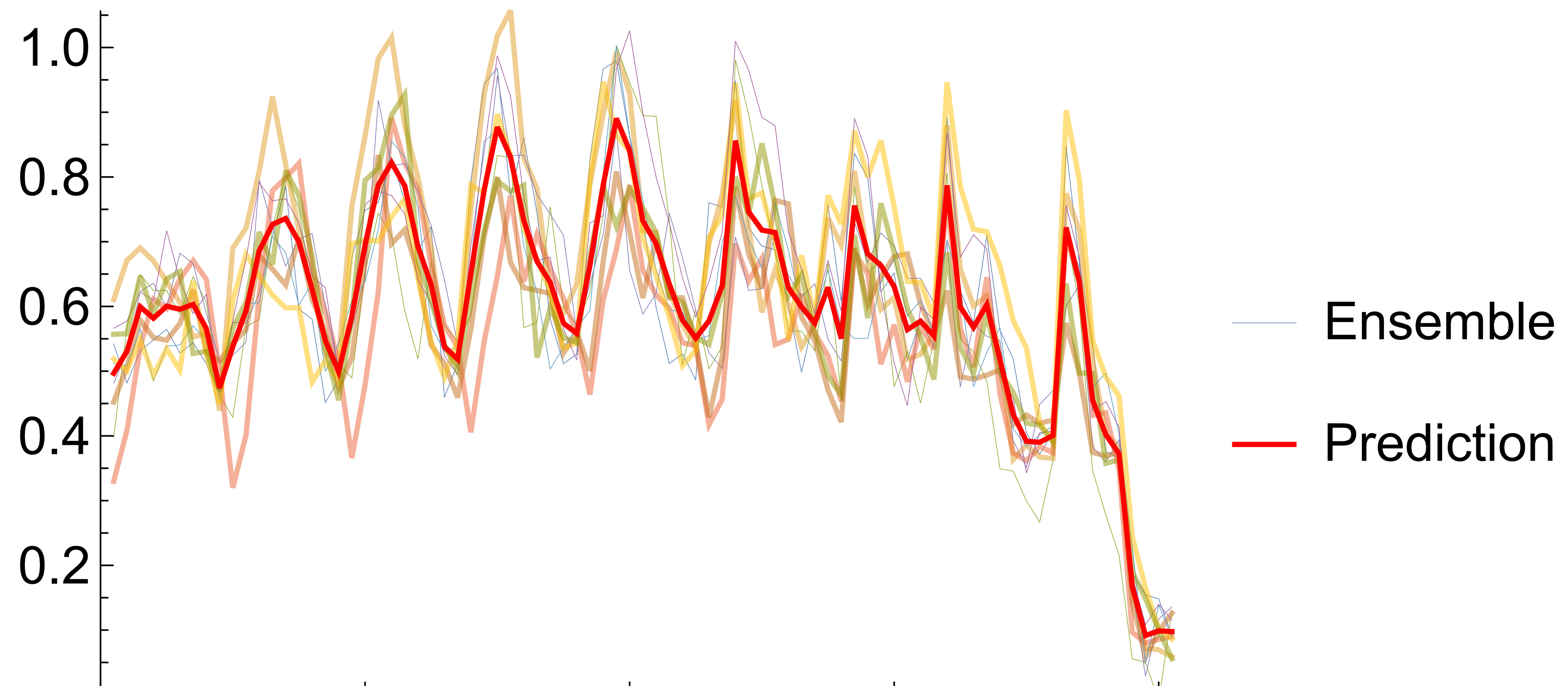


Statistical loss

- Predictions:

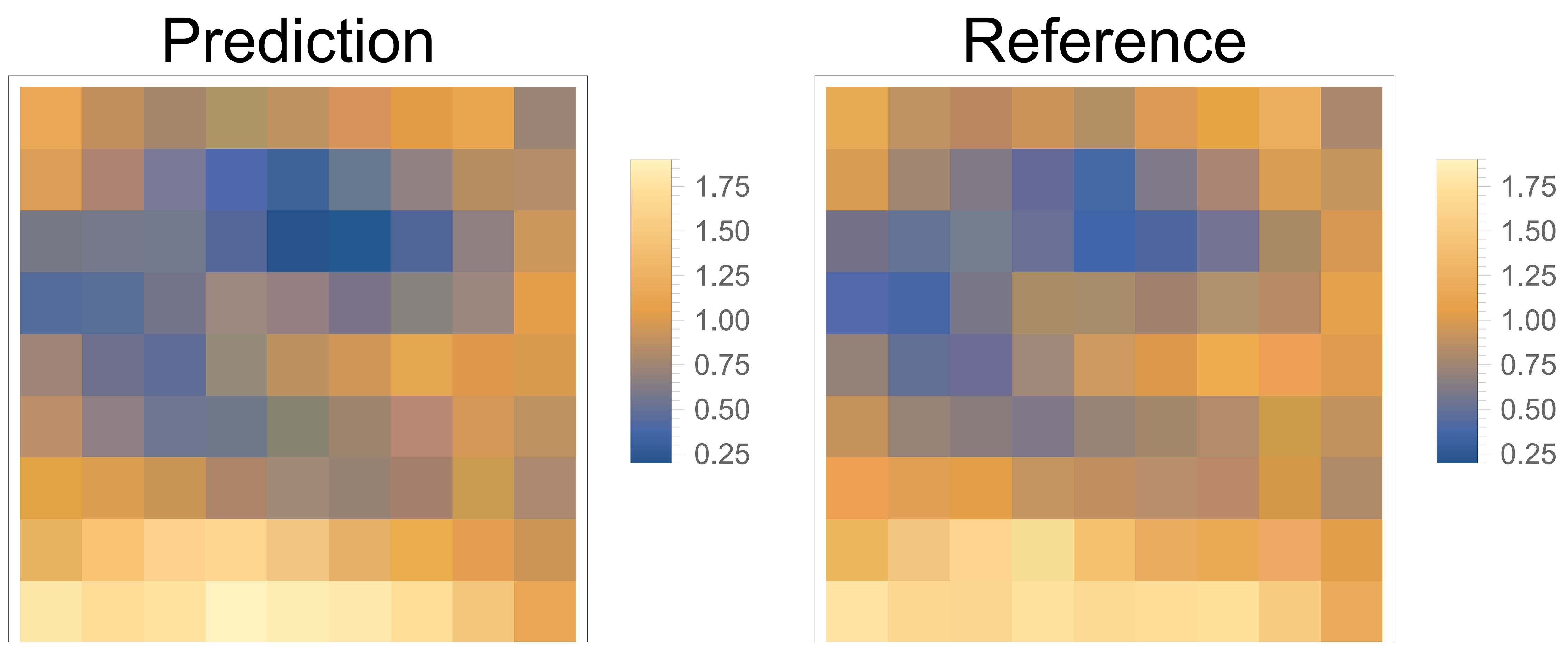


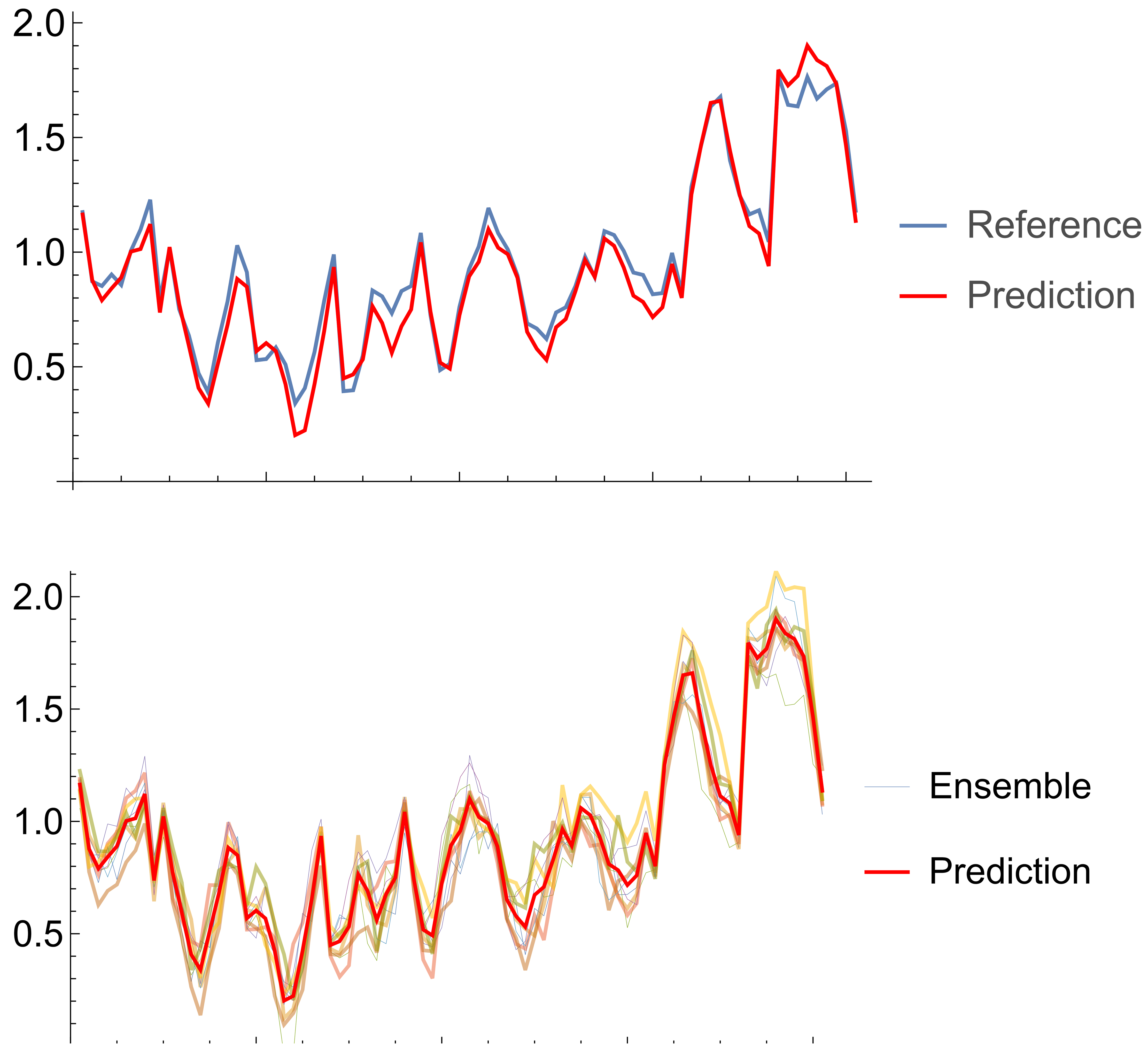




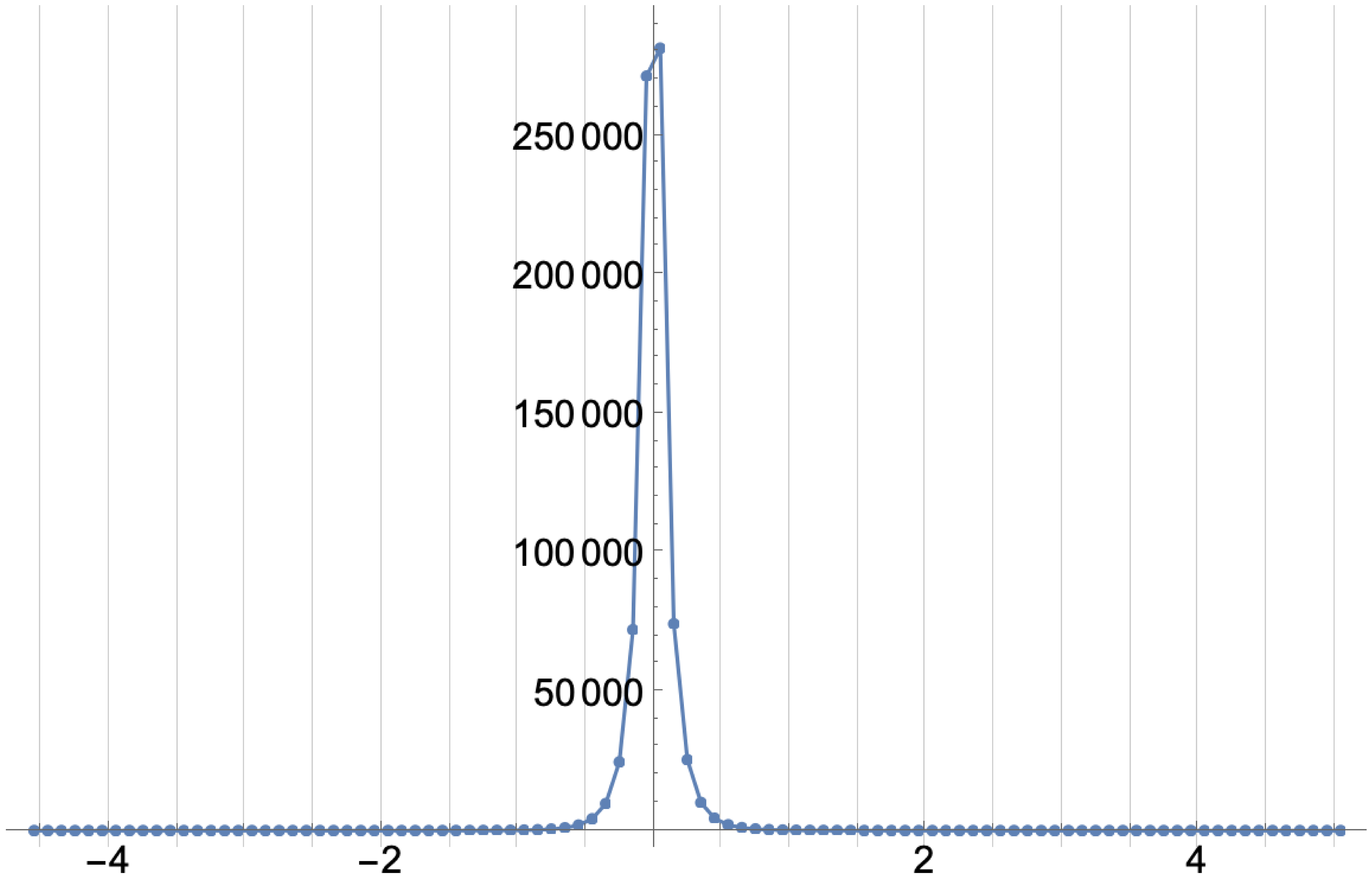
Statistical loss

- Predictions:



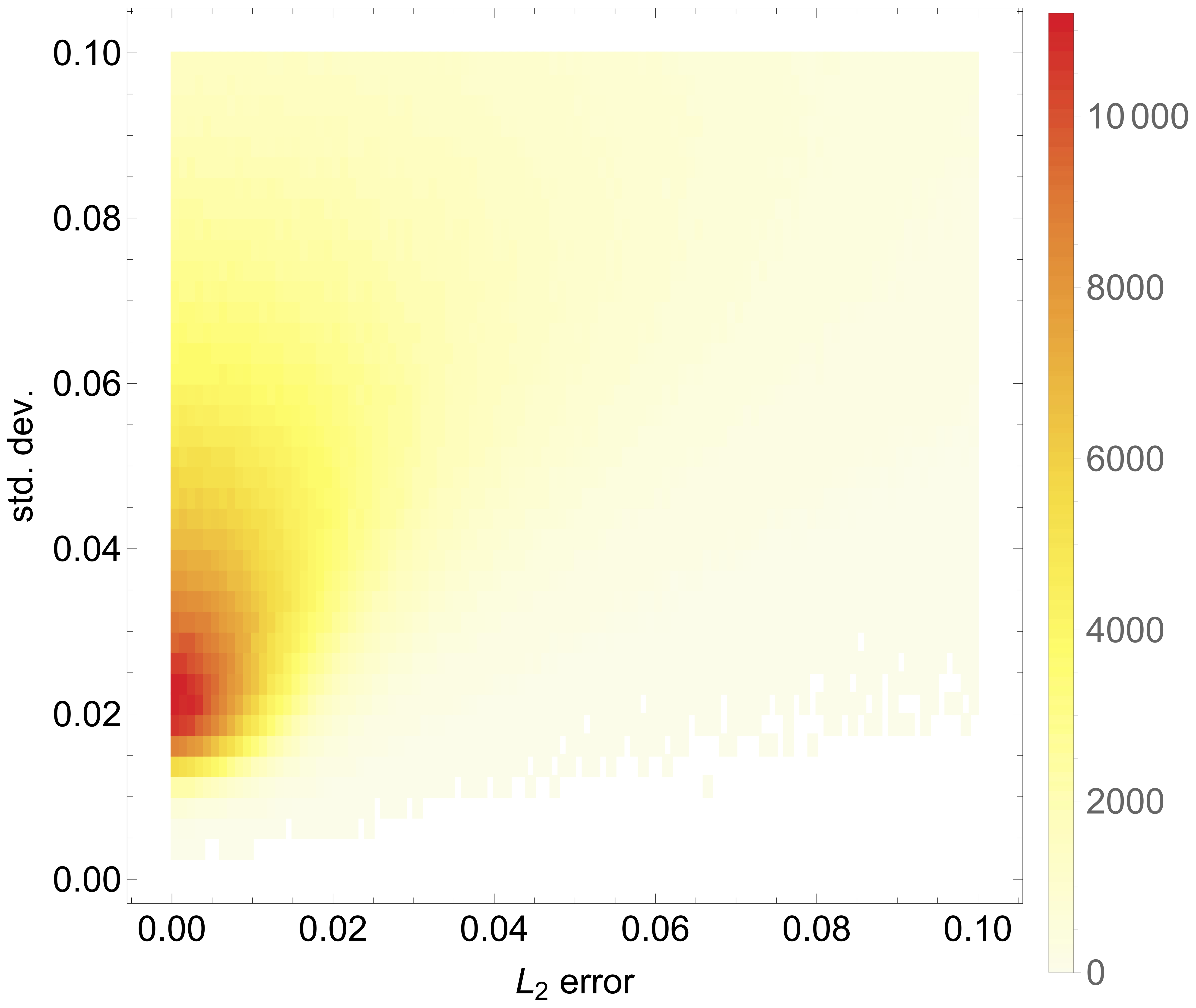


Statistical loss



Histogram of ensemble errors

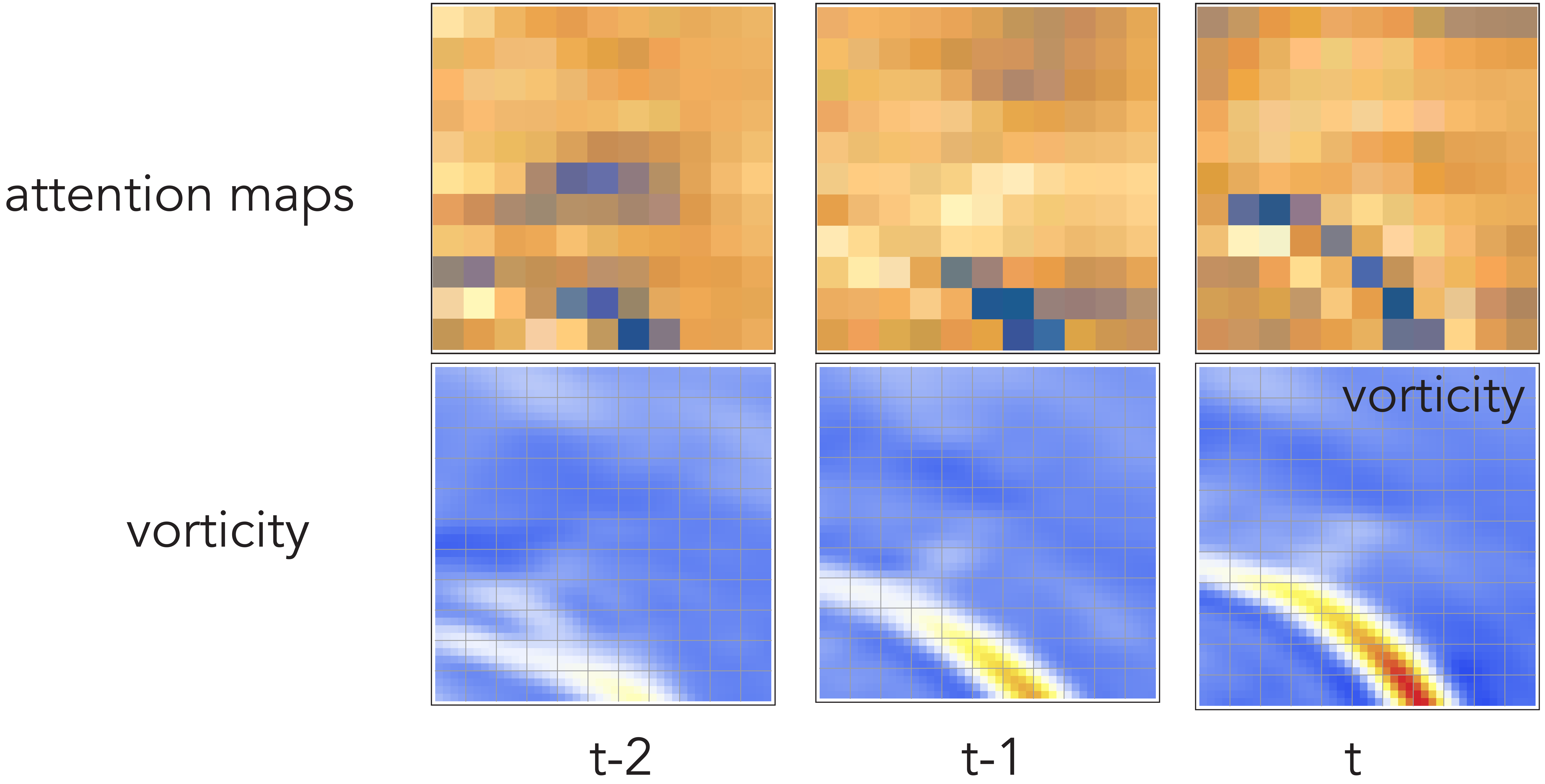
Statistical loss



2D Histogram of L_2 error vs. std. dev.

Statistical loss

- Attention maps:



Zero shot evaluation

- Evaluate performance on representation network *as is*

Zero shot evaluation

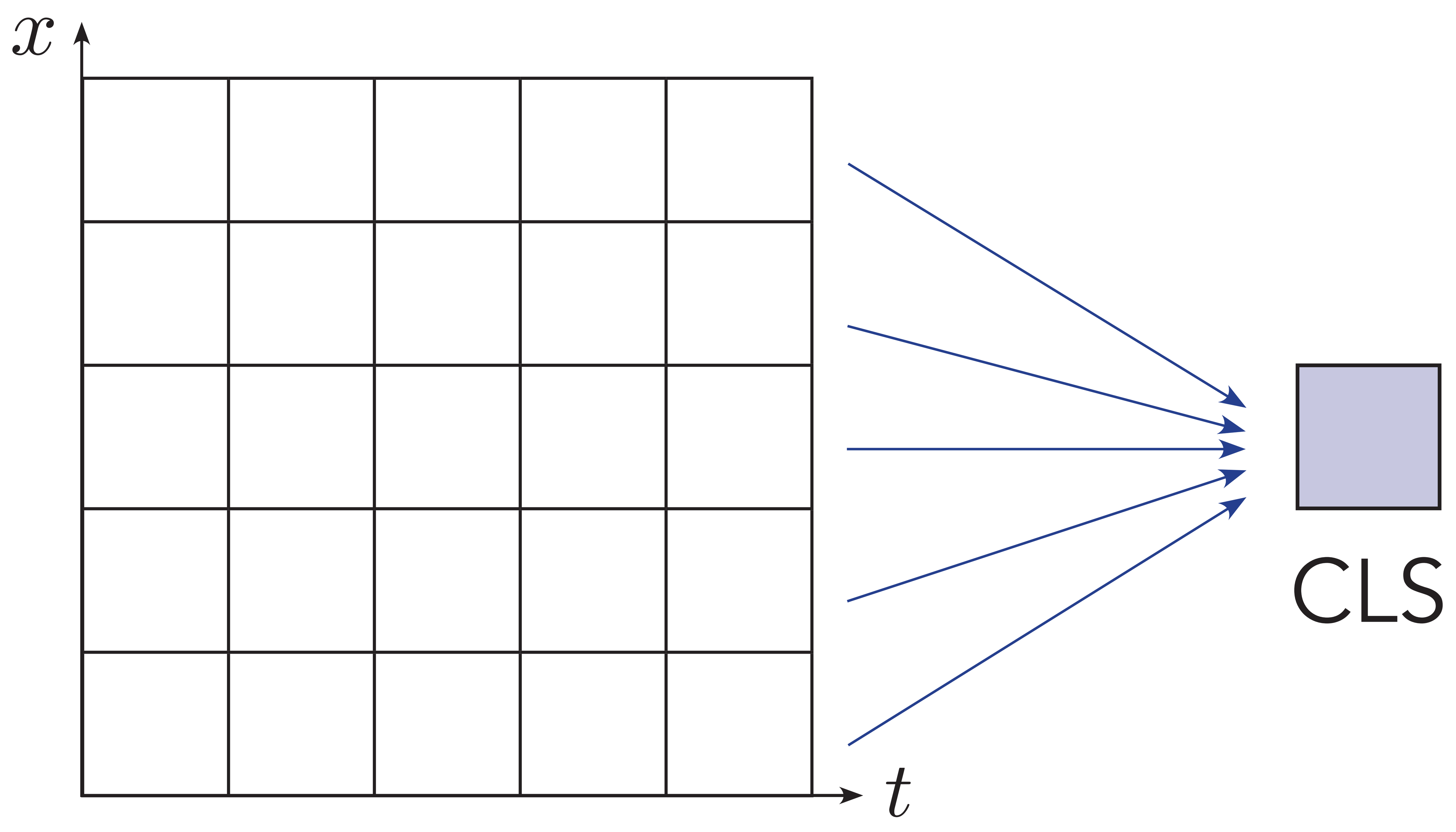
- Evaluate performance on representation network *as is*

Proposal for atmospheric representation learning:

Short term now-/forecasting

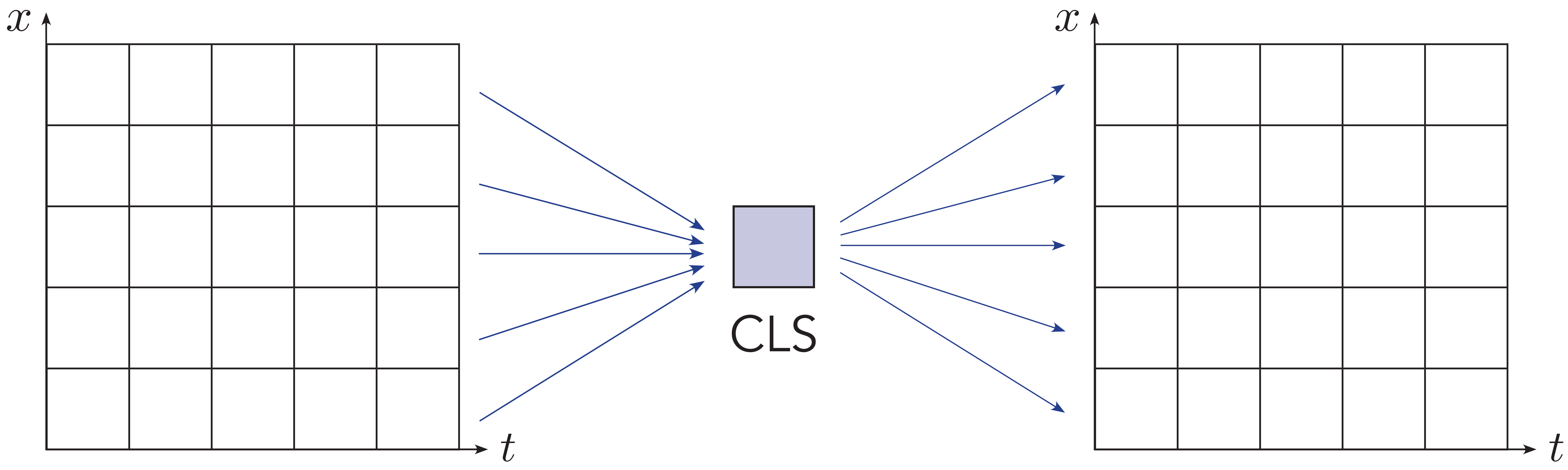
Zero shot evaluation

Embedding

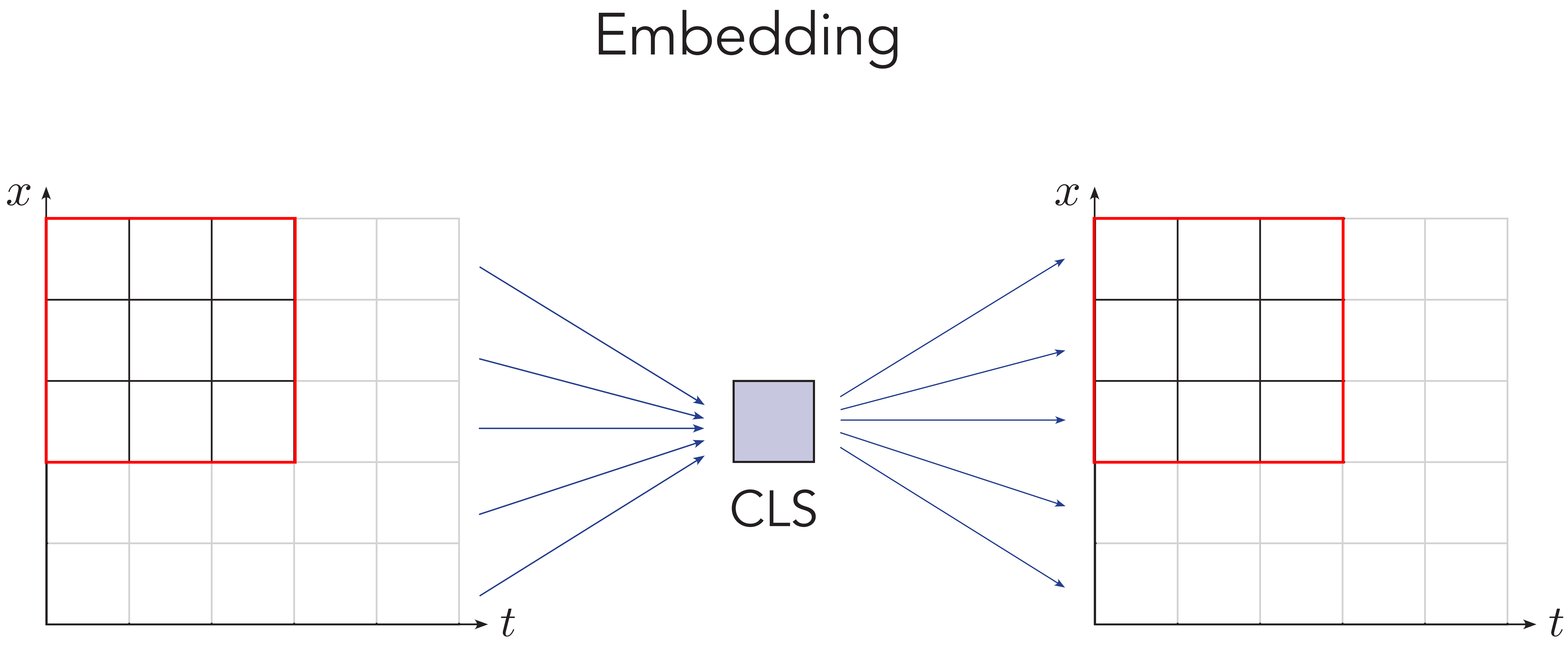


Zero shot evaluation

Embedding

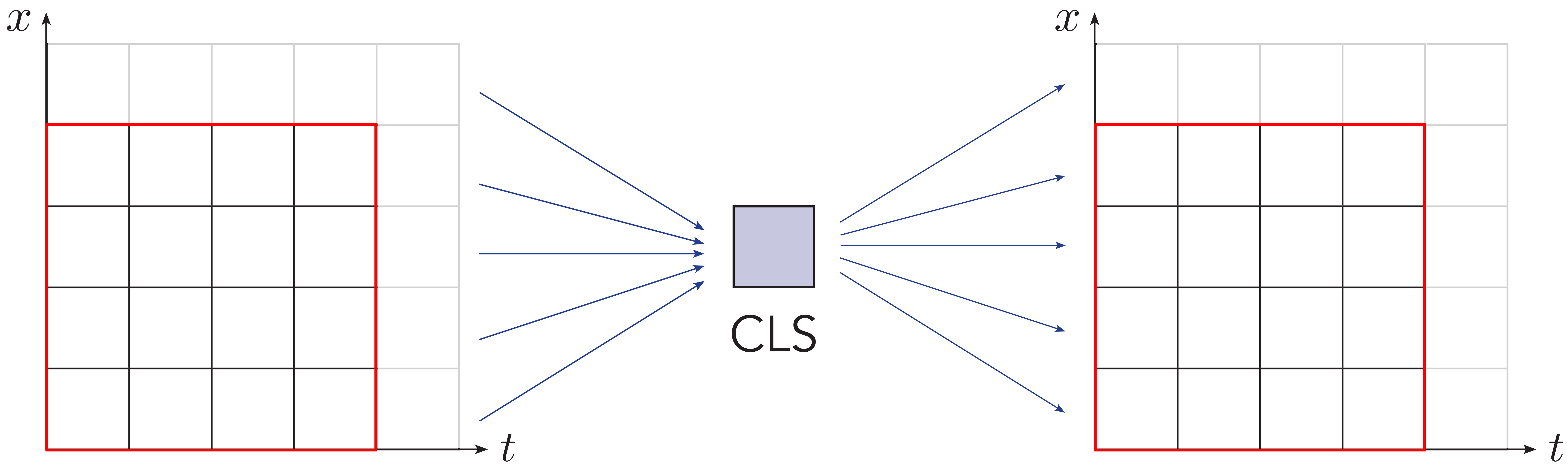


Zero shot evaluation



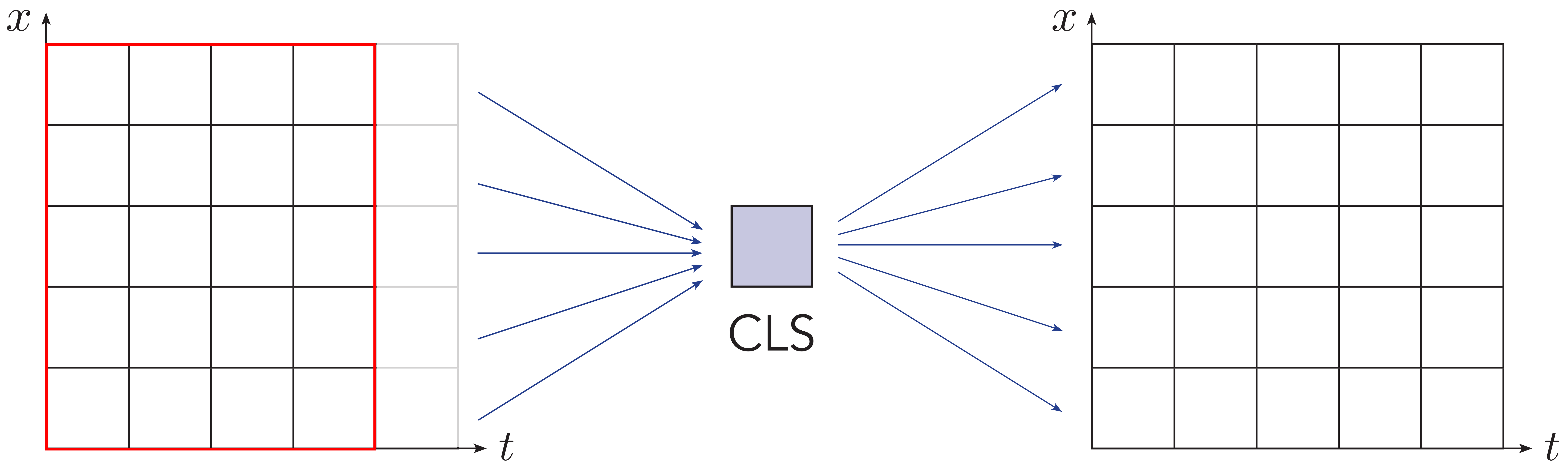
Zero shot evaluation

Embedding

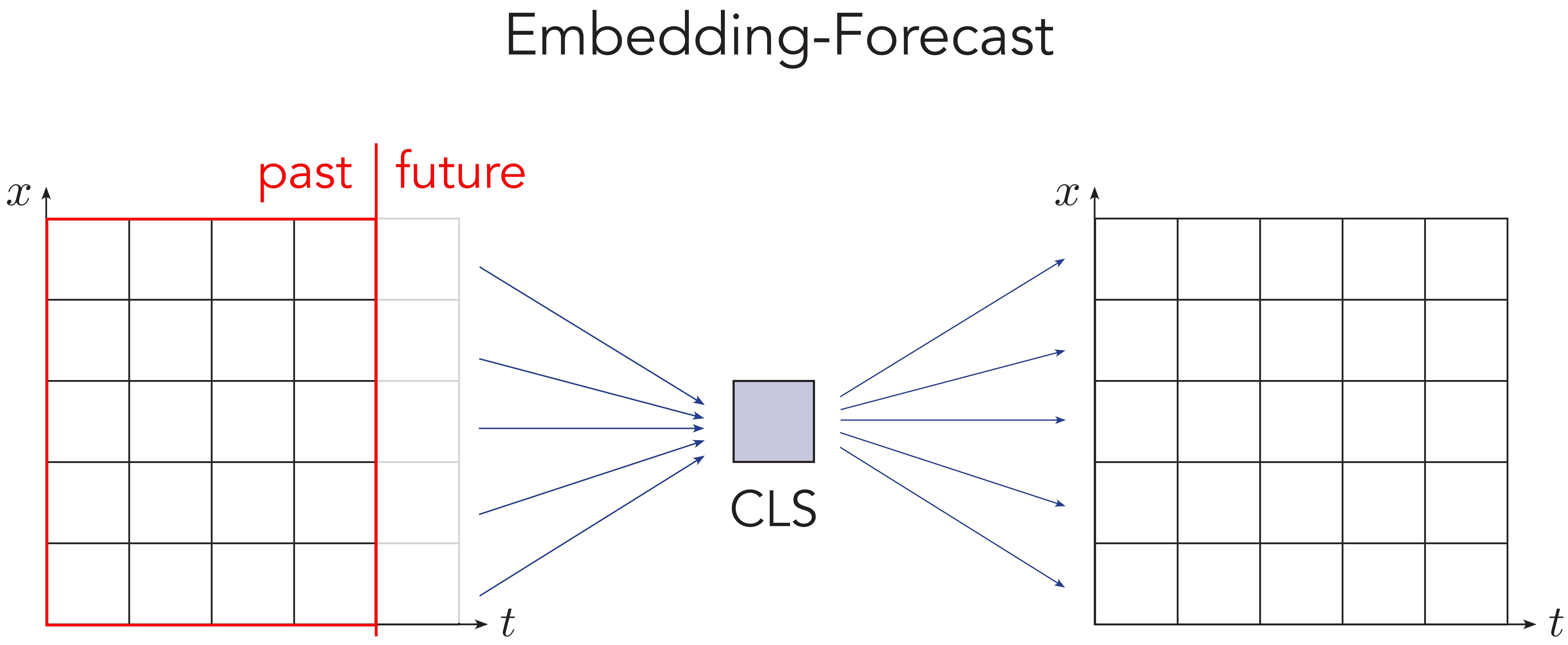


Zero shot evaluation

Embedding-Forecast

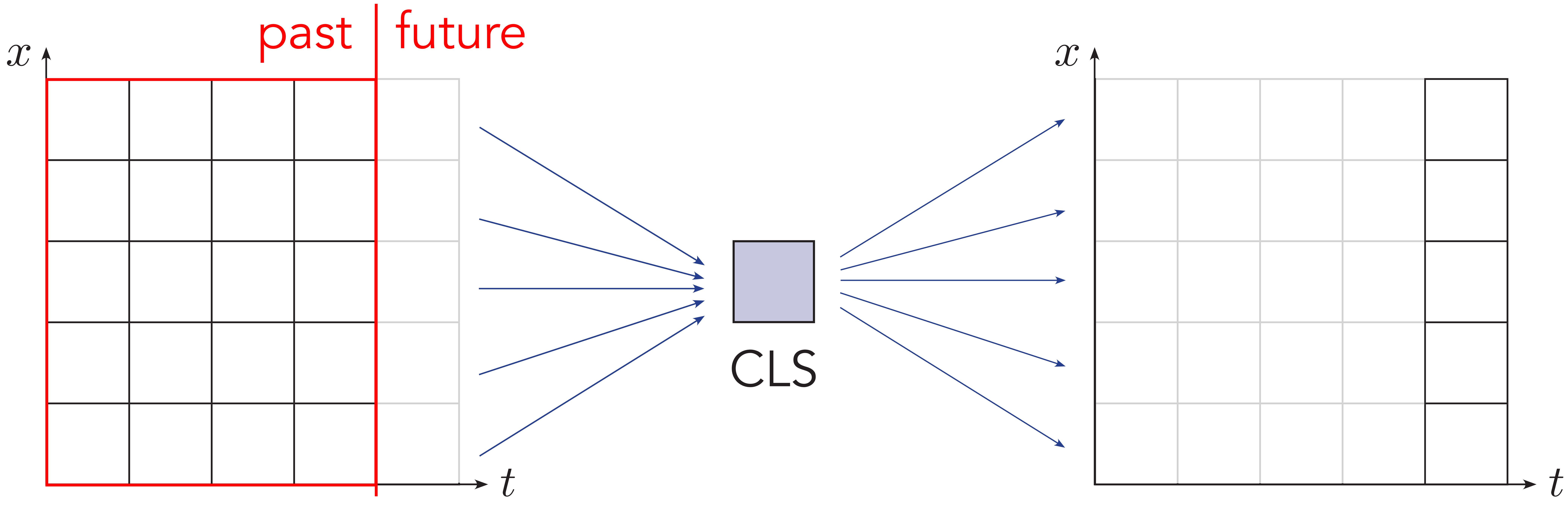


Zero shot evaluation



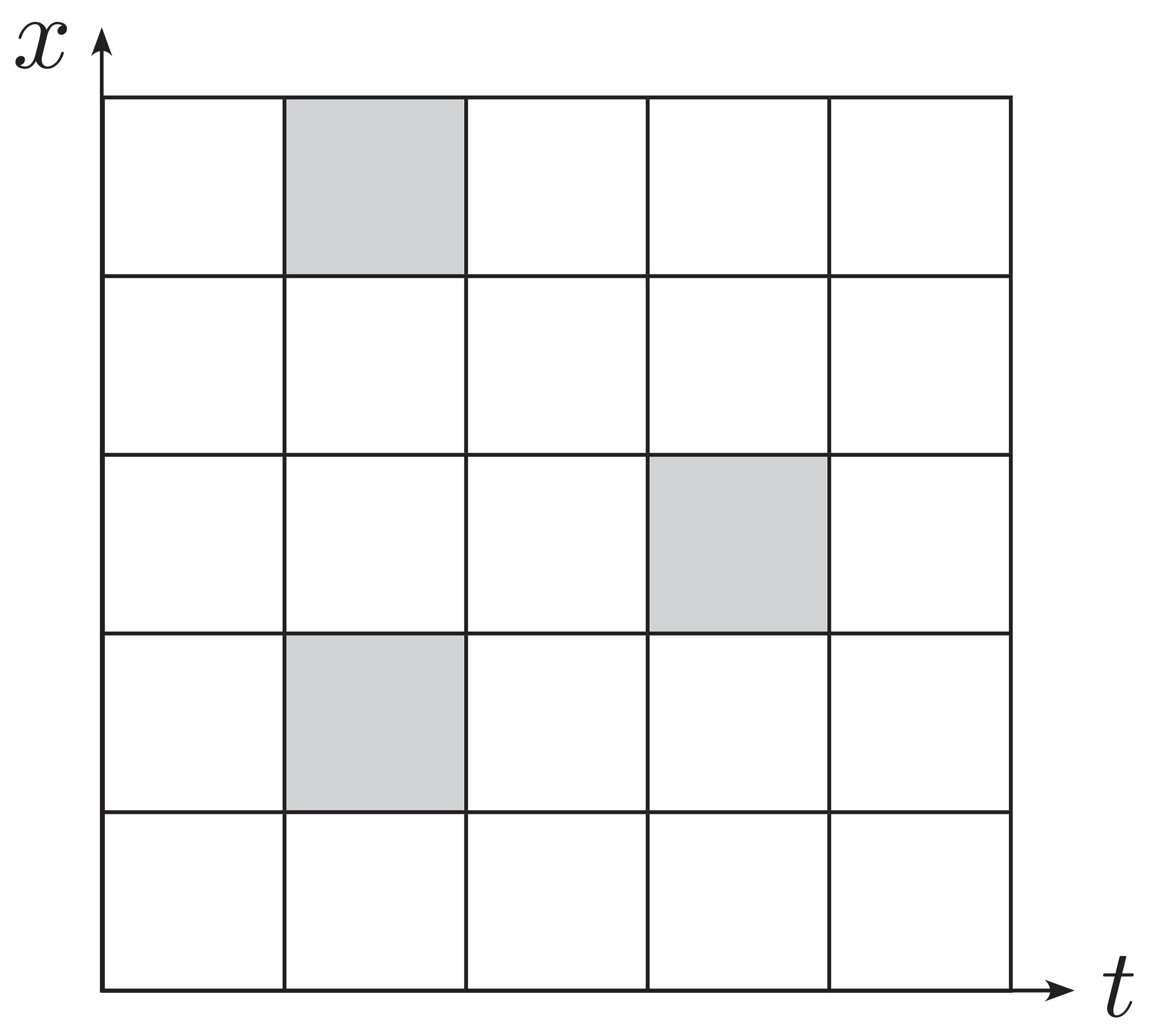
Zero shot evaluation

Embedding-Forecast



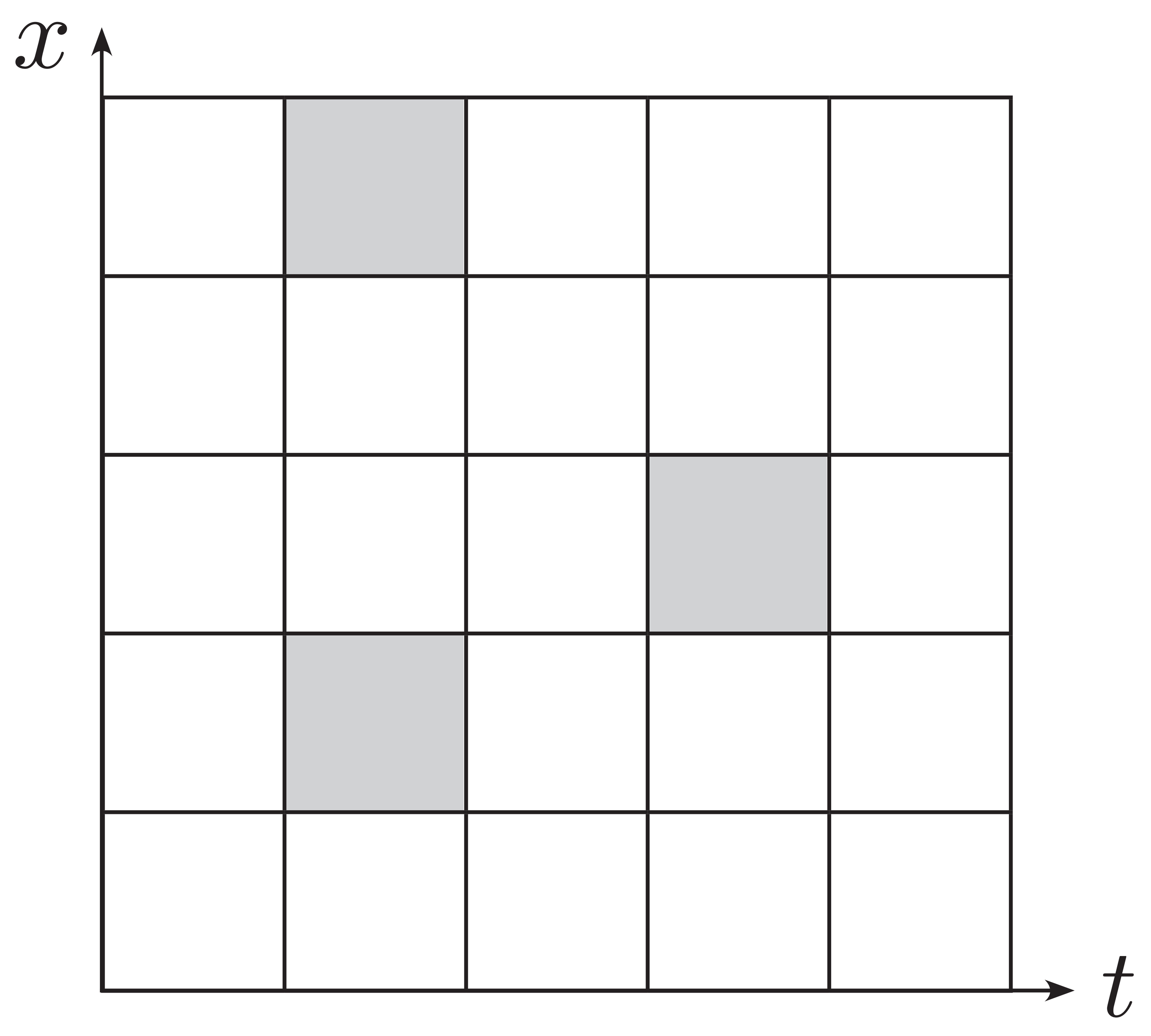
Zero shot evaluation

BERT

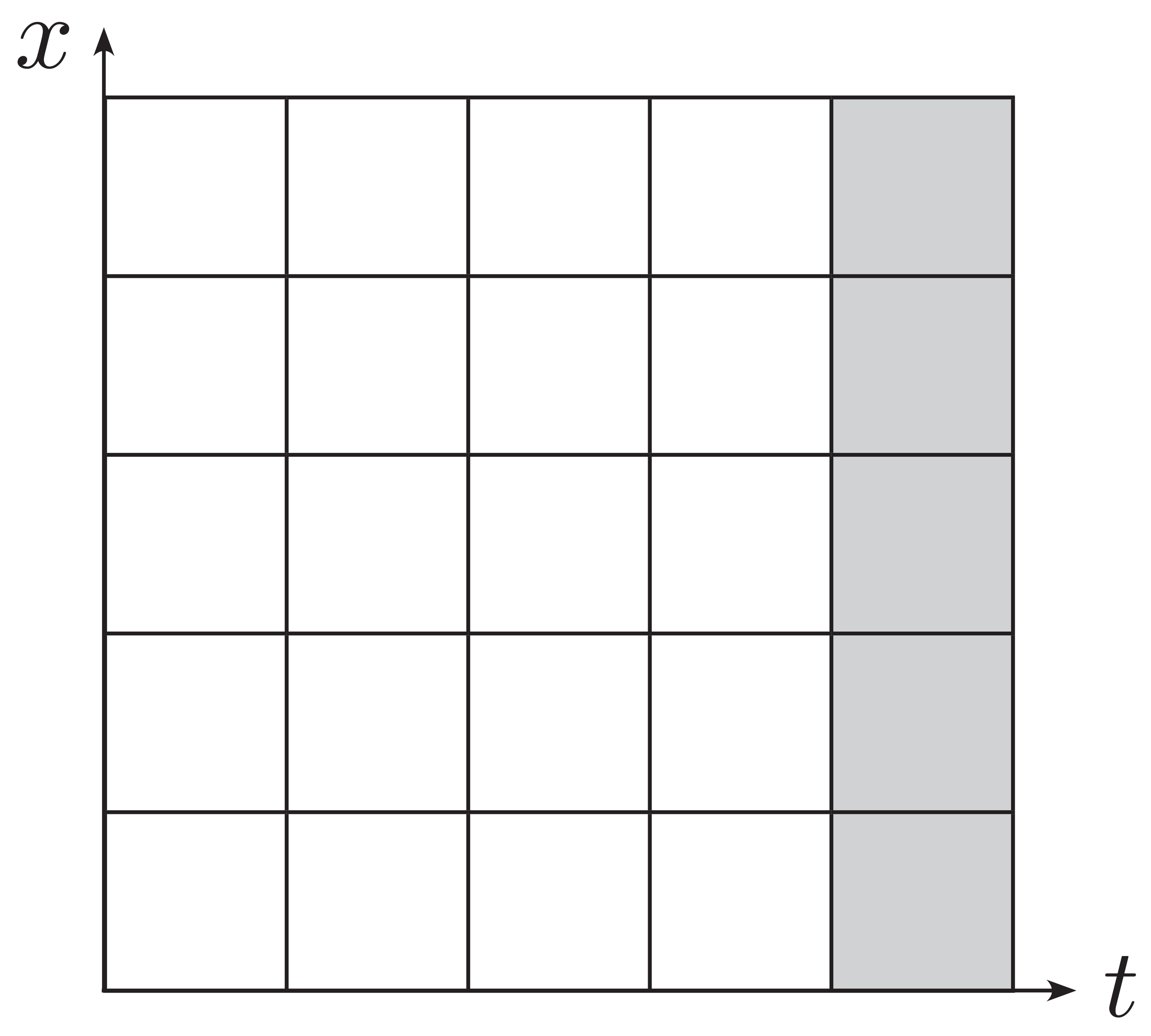


Zero shot evaluation

BERT

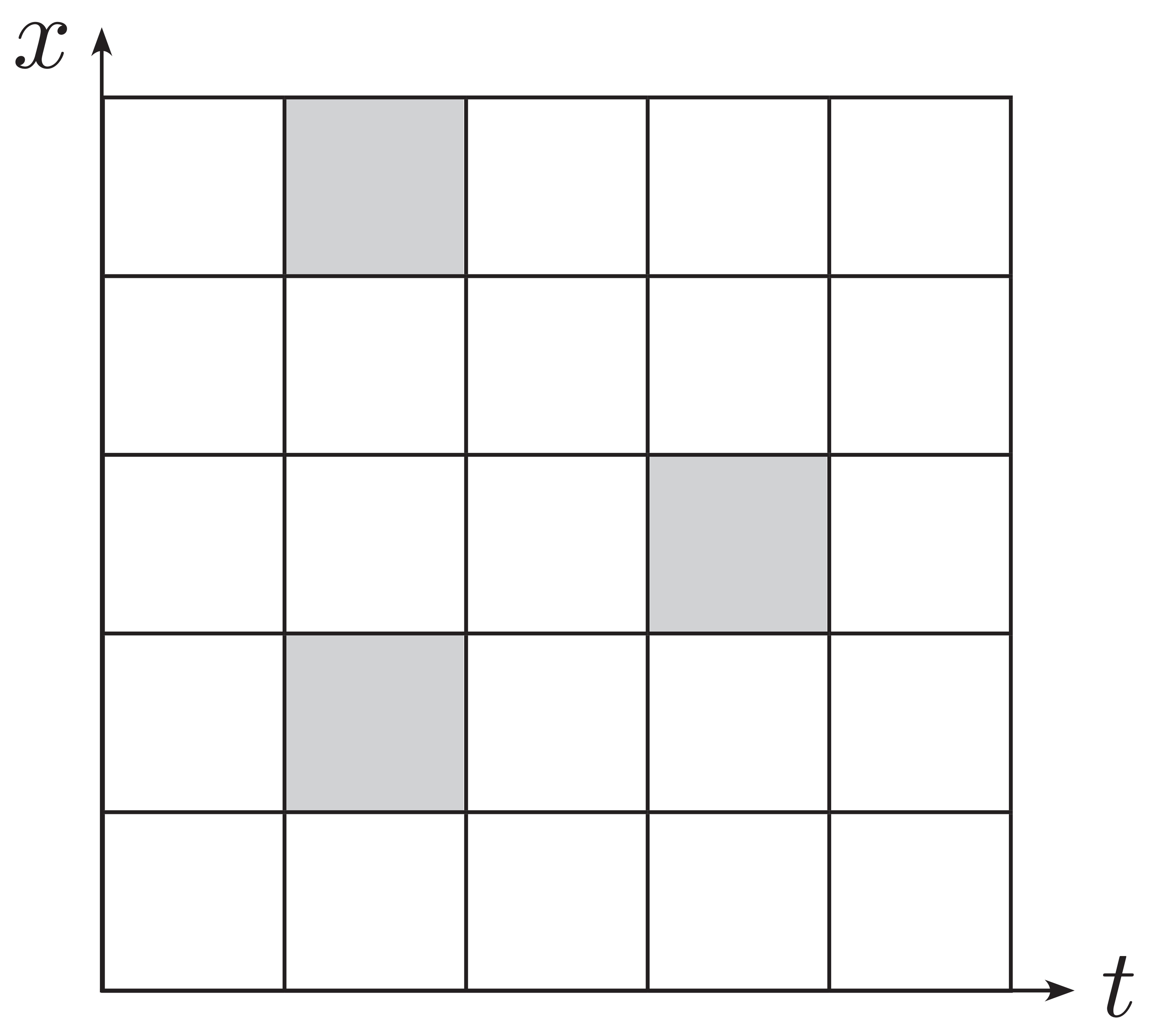


BERT-Forecast

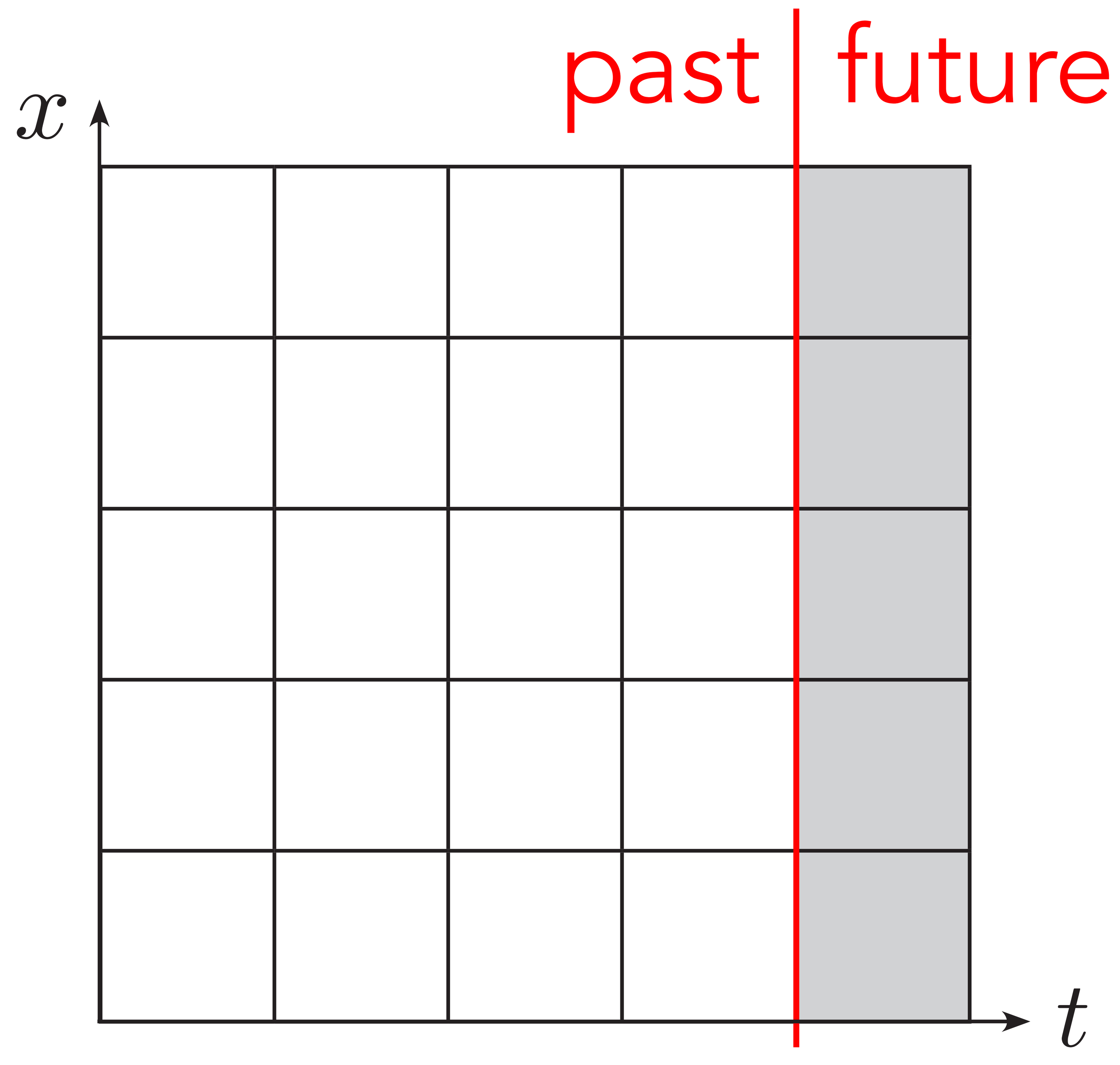


Zero shot evaluation

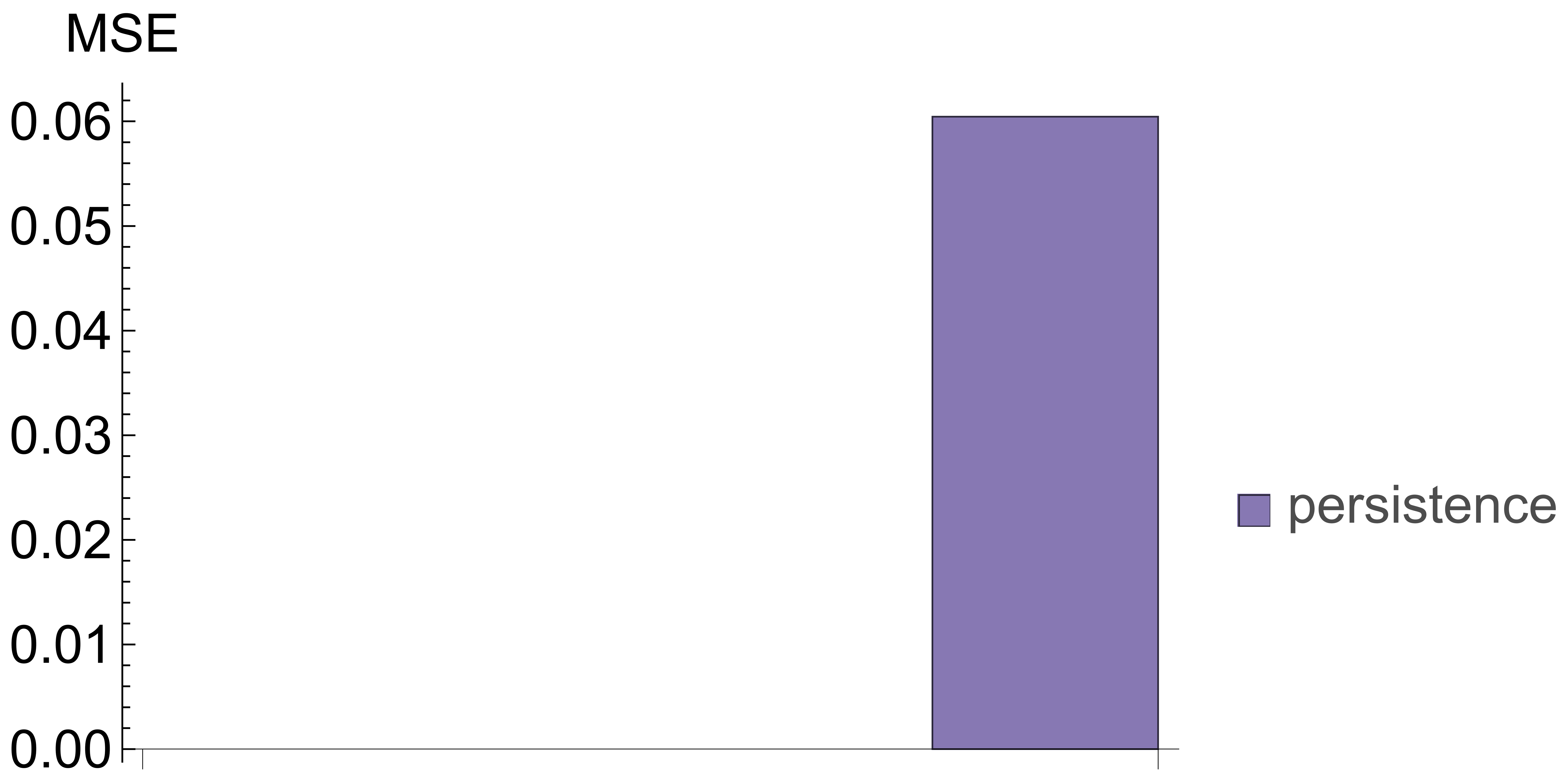
BERT



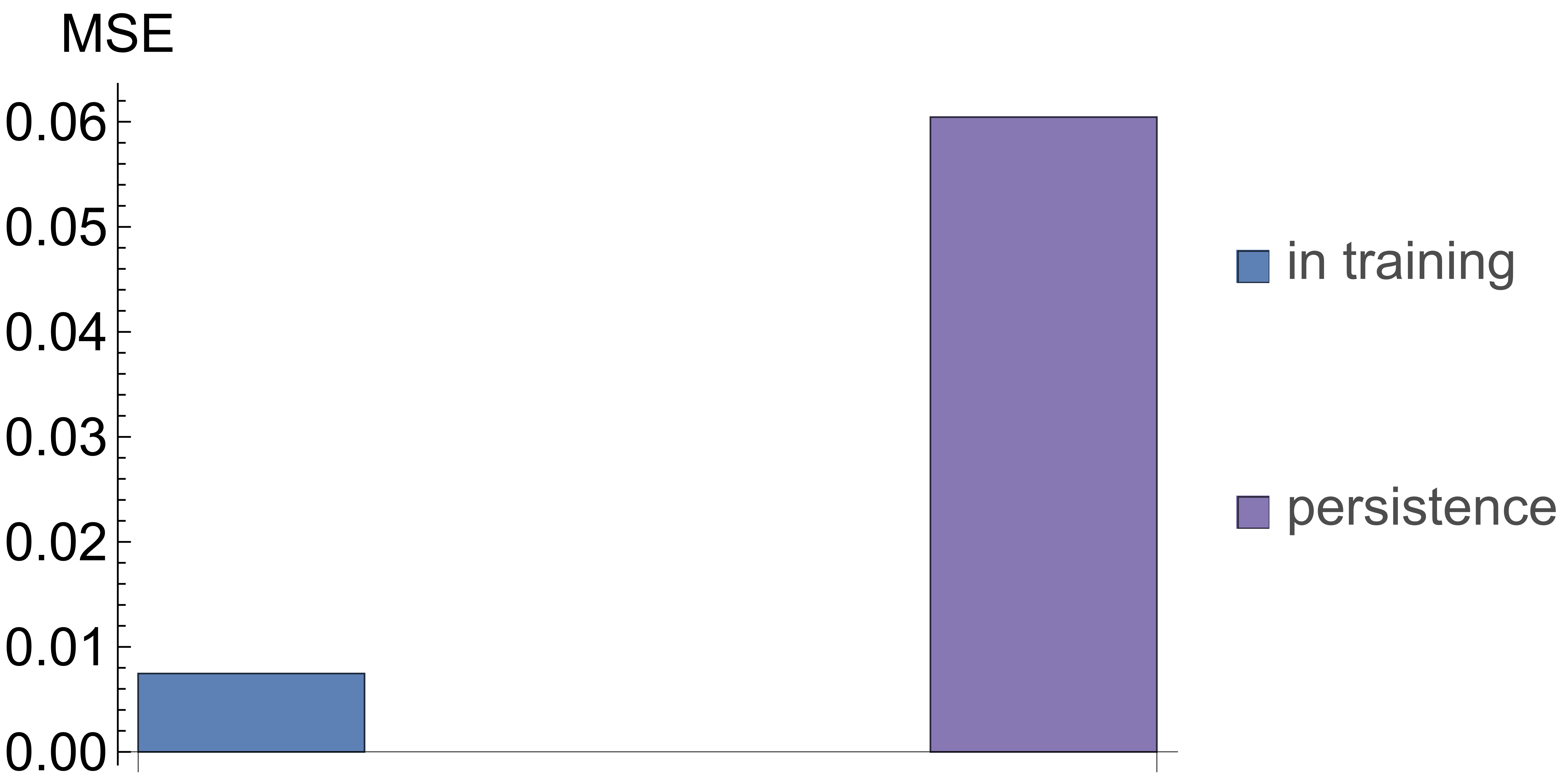
BERT-Forecast



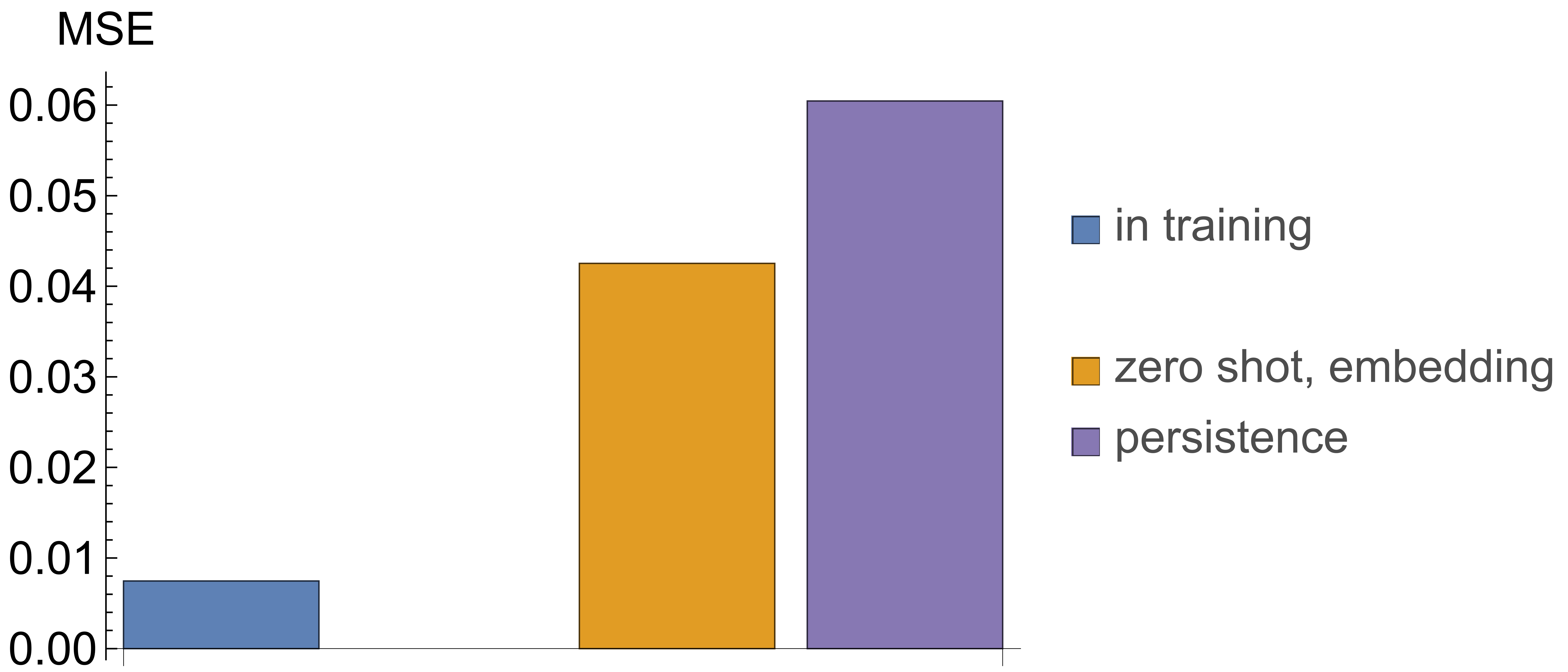
Zero shot performance



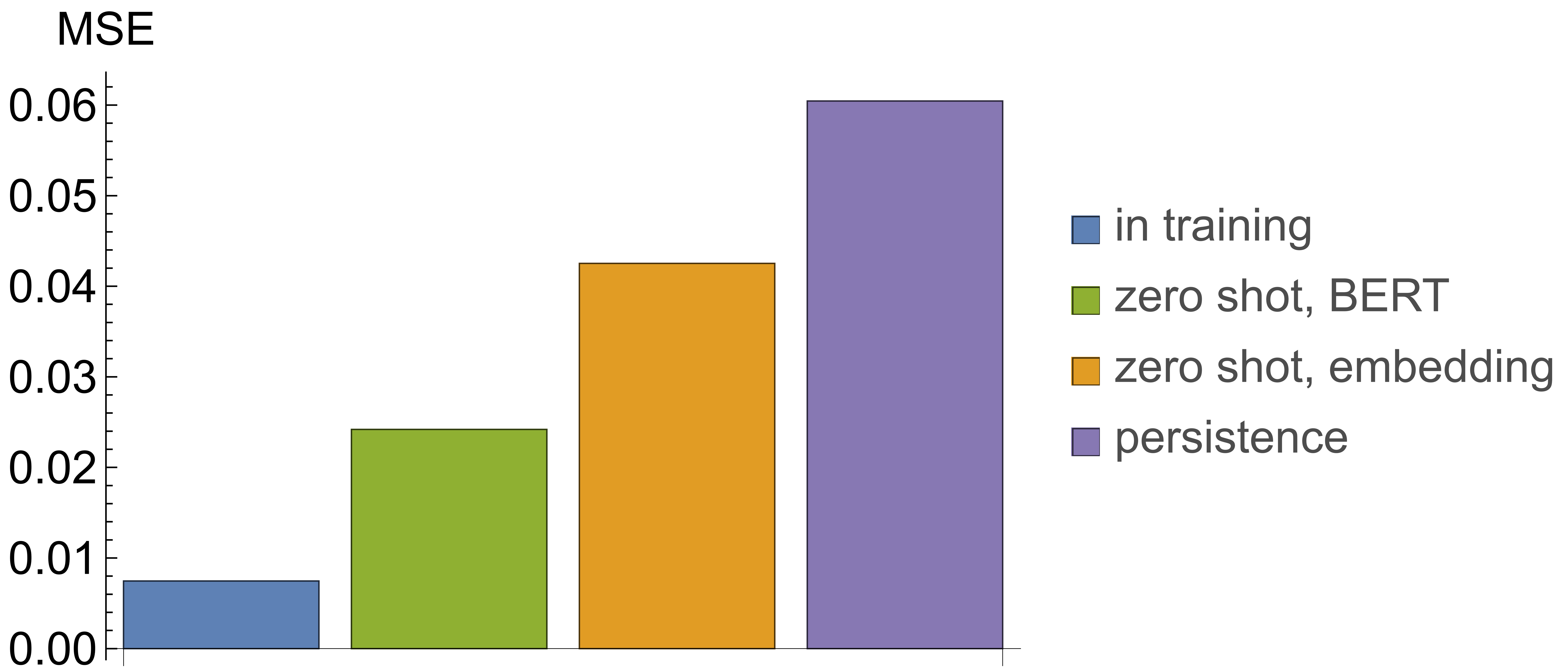
Zero shot performance



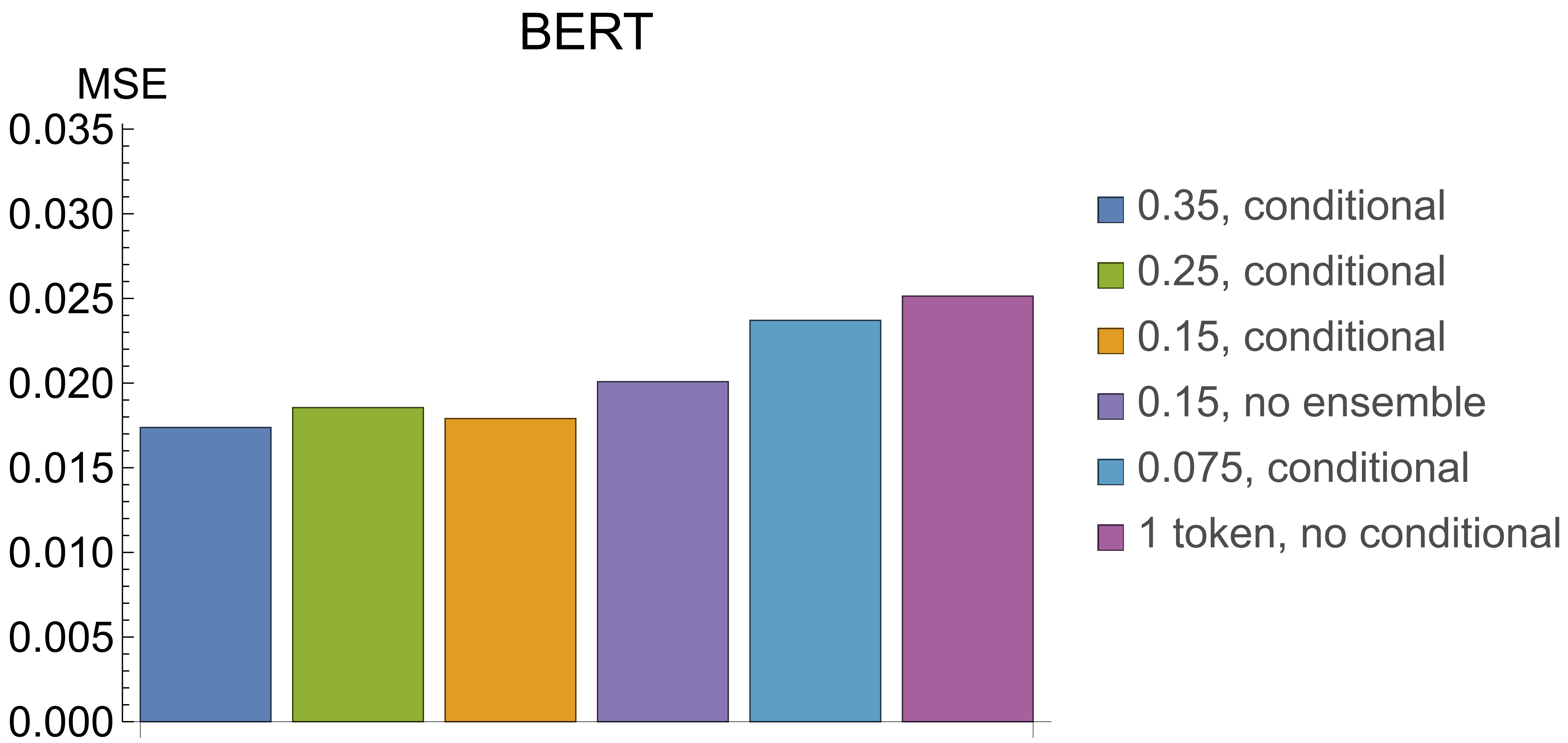
Zero shot performance



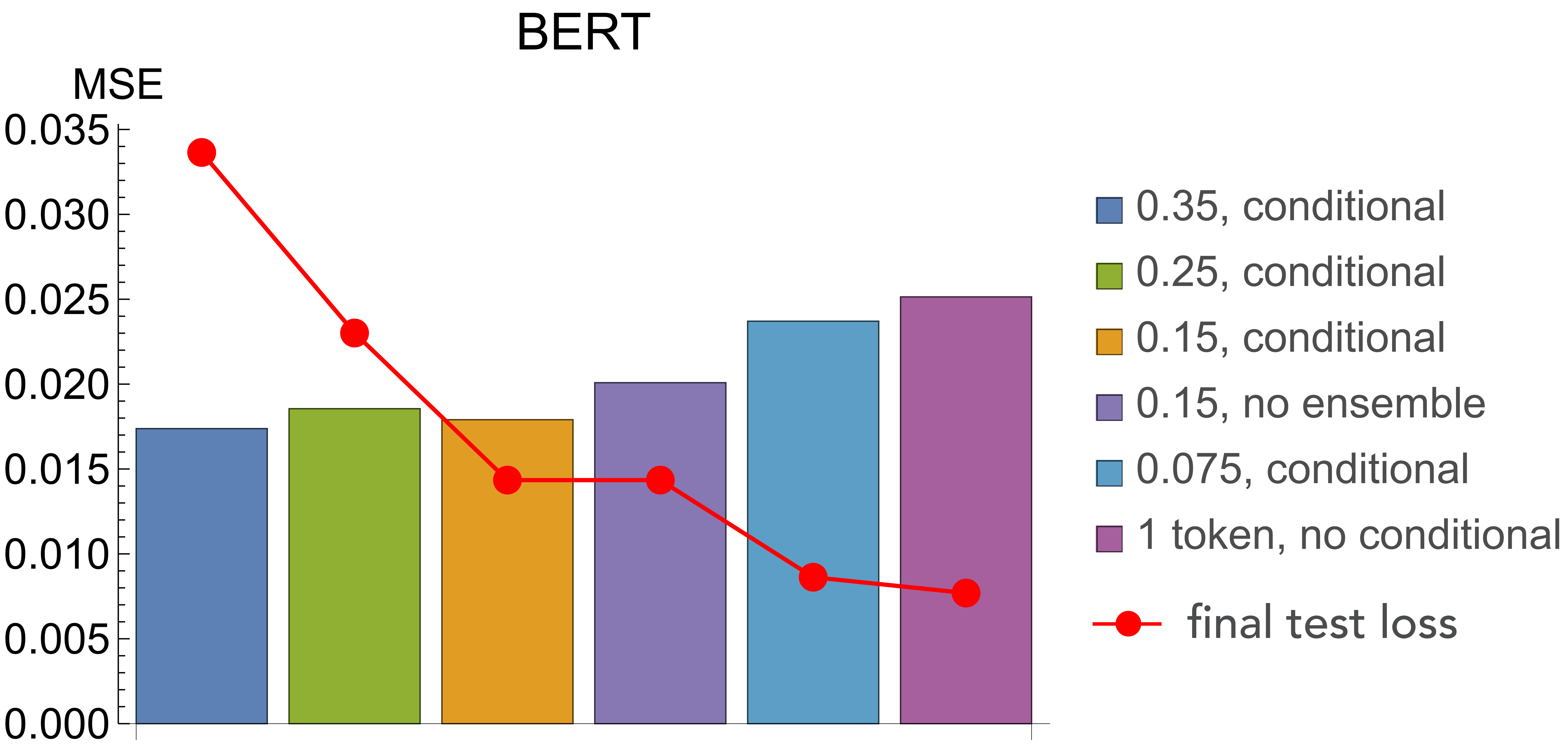
Zero shot performance



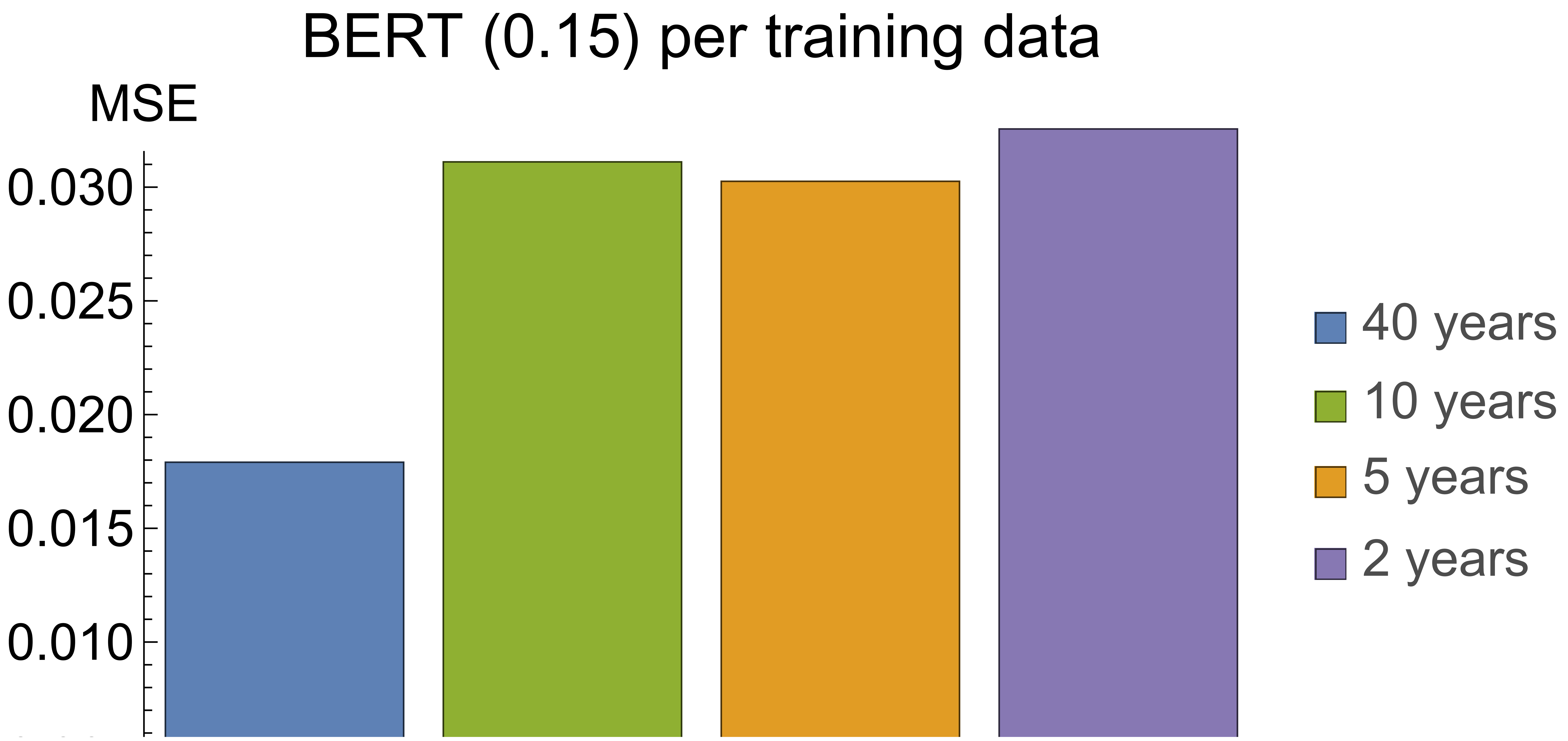
Zero shot performance



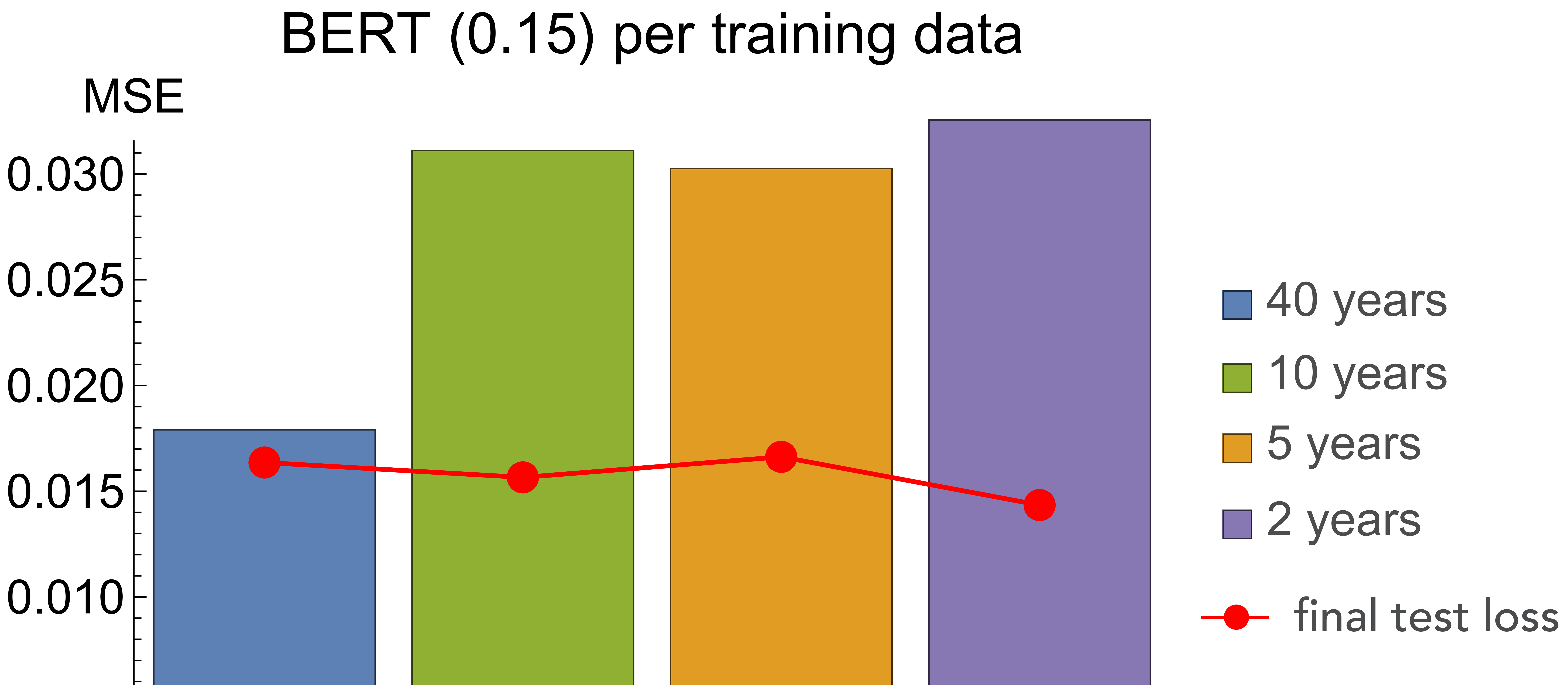
Zero shot performance



Zero shot performance



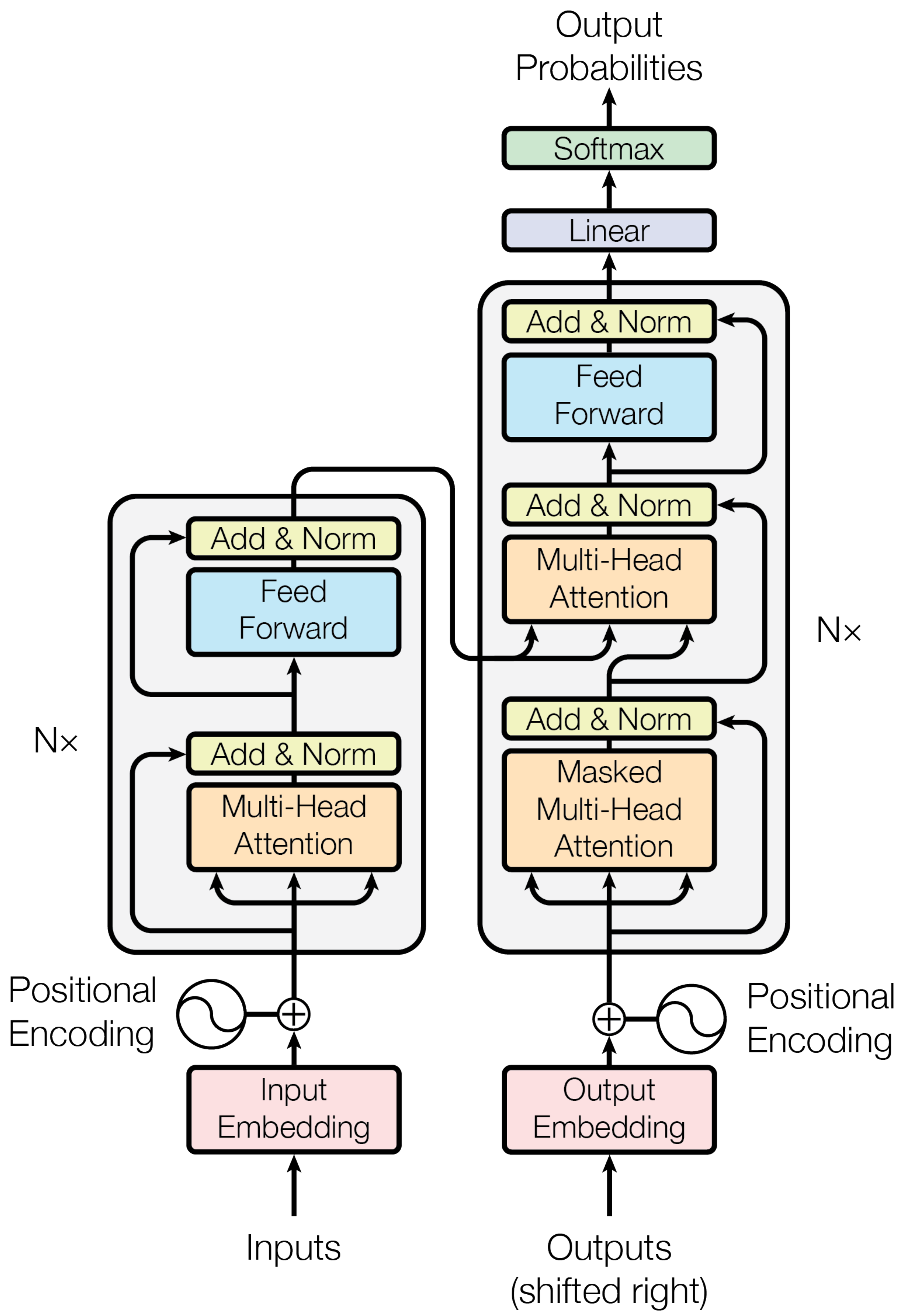
Zero shot performance



AtmoRep: longer term objectives

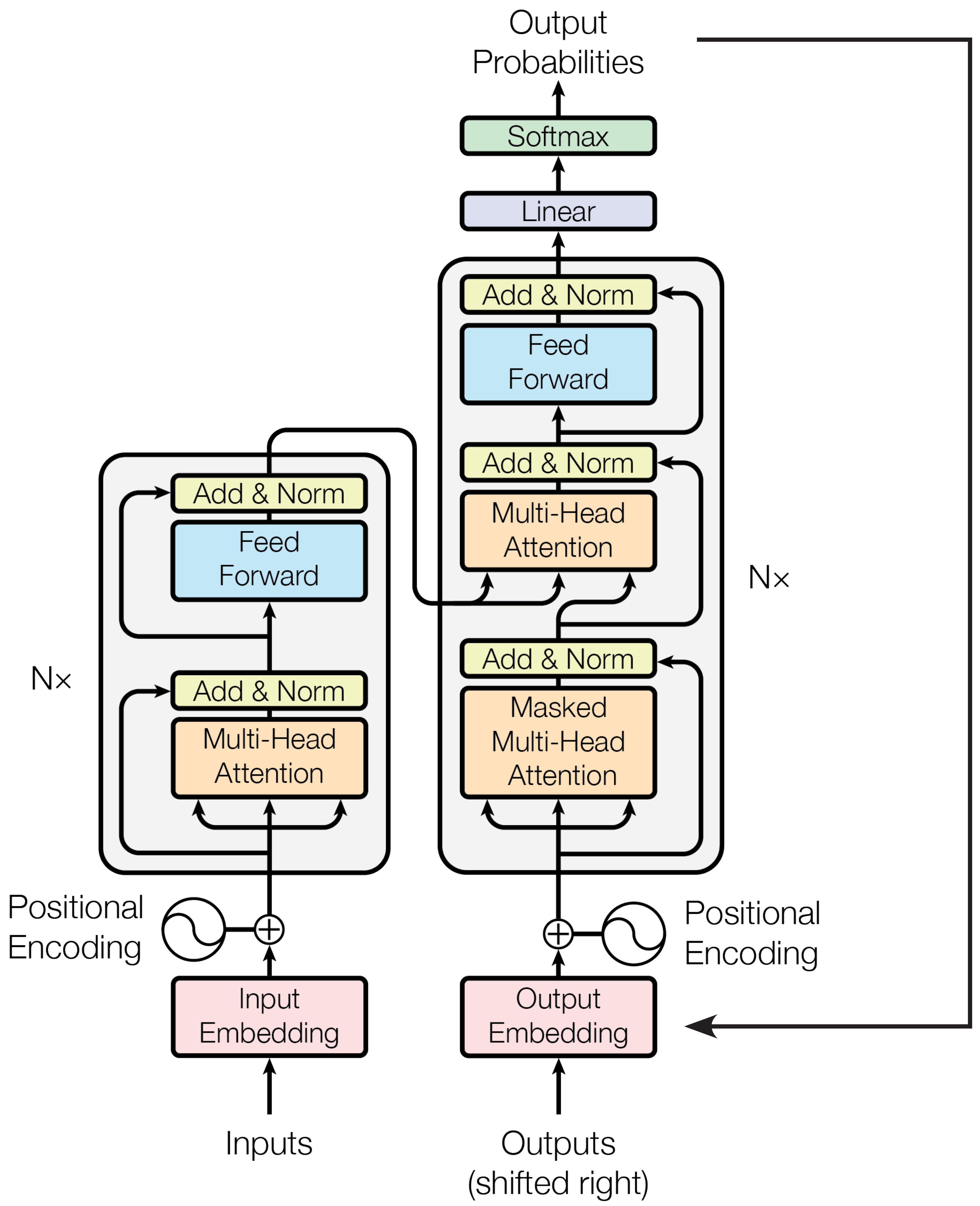
- Weather forecasting
- Climate projections
- Coupled Earth system
- Scientific model
- Training/fine-tuning on direct observational data

Forecasting / projections



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

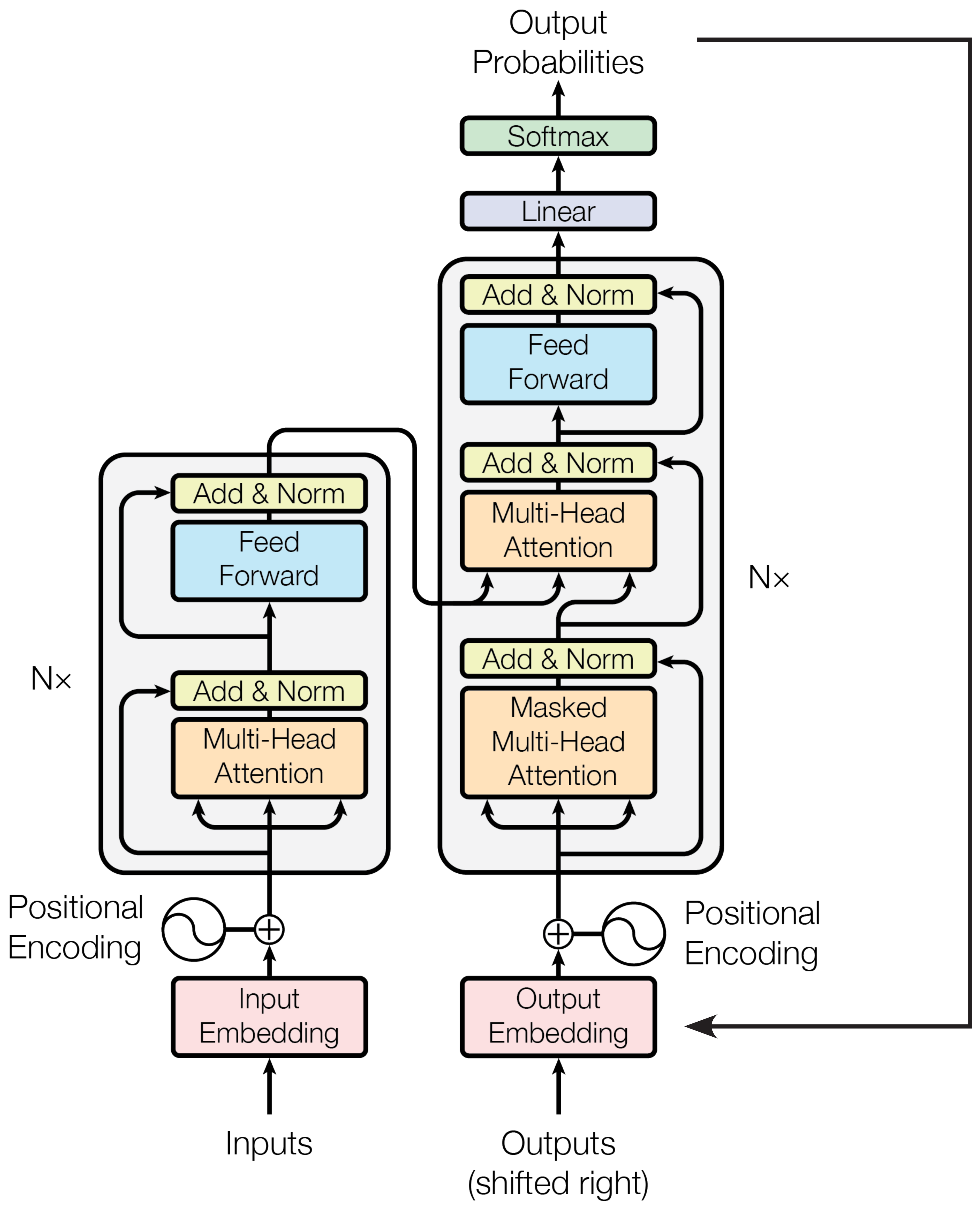
Forecasting / projections



autoregressive,
generative modeling

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Forecasting / projections



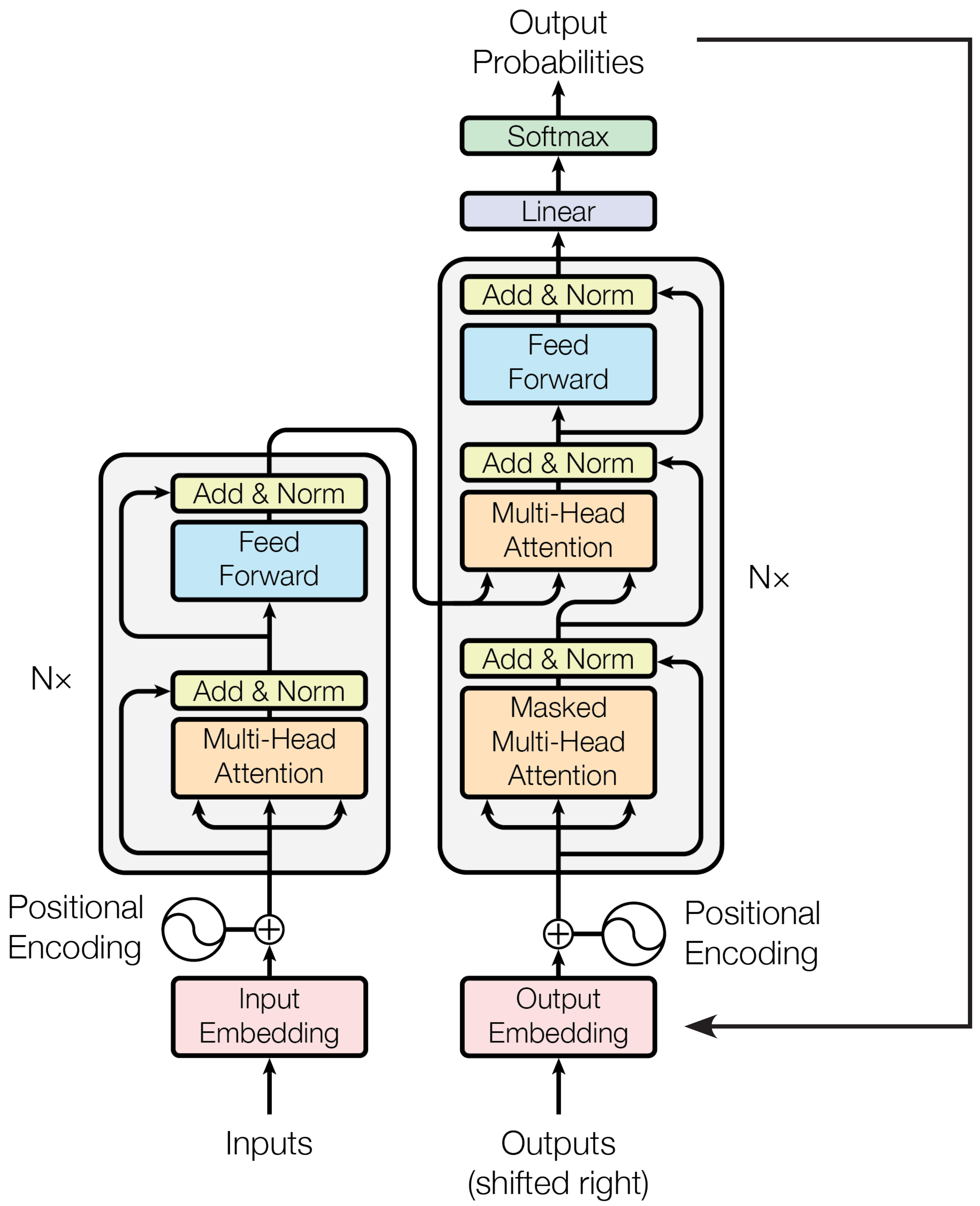
autoregressive,
generative modeling

akin to time stepping
loop (roll out) for fore-
casting/projections

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Forecasting / projections

coarse scale/simple classical model



autoregressive, generative modeling

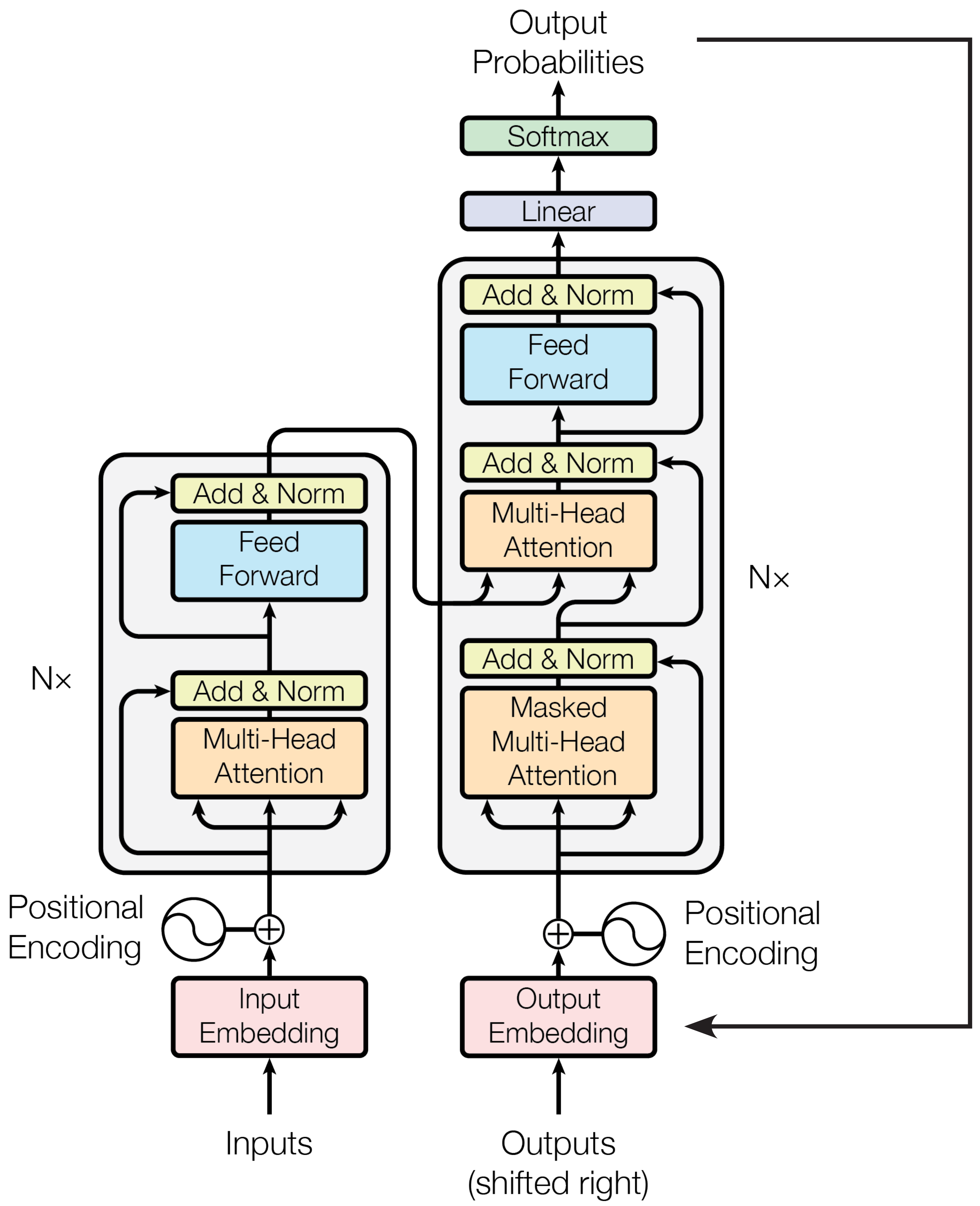
akin to time stepping loop (roll out) for fore-casting/projections

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Forecasting / projections

coarse scale/simple classical model

slow climate variables



autoregressive, generative modeling

akin to time stepping loop (roll out) for fore-casting/projections

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Forecasting / projections

- Local model: apply model locally but couple through loss to get consistent forecast[1]
  - Memory efficient and easily parallelizable

[1] A. Blanchard, N. Parashar, B. Dodov, C. Lessig, and T. Sapsis. *A multi-scale deep learning framework for projecting weather extremes*. In NEURIPS 2021 Workshop on Climate Change (spotlight talk). 2022.

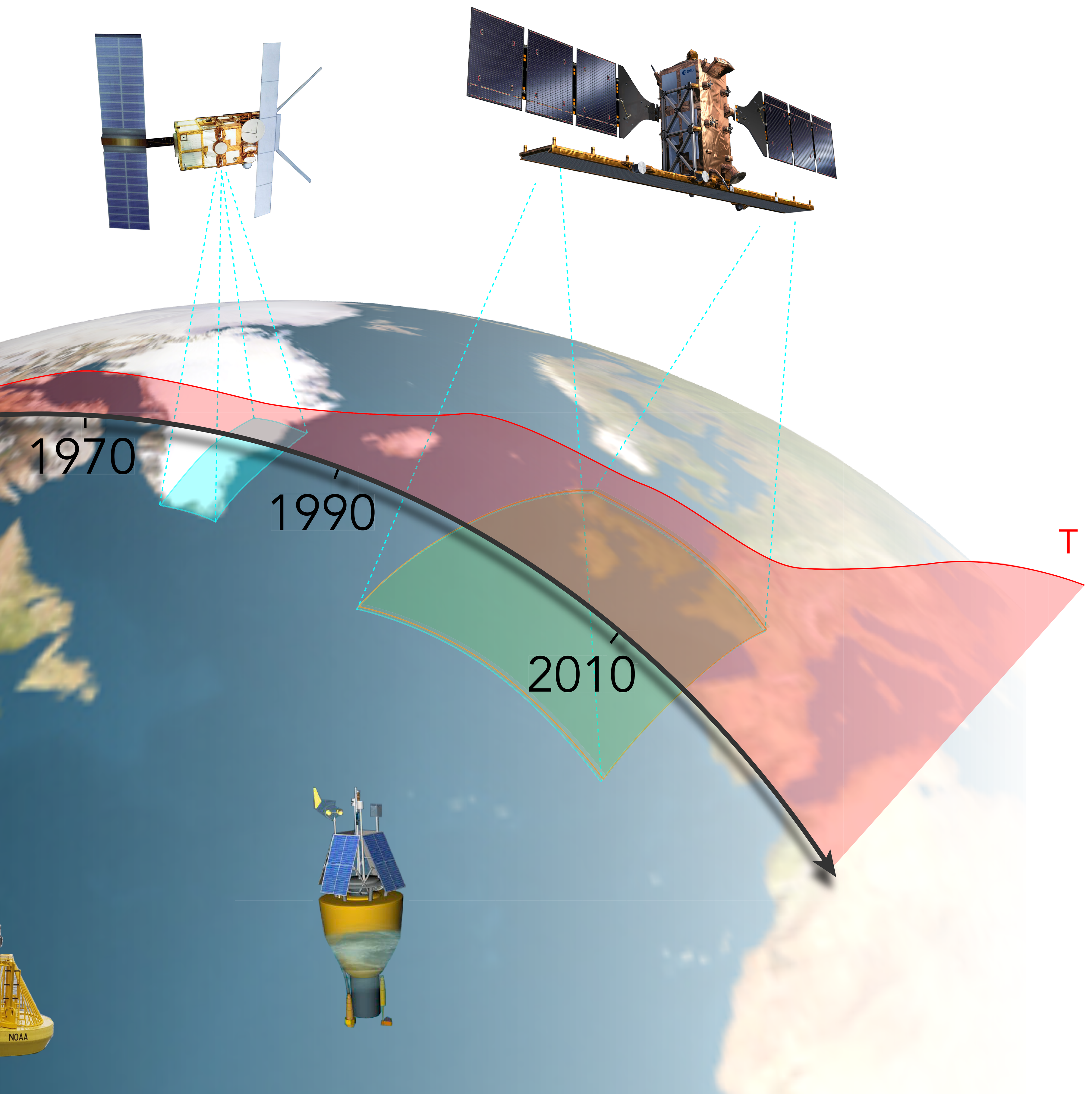
Forecasting / projections

- Local model: apply model locally but couple through loss to get consistent forecast[1]
  - Memory efficient and easily parallelizable
- How to represent multi-res structure in a transformer?
  - Number of tokens or embedding dimension?
  - Couple it in UNet like structure between encoder and decoder

[1] A. Blanchard, N. Parashar, B. Dodov, C. Lessig, and T. Sapsis. *A multi-scale deep learning framework for projecting weather extremes*. In NEURIPS 2021 Workshop on Climate Change (spotlight talk). 2022.

Current / next steps

- Complete representation learning model
  - Scale data and network size
  - Different training tasks and protocols



Current / next steps

◦ Complete representation learning model

◦ Downstream applications

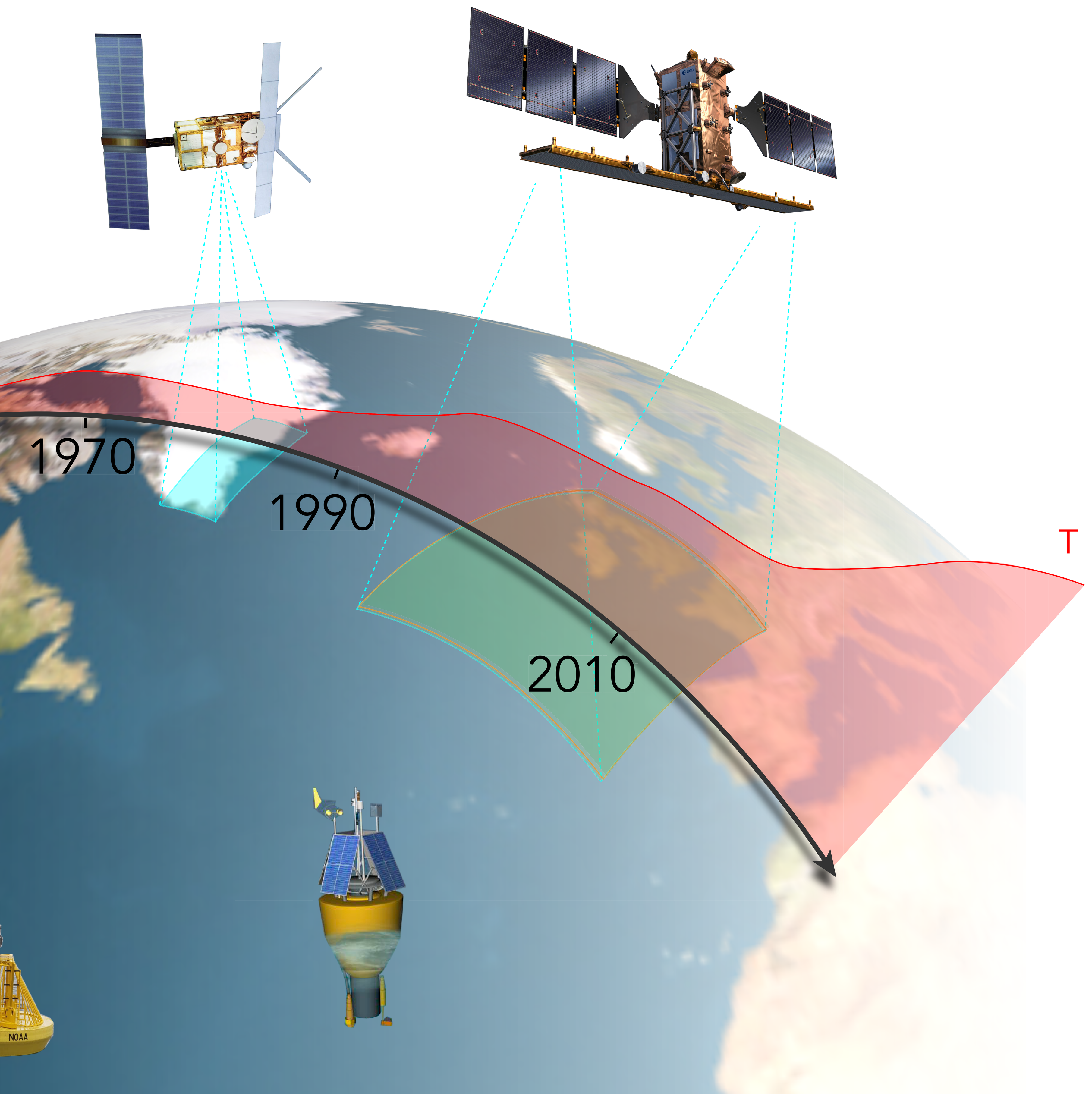
  › Weather forecasting

  › Hurricane tracking

  › Wind farm placement

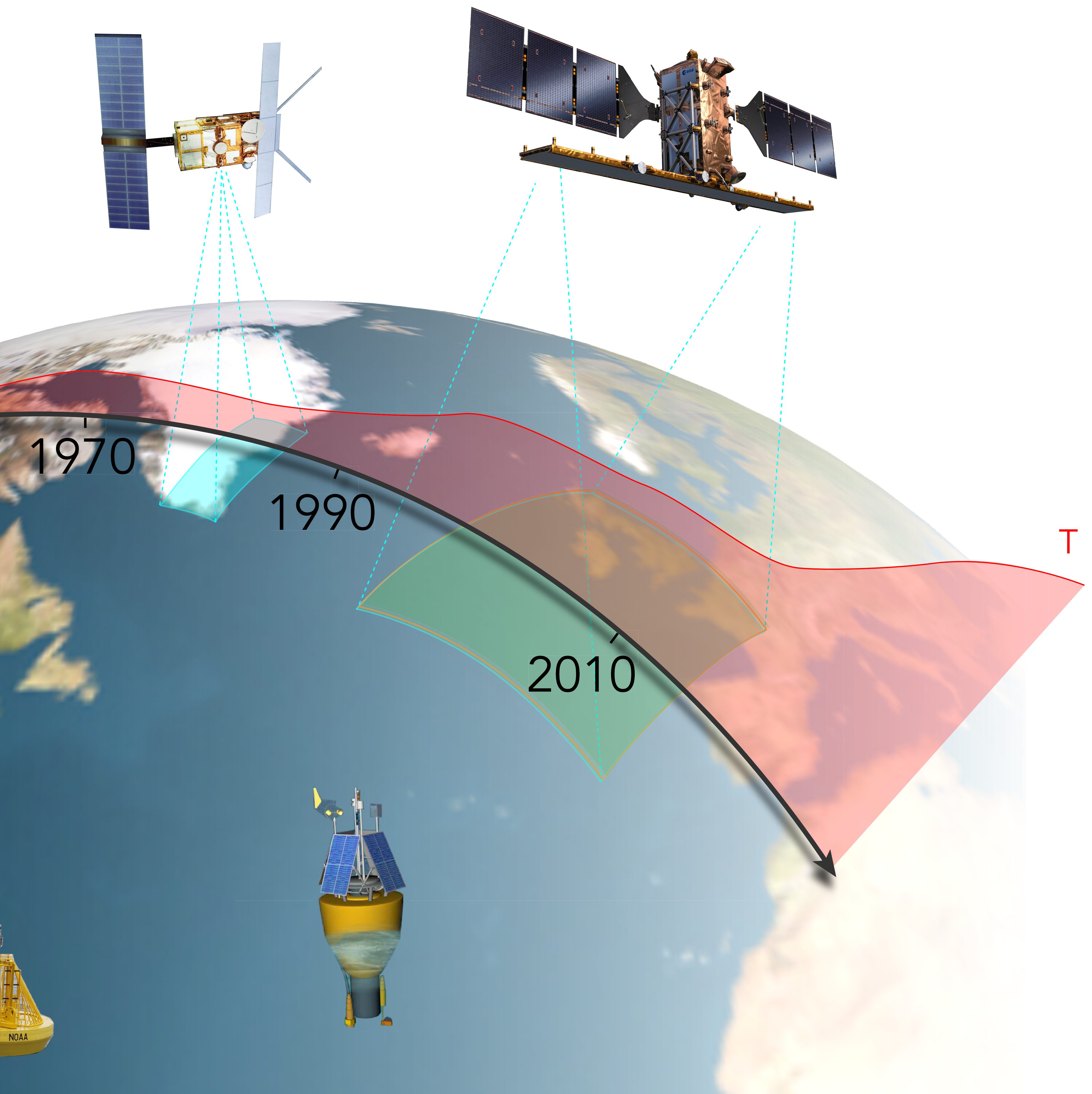
  › Air pollution forecasting

  › ...



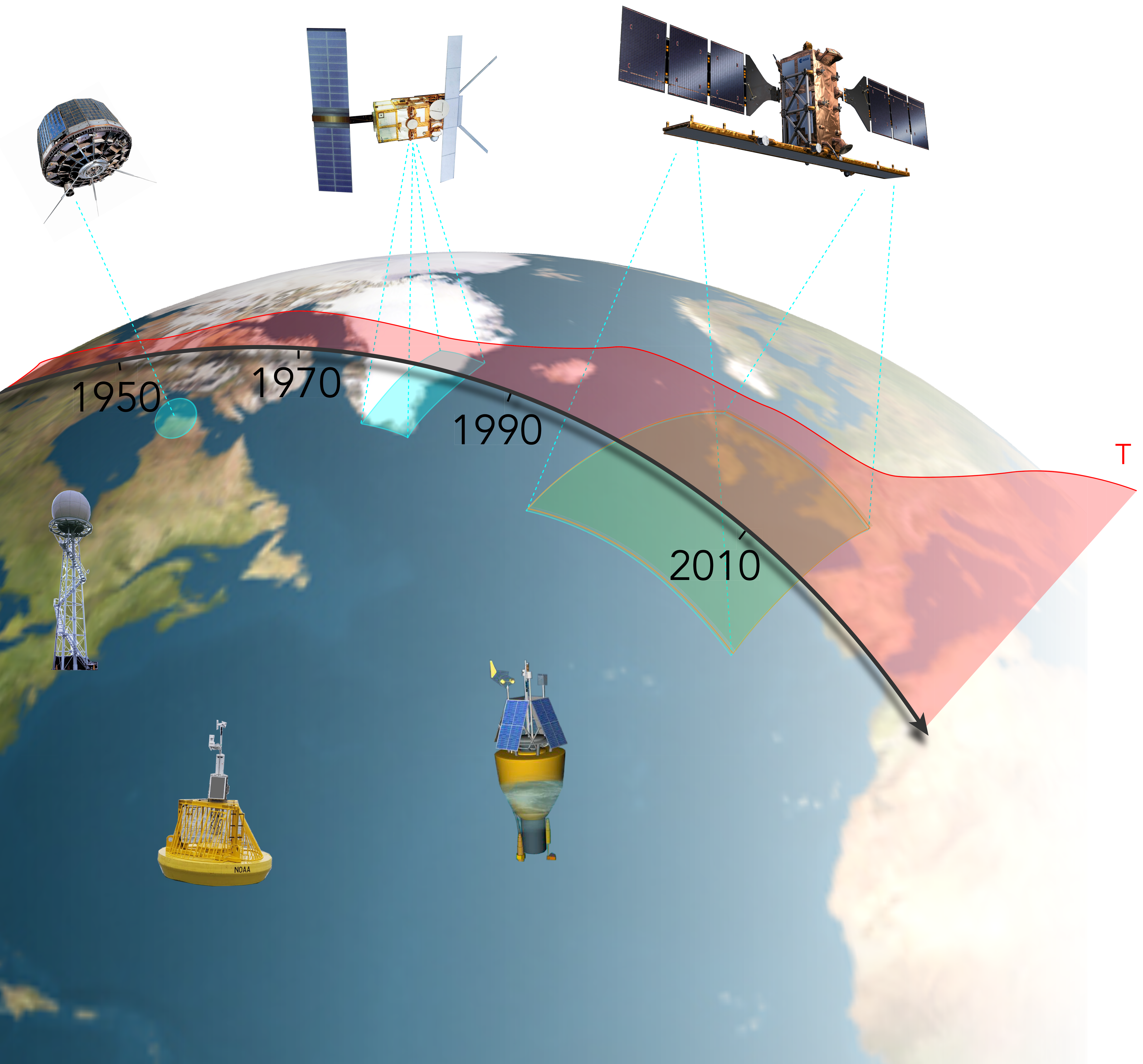
Current / next steps

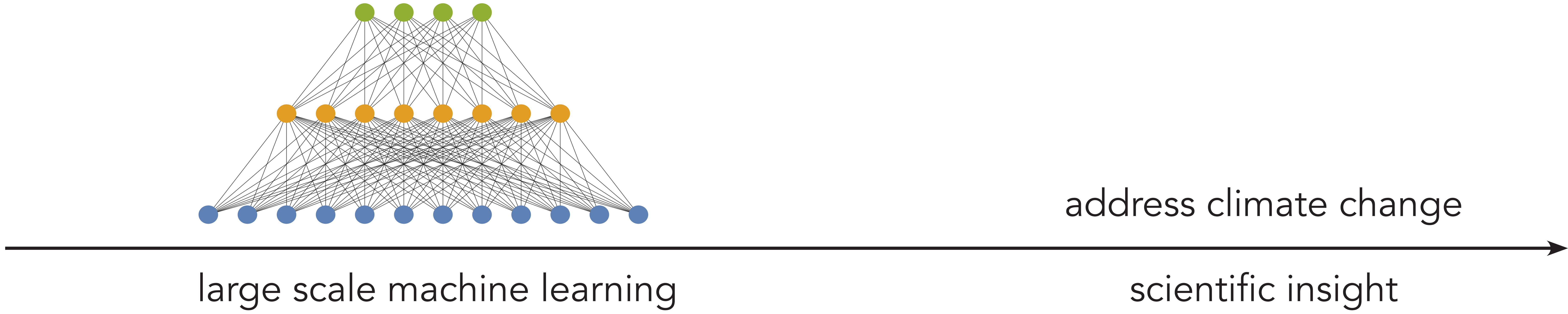
- Complete representation learning model
- Downstream applications
- Strong convection events
  - Improve forecasting
  - Study representation of physics in the network



AtmoRep

Large scale representation learning of atmospheric dynamics





AtmoDist: evaluation