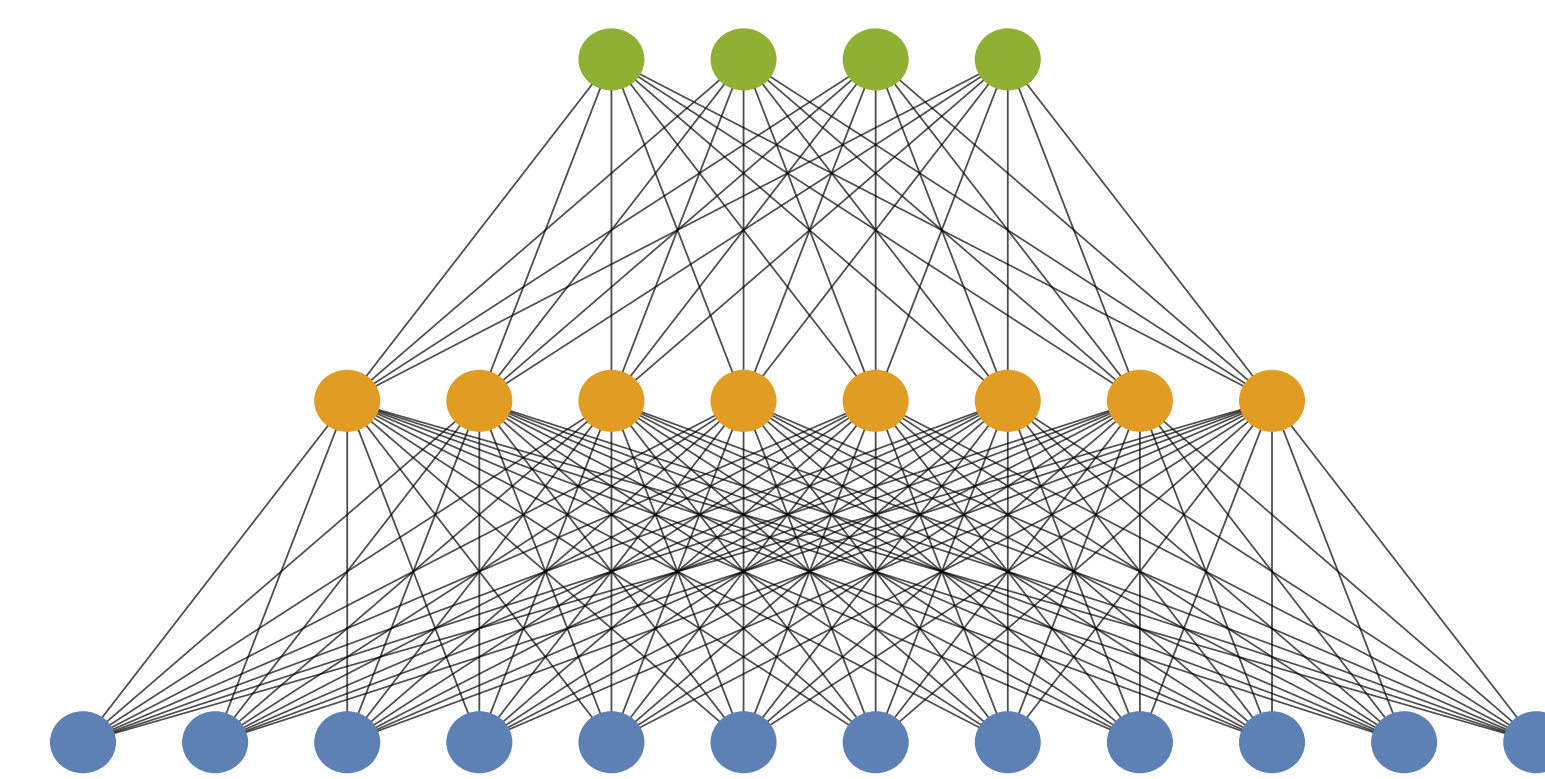
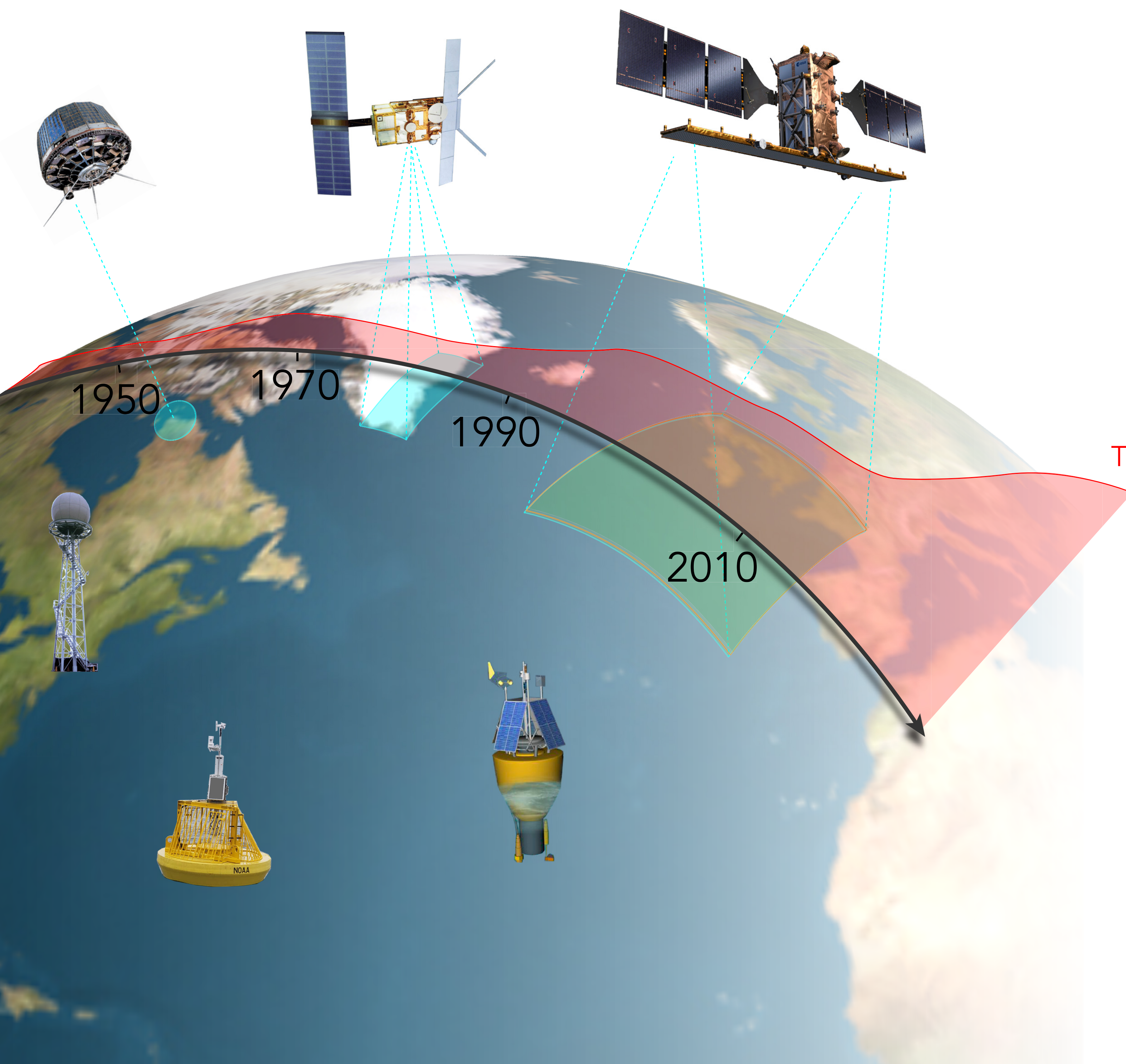


AtmoRep

Large scale representation learning of atmospheric dynamics

Christian Lessig, Ilaria Luise, Martin Schultz, et al.



large scale representation learning

applications

scientific insight

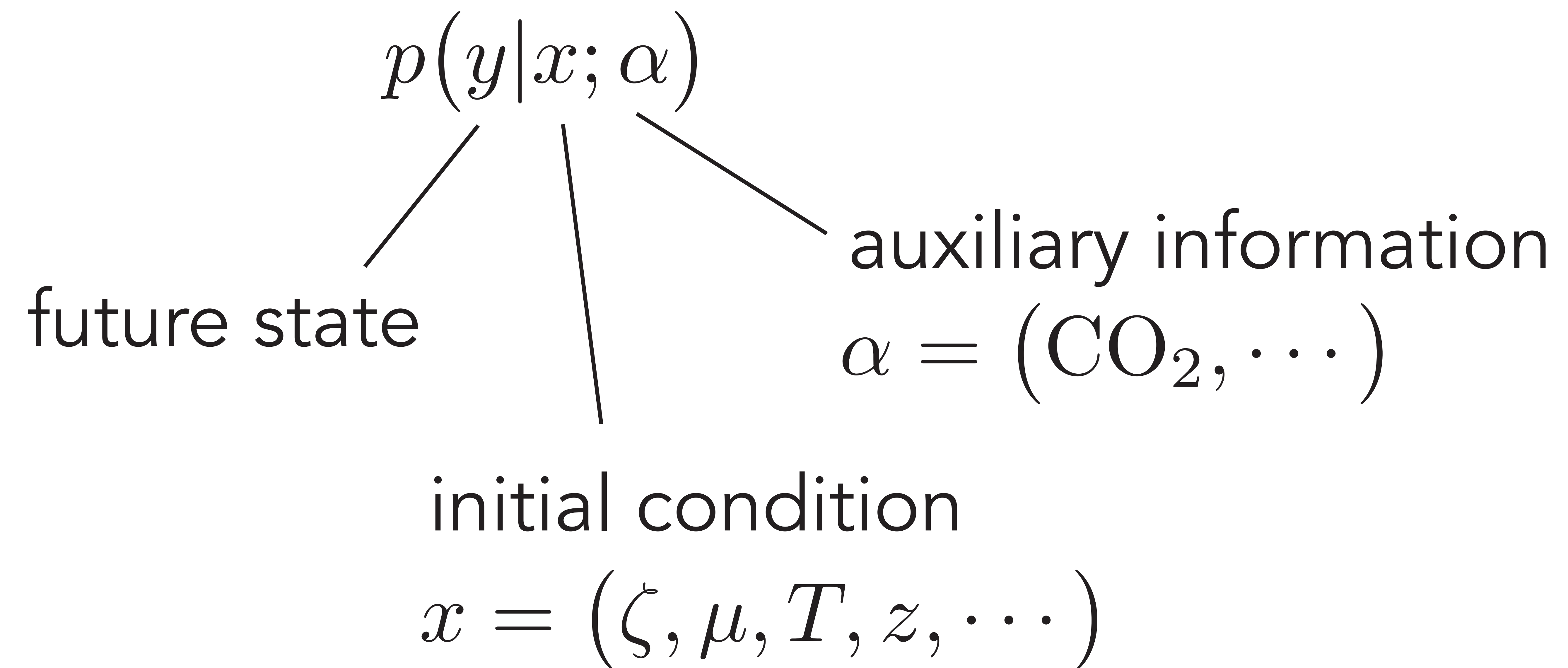
Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

Introduction

- Atmosphere as stochastic dynamical system:



Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- › Forecasting, downscaling, interpolation: $p(y|x; \alpha)$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- › Forecasting, downscaling, interpolation: $p(y|x; \alpha)$
- › Counterfactuals: $p(y|x; \hat{\alpha})$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- › Forecasting, downscaling, interpolation: $p(y|x; \alpha)$

- › Counterfactuals: $p(y|x; \hat{\alpha})$

- › Climate: $p_{\alpha}(y) = \int p(y|x; \alpha) p(x) dx$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha) \quad \text{highly complex, in-} \\ \text{stationary distribution}$$

- › Forecasting, downscaling, interpolation: $p(y|x; \alpha)$

- › Counterfactuals: $p(y|x; \hat{\alpha})$

- › Climate: $p_{\alpha}(y) = \int p(y|x; \alpha) p(x) dx$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- Numerical statistical atmospheric model:

$$\tilde{p}(\tilde{y}|\tilde{x}; \tilde{\alpha})$$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- Numerical statistical atmospheric model:

$$\tilde{p}(\tilde{y}|\tilde{x}; \tilde{\alpha}) \approx p(y|x; \alpha)$$

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- Numerical statistical atmospheric model:

$$\tilde{p}(\tilde{y}|\tilde{x}; \tilde{\alpha}) \approx p(y|x; \alpha)$$

approx. future state

approx. initial condition

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- Numerical statistical atmospheric model:

$$\tilde{p}(\tilde{y}|\tilde{x}; \tilde{\alpha}) \approx p(y|x; \alpha)$$

highly complex,
instationary
distribution

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

- Numerical statistical atmospheric model:

most complex
model available:
very large neural
network

$$\tilde{p}(\tilde{y}|\tilde{x}; \tilde{\alpha}) \approx p(y|x; \alpha)$$

highly complex,
instationary
distribution

Introduction

- Atmosphere as stochastic dynamical system:

$$p(y|x; \alpha)$$

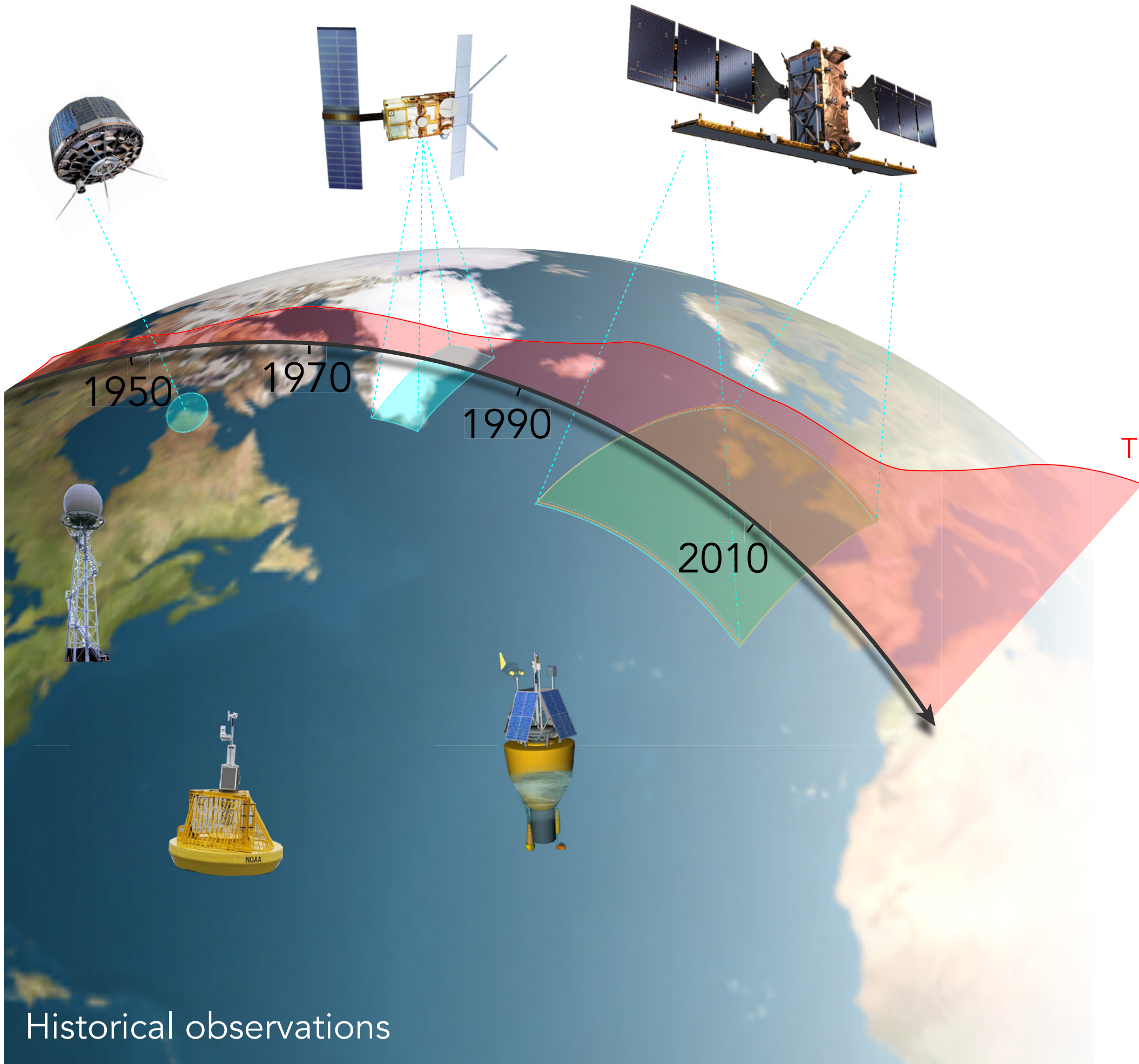
- Numerical statistical atmospheric model:

most complex
model available:
very large neural
network

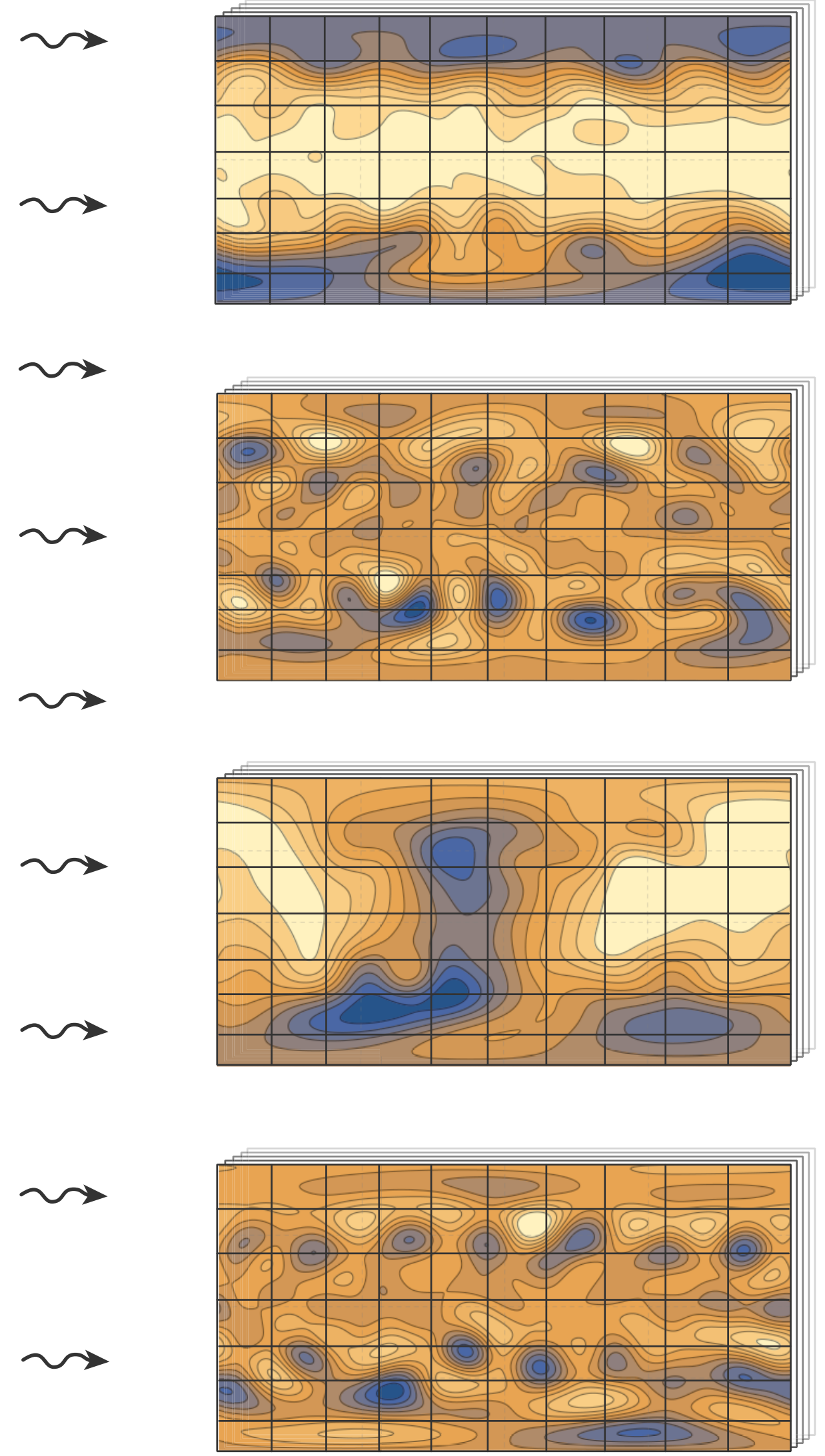
$$\tilde{p}_\theta(\tilde{y}|\tilde{x}; \tilde{\alpha}) \approx p(y|x; \alpha)$$

highly complex,
instationary
distribution

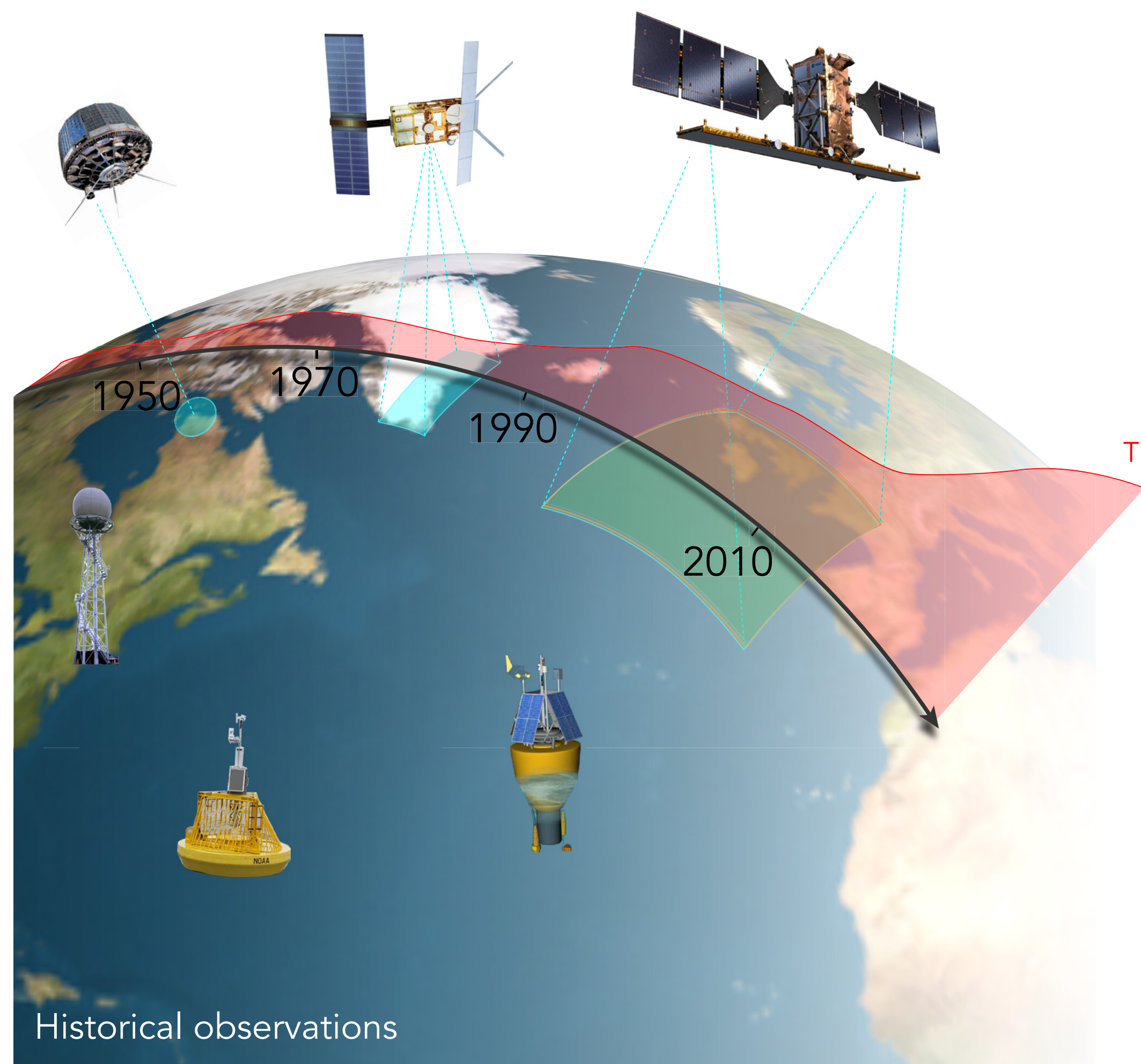
AtmoRep



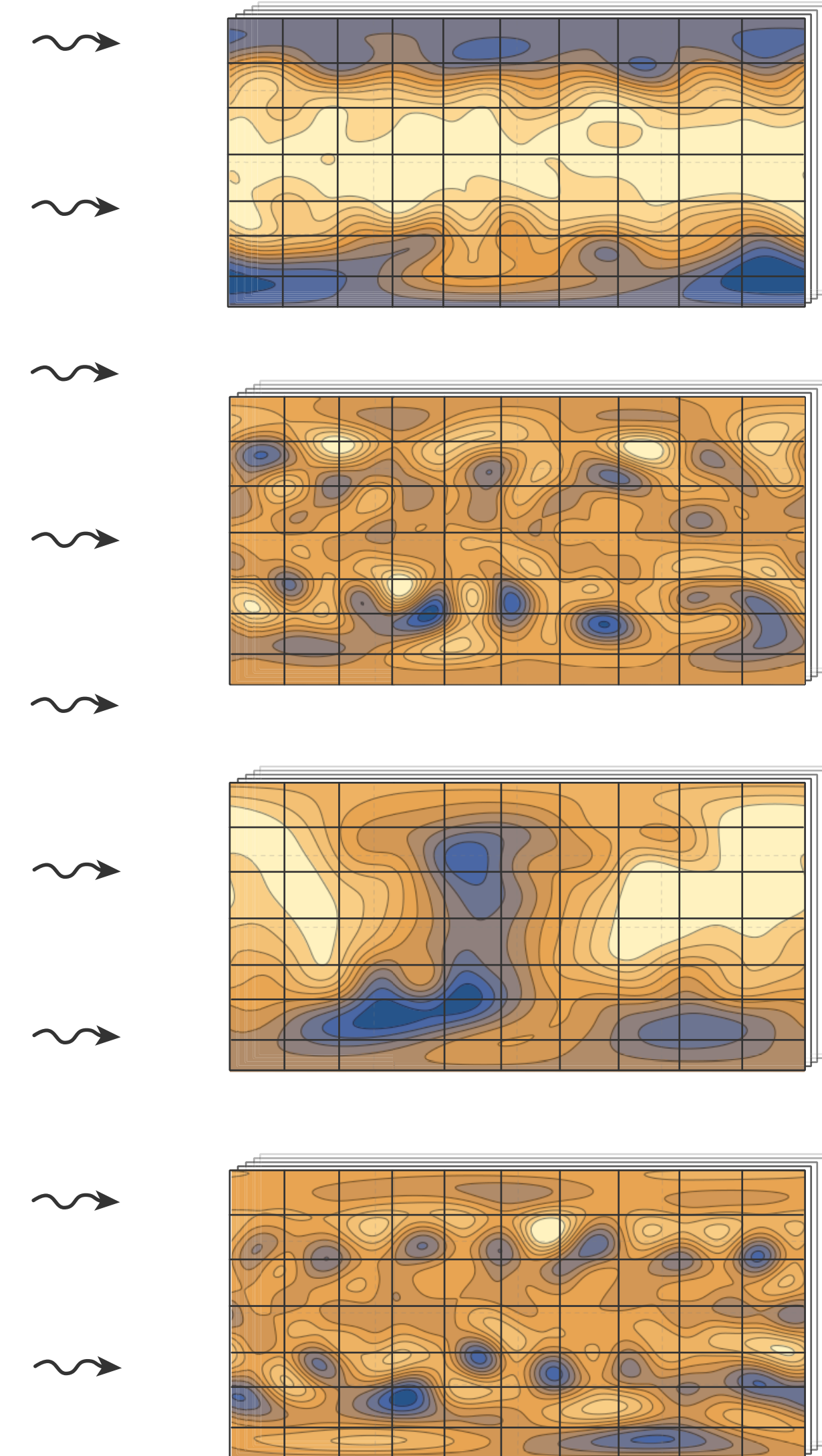
ERA5 reanalysis



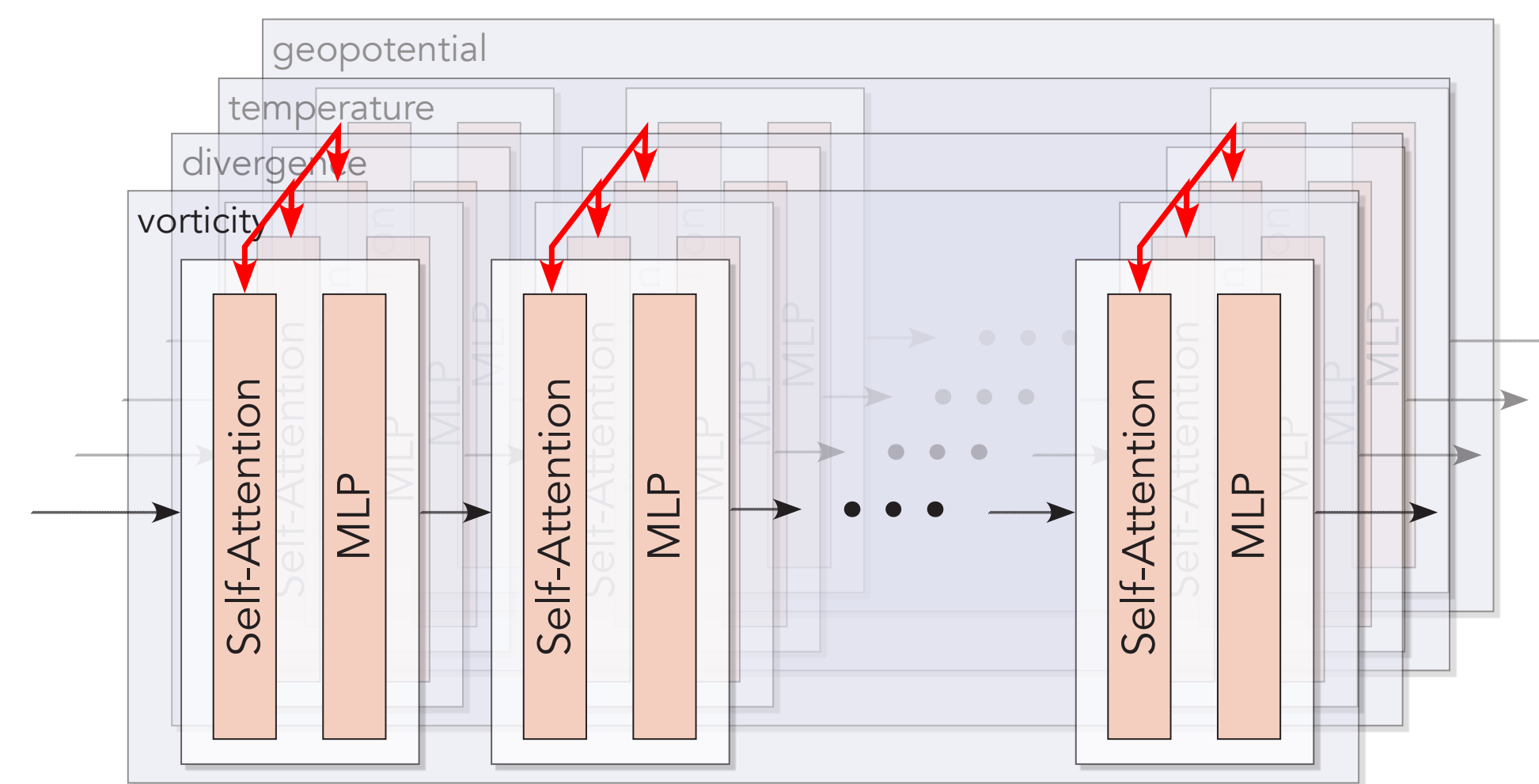
AtmoRep



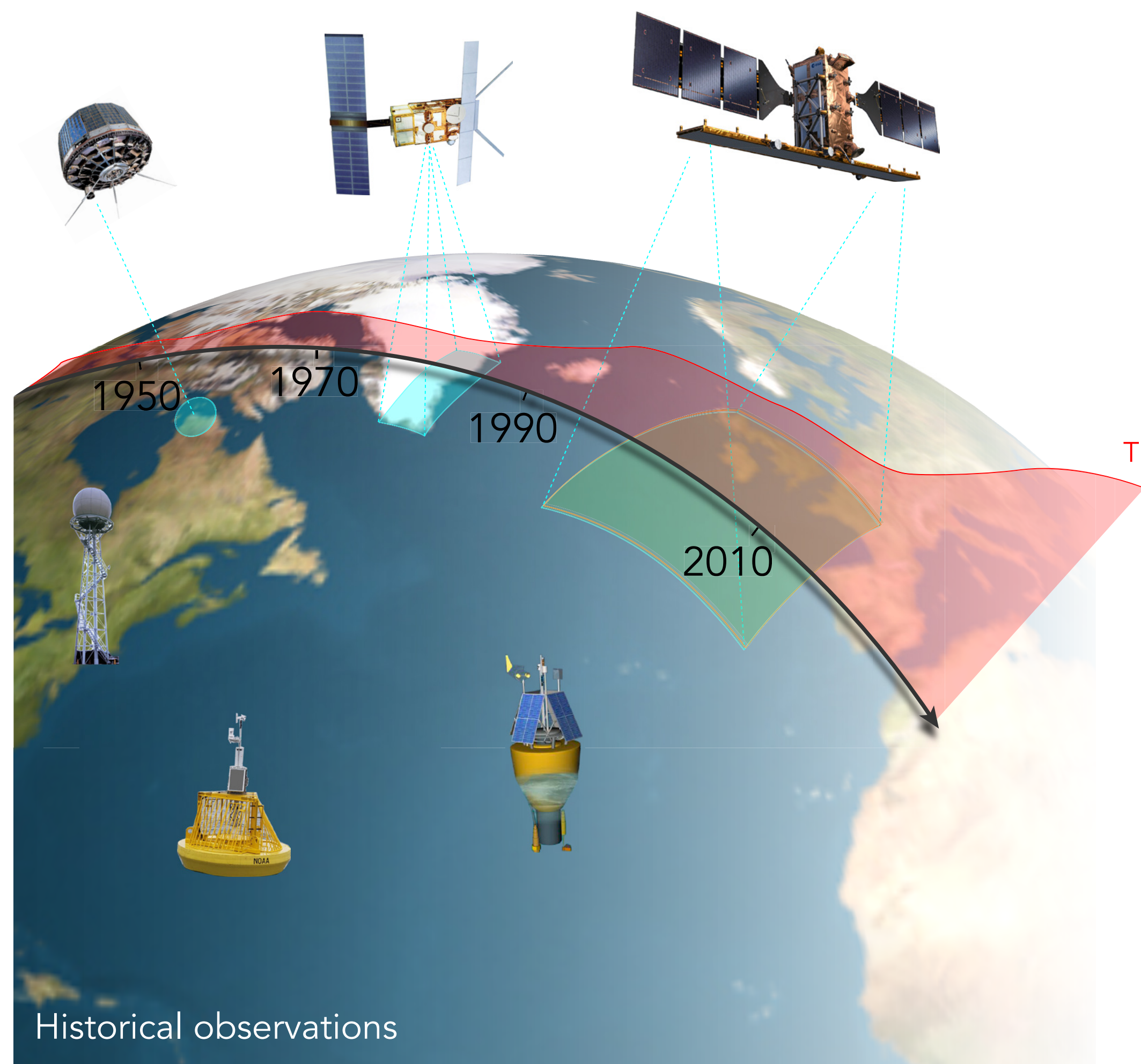
ERA5 reanalysis



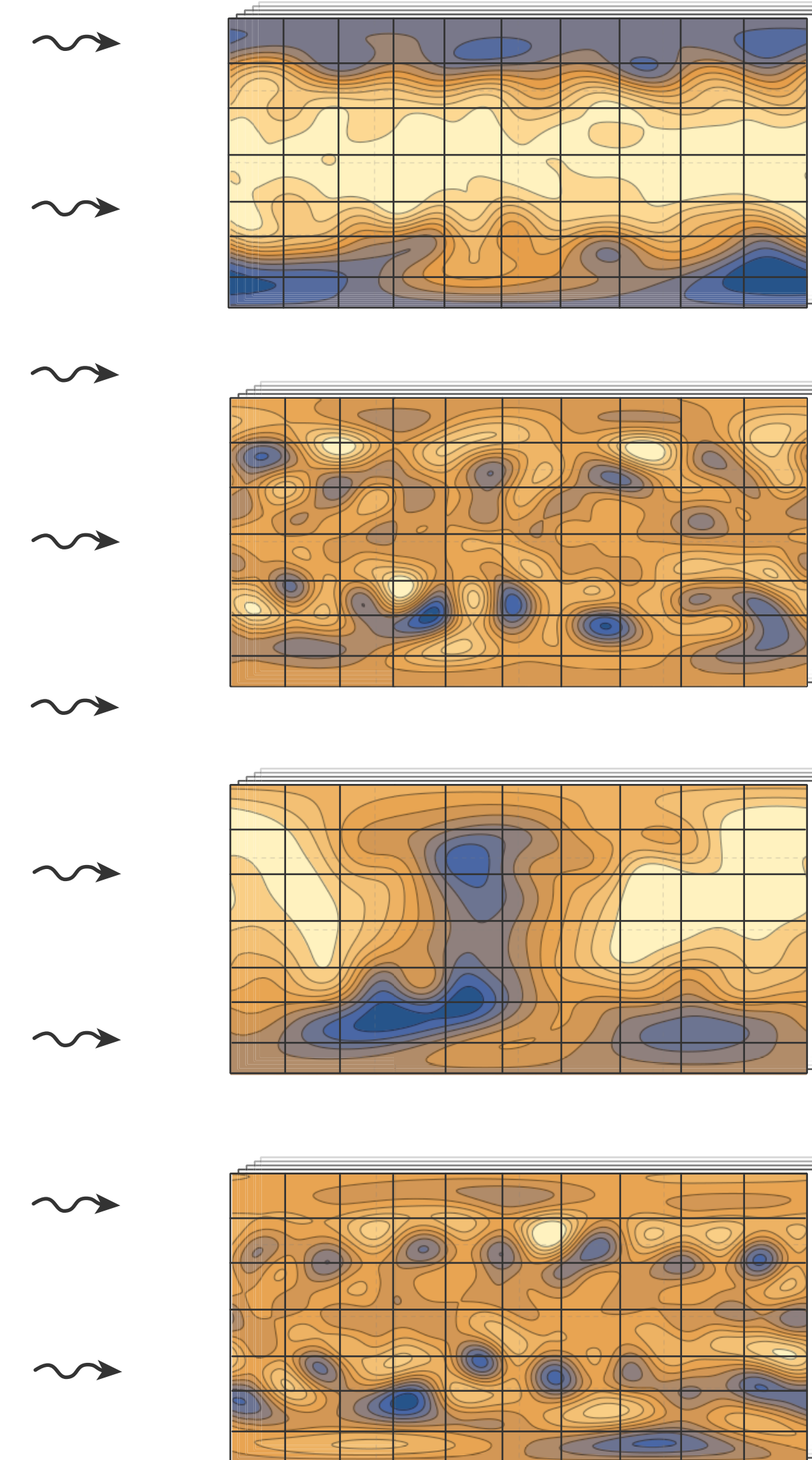
$$p_{\theta}(y|x)$$



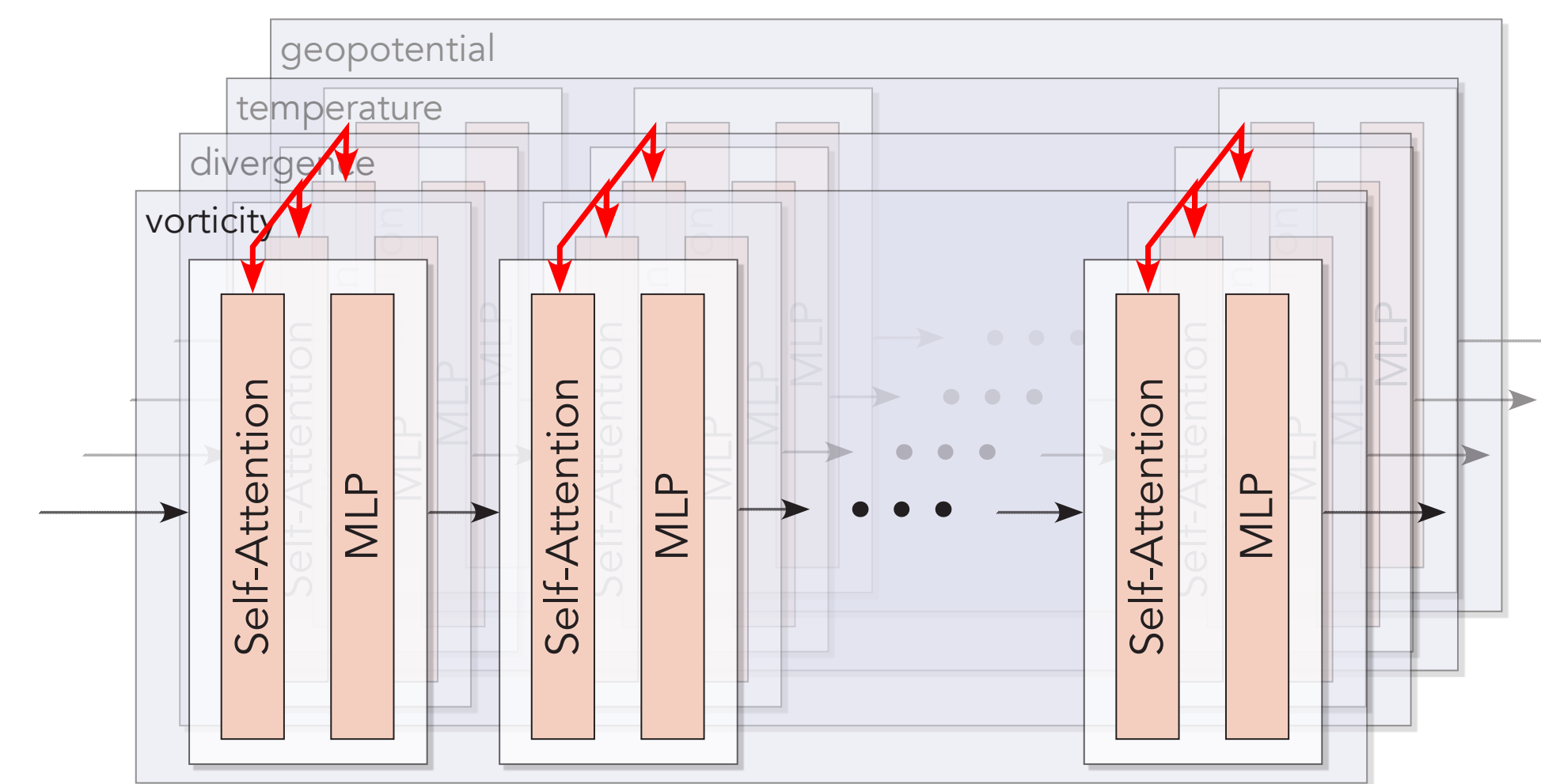
AtmoRep



ERA5 reanalysis

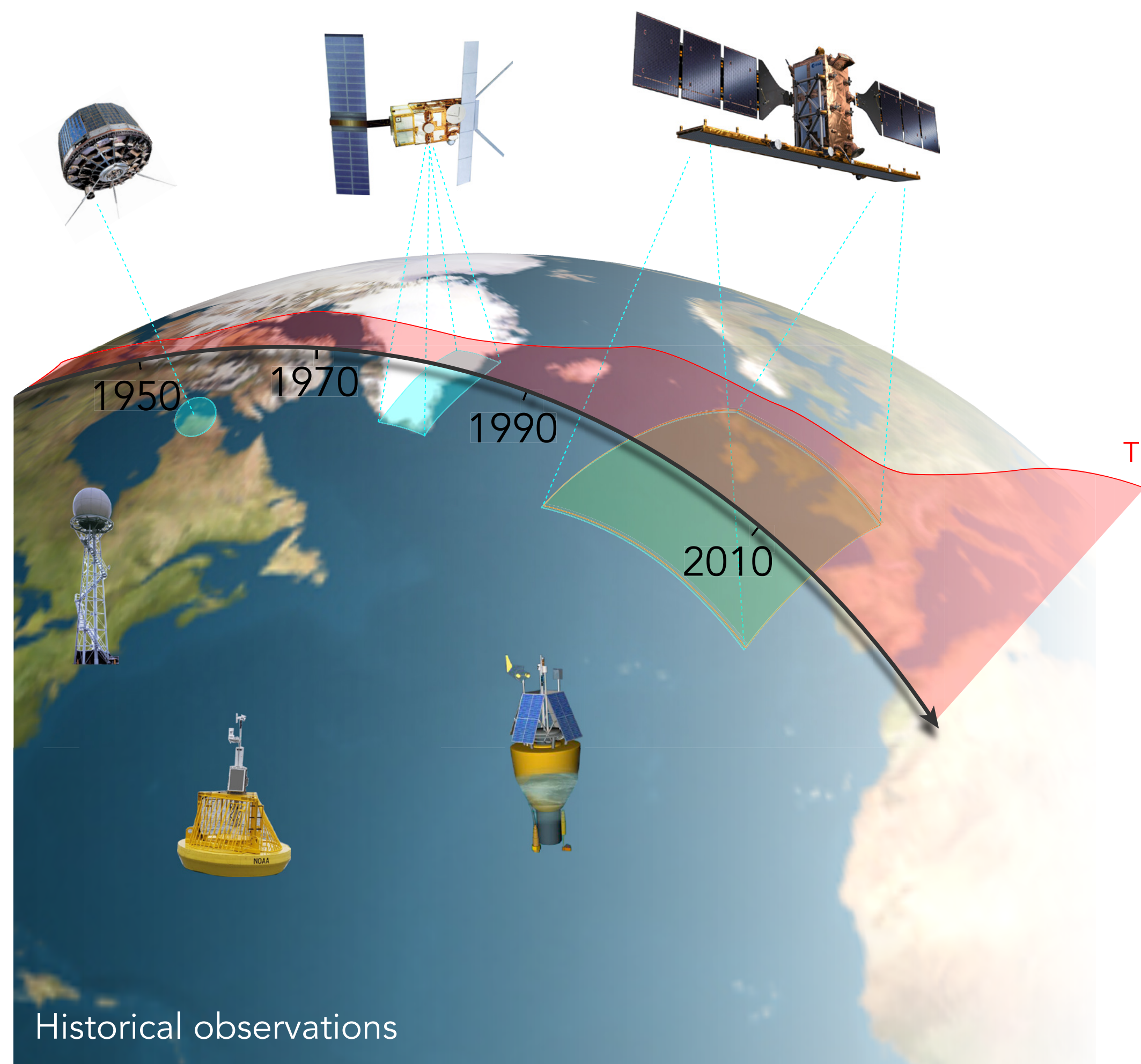


$$p_{\theta}(y|x)$$

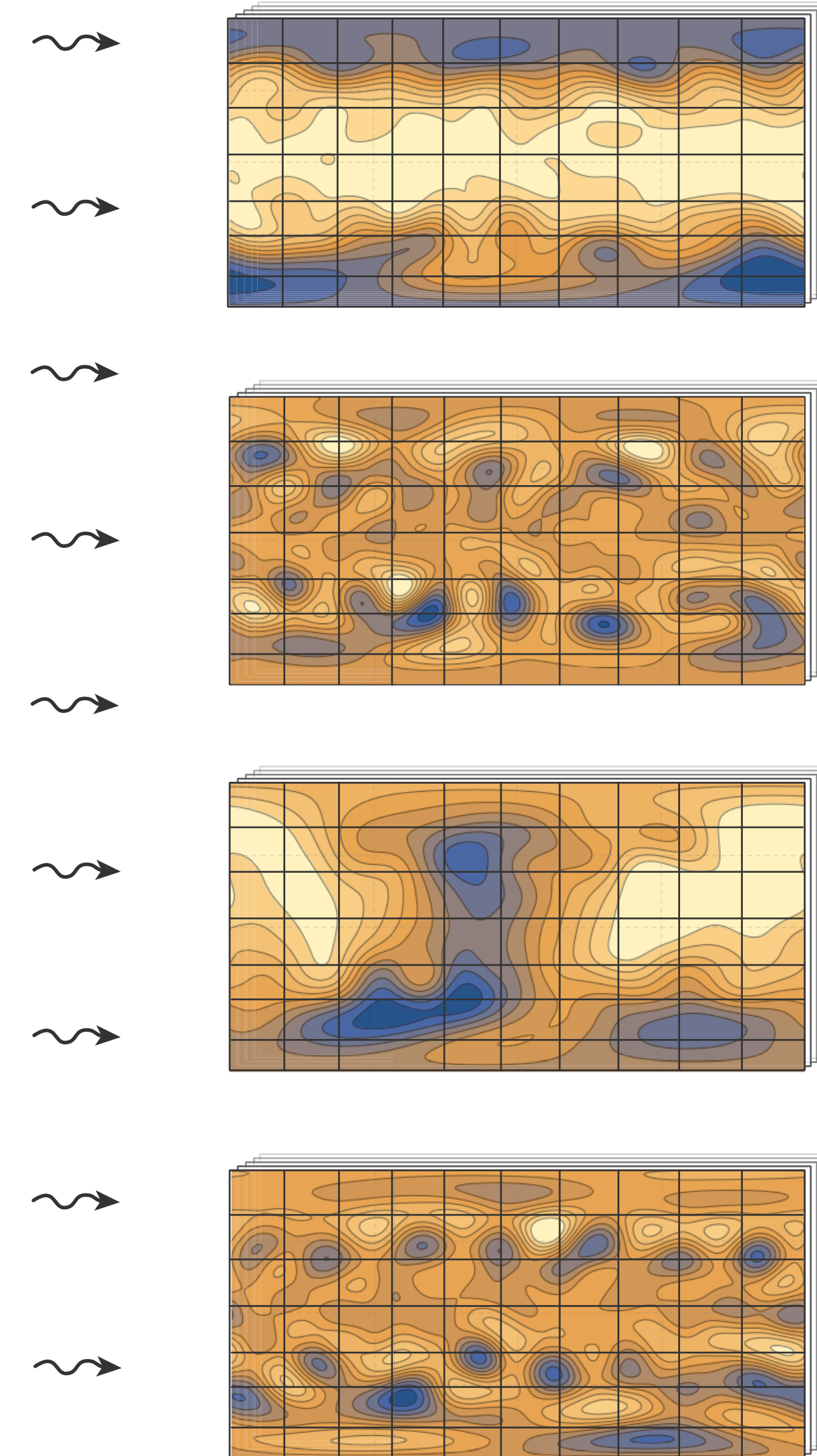


large transformer
with 4×10^9 parameters

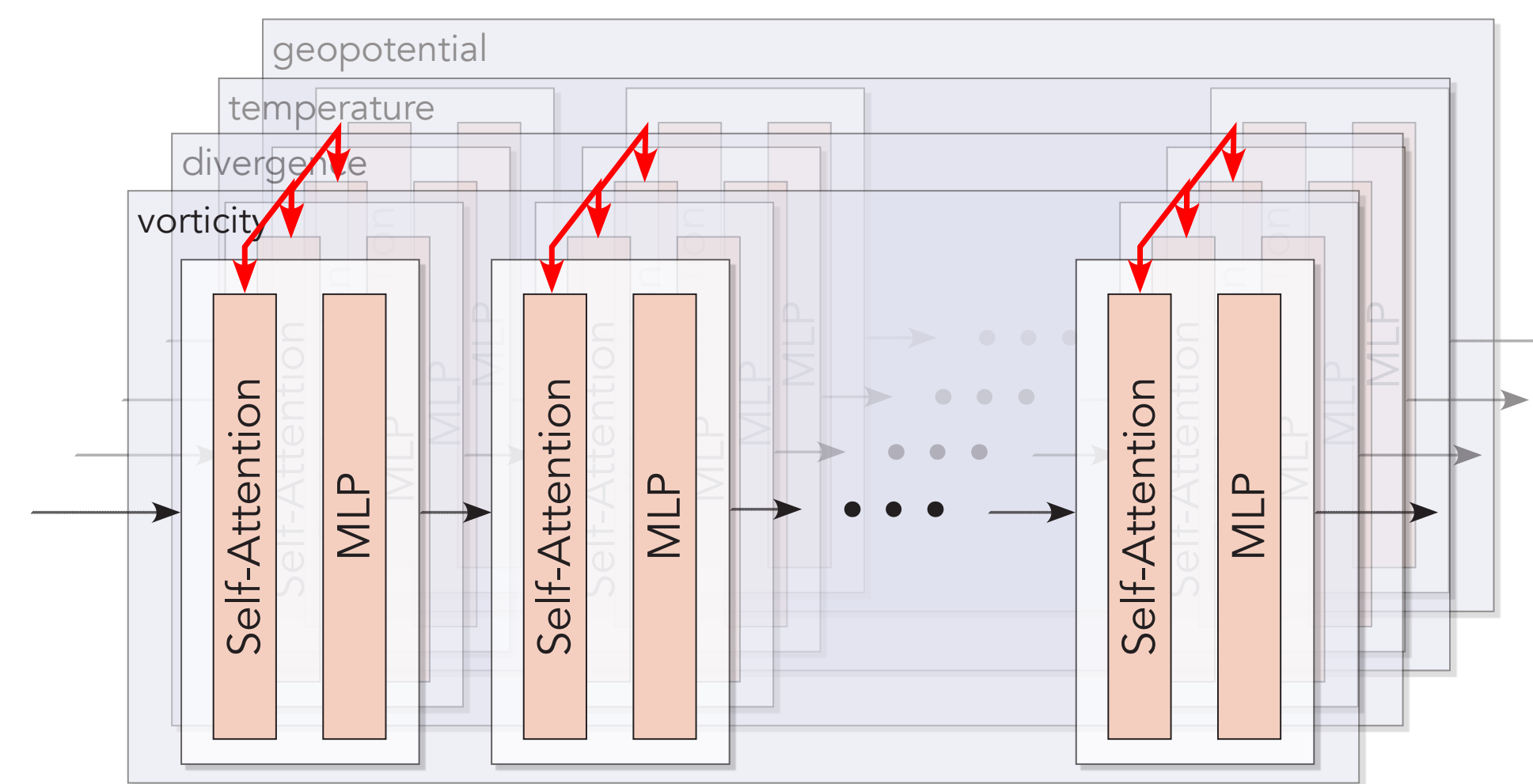
AtmoRep



ERA5 reanalysis

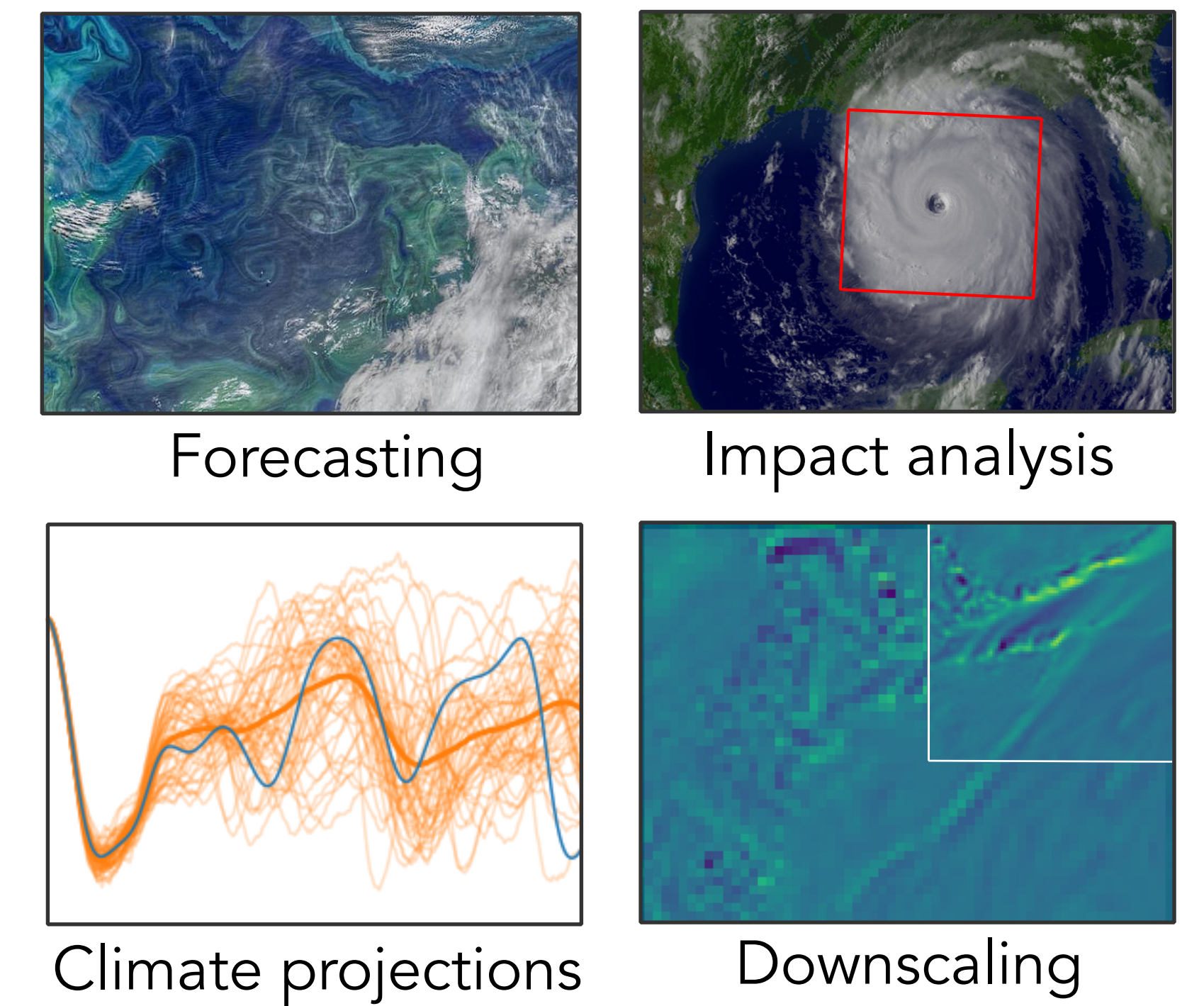


$$p_{\theta}(y|x)$$

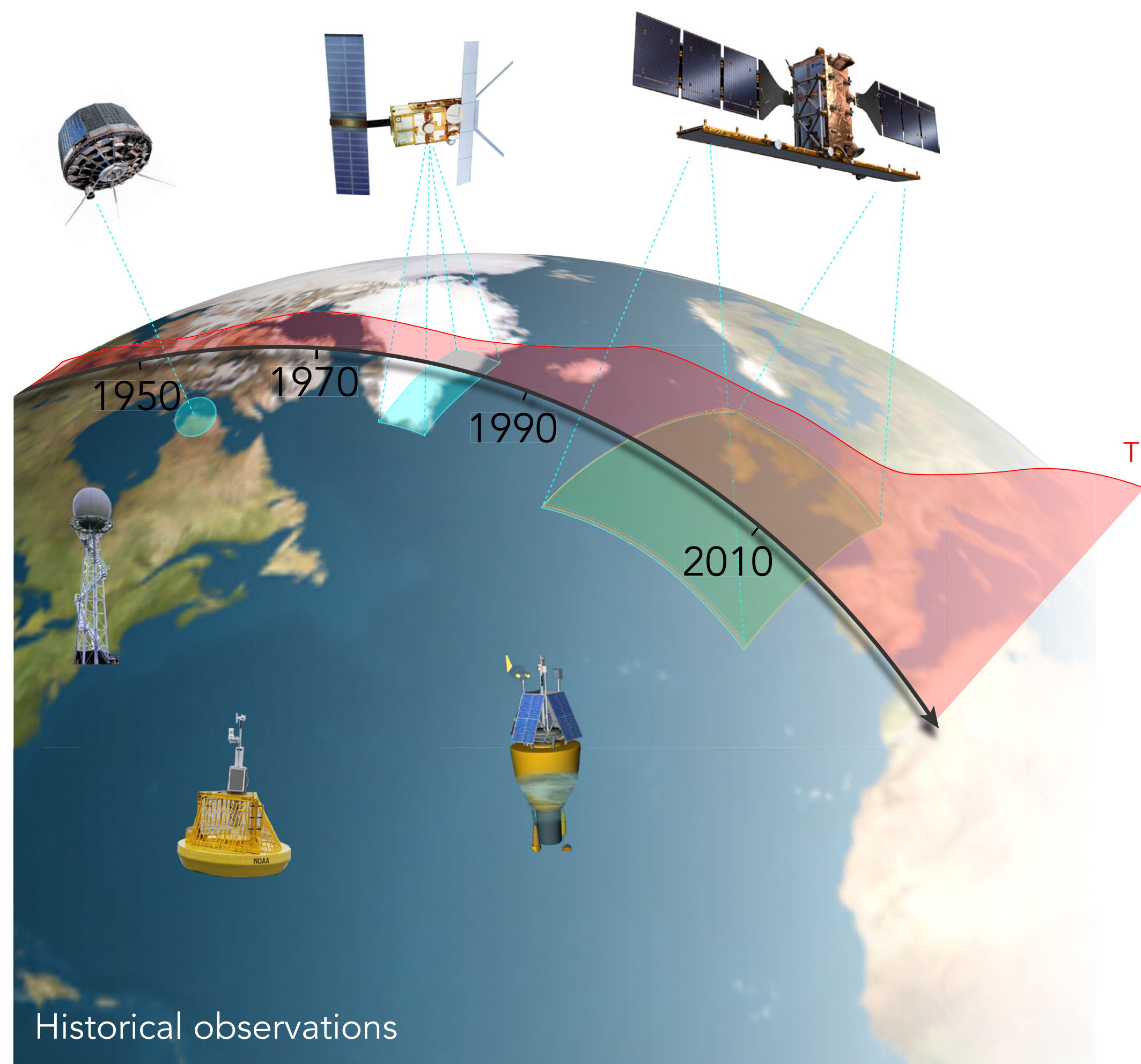


large transformer
with 4×10^9 parameters

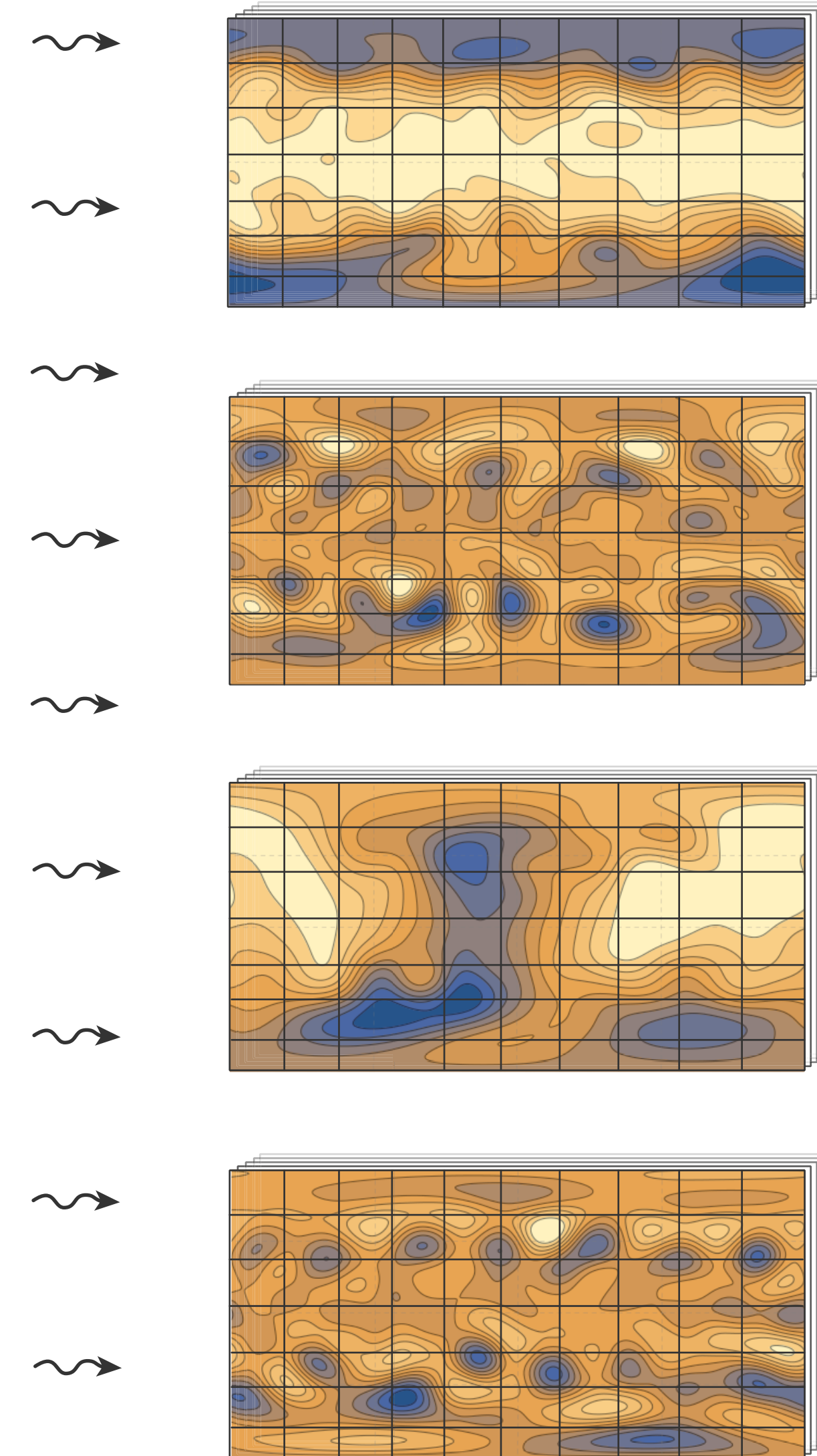
applications



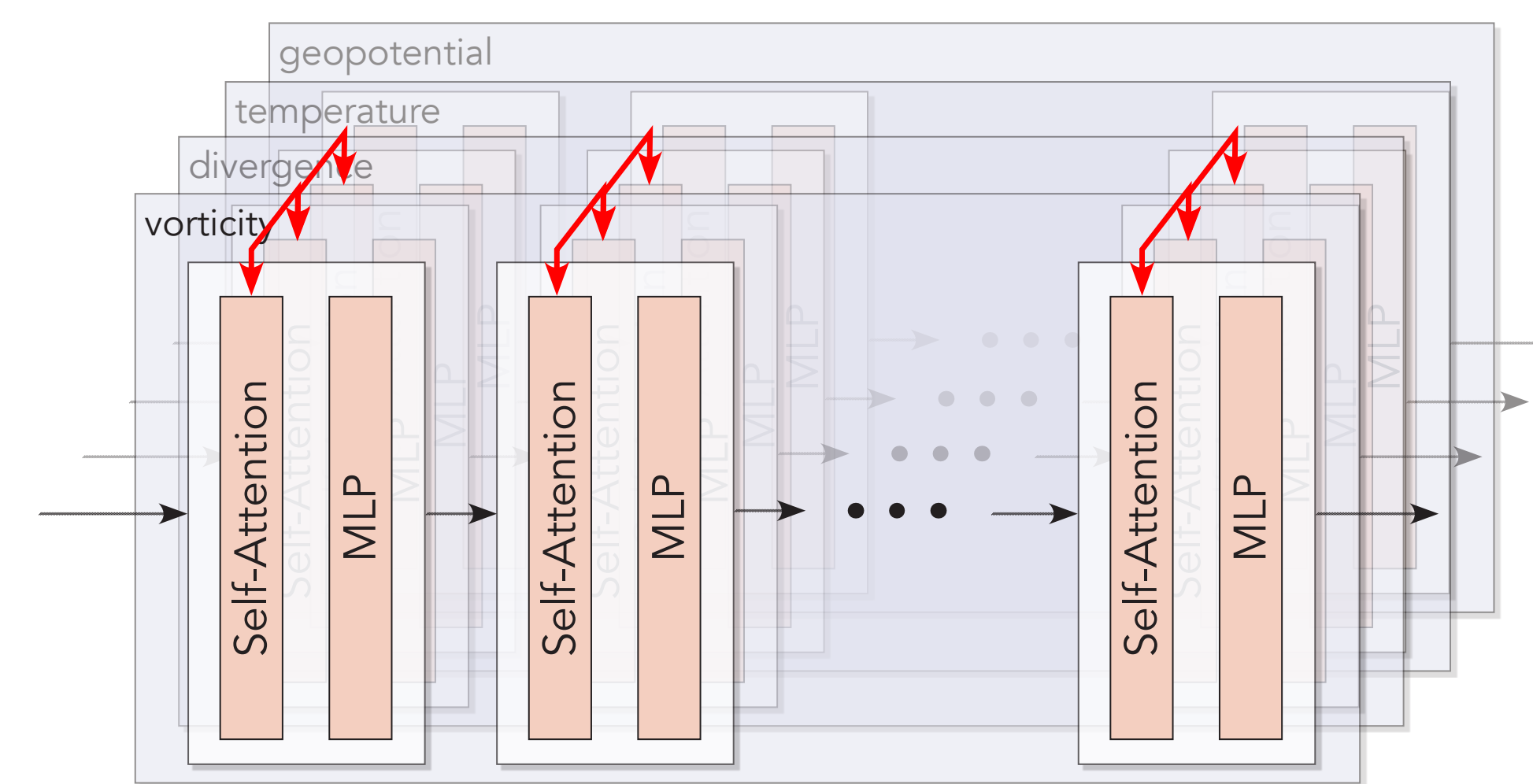
AtmoRep



ERA5 reanalysis

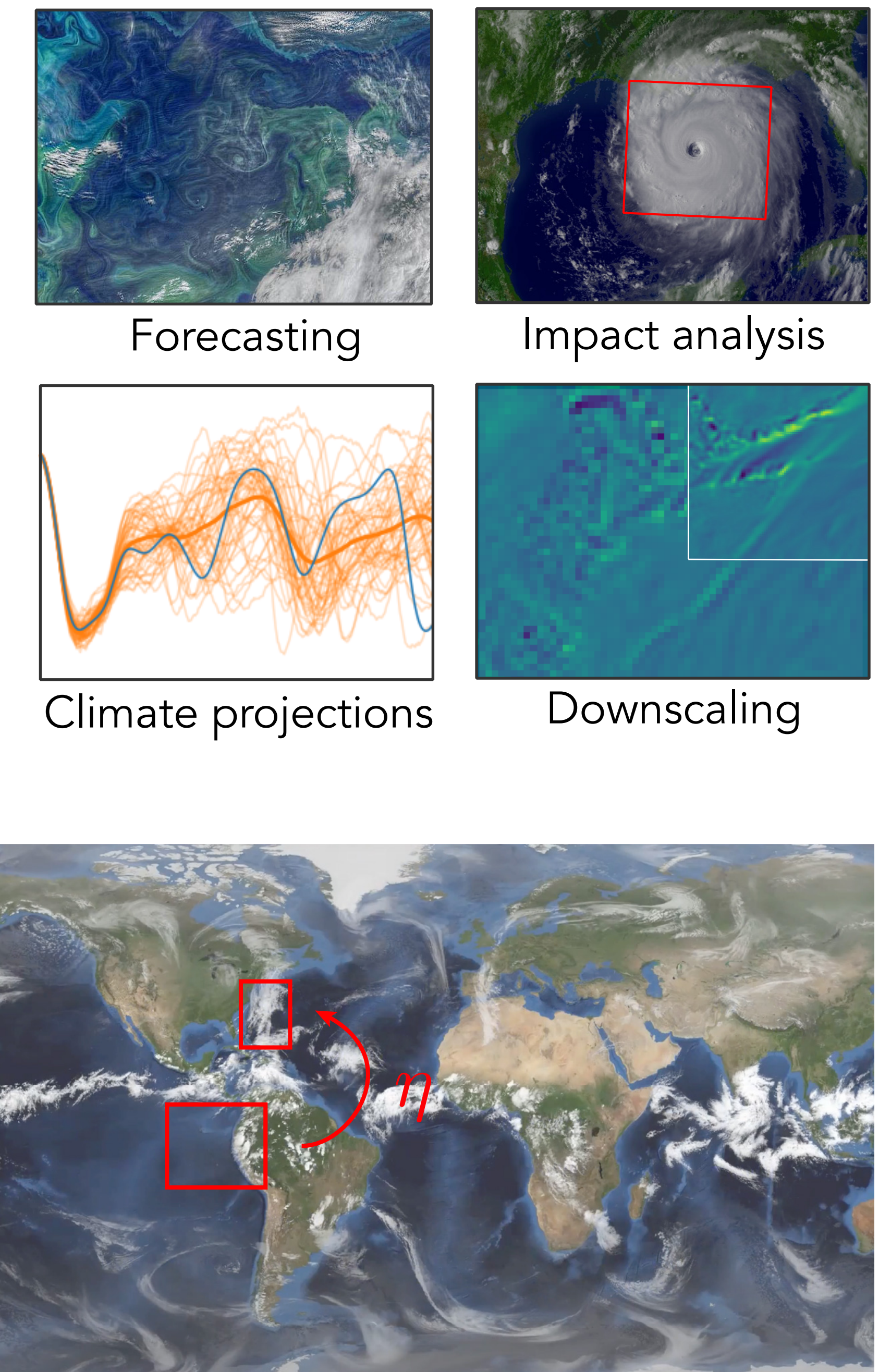


$$p_{\theta}(y|x)$$



large transformer
with 4×10^9 parameters

applications

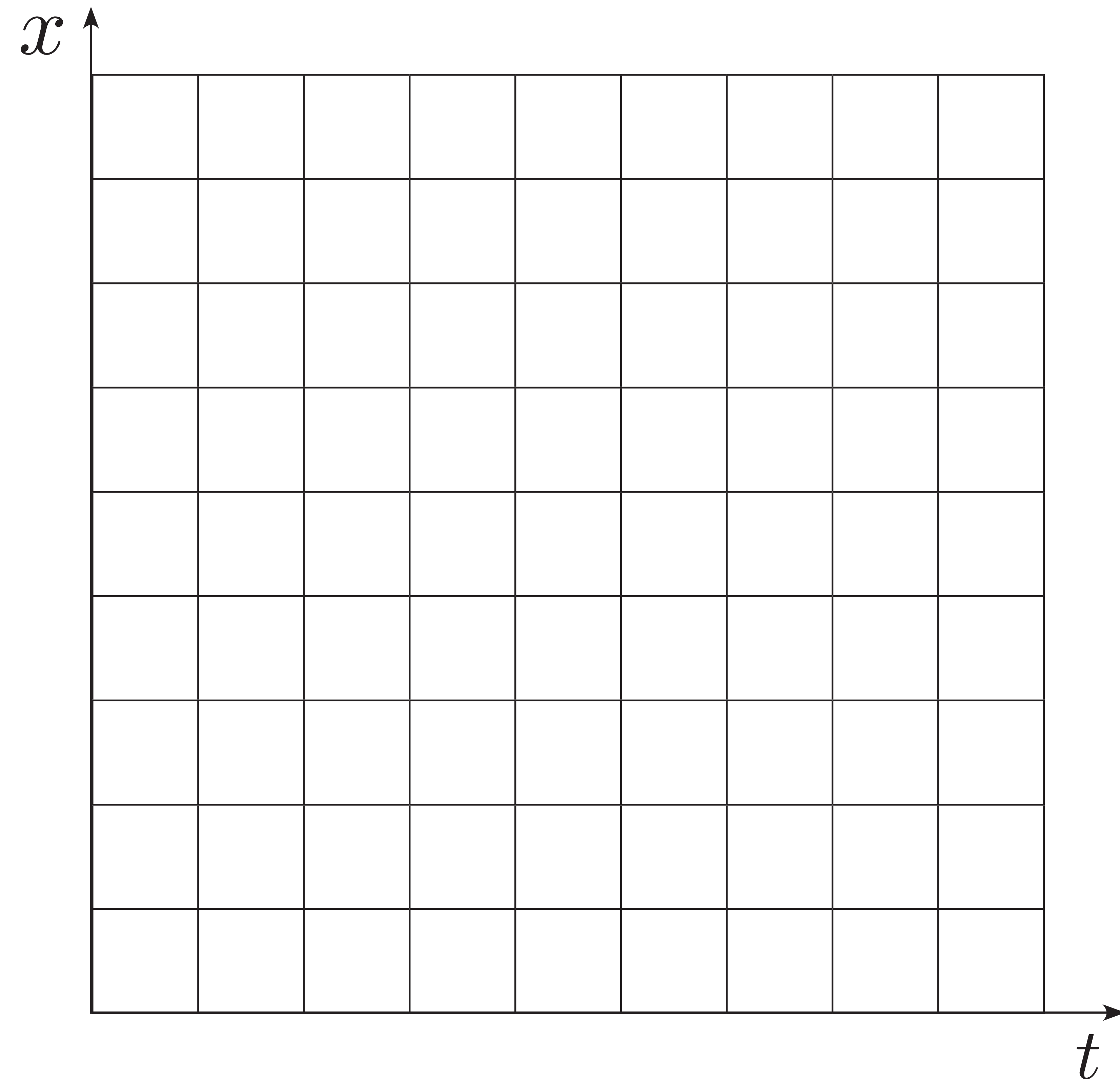


scientific insight

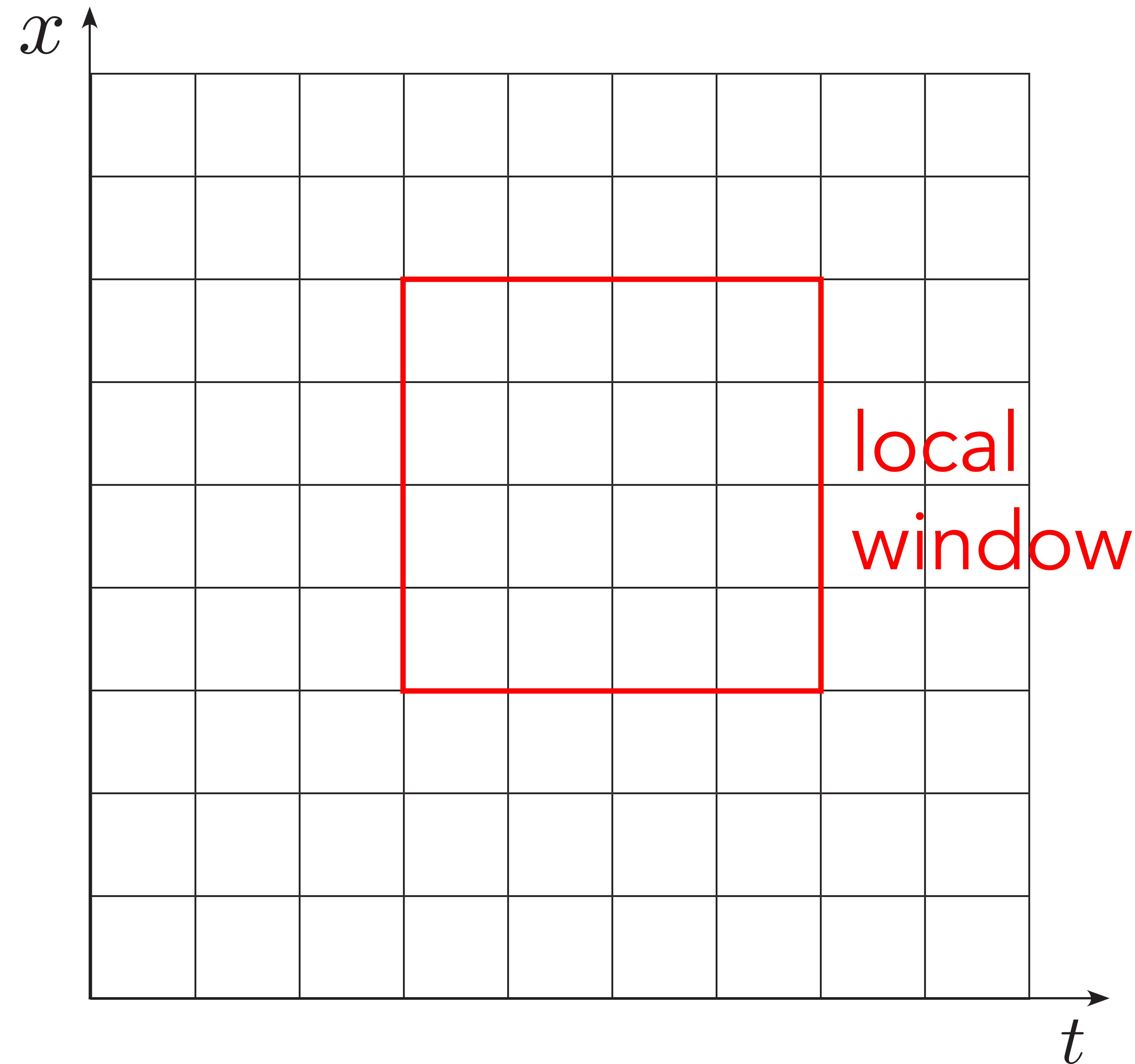
AtmoRep Training

- Statistical atmospheric model $p_{\theta}(y|x)$
 - › Training models spatio-temporal relationship between arbitrary state x and y

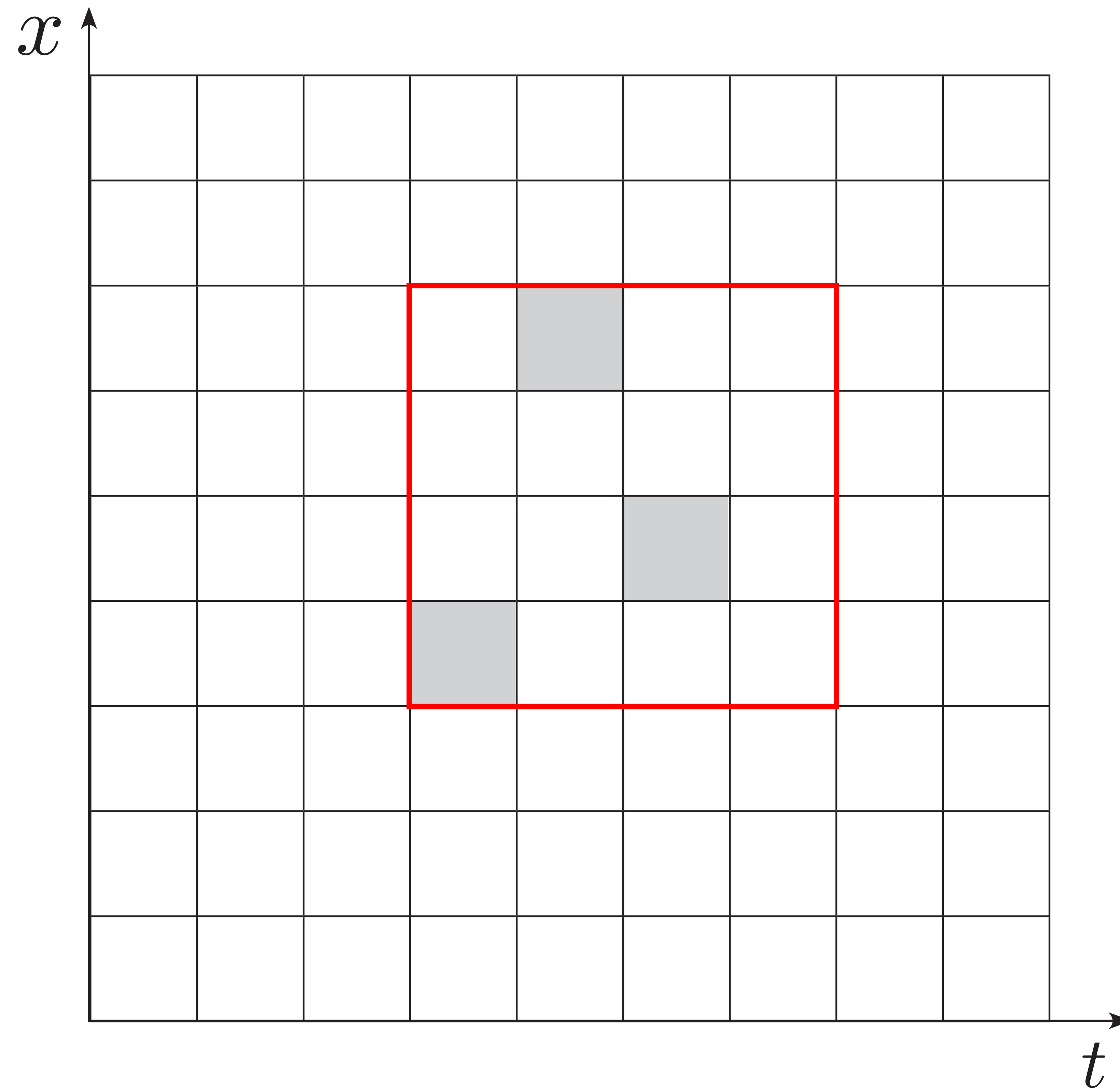
AtmoRep Training



AtmoRep Training



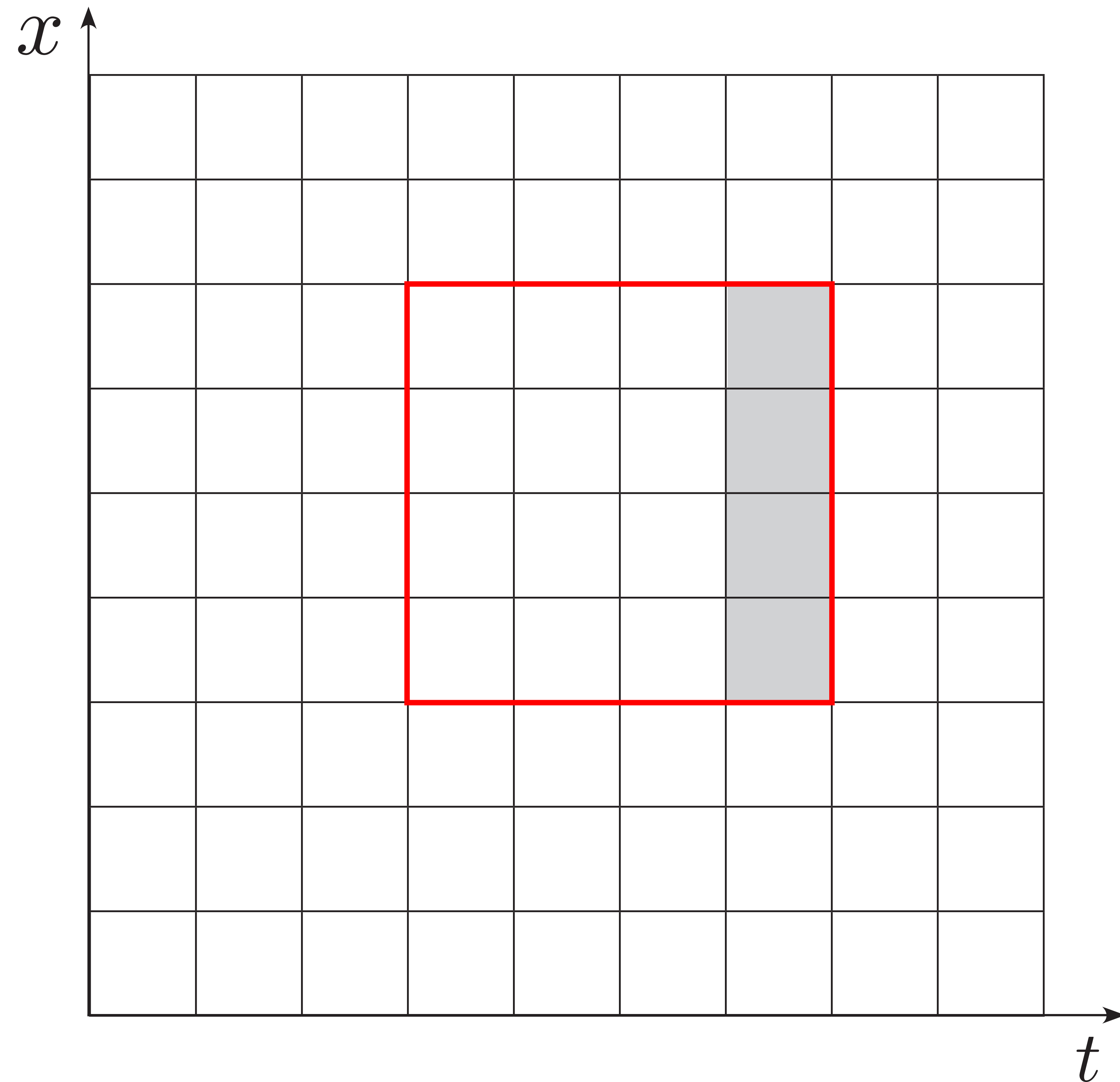
AtmoRep Training



Training task:
predict randomly
masked neighbor-
hoods in space-
time

AtmoRep Training

Forecasting

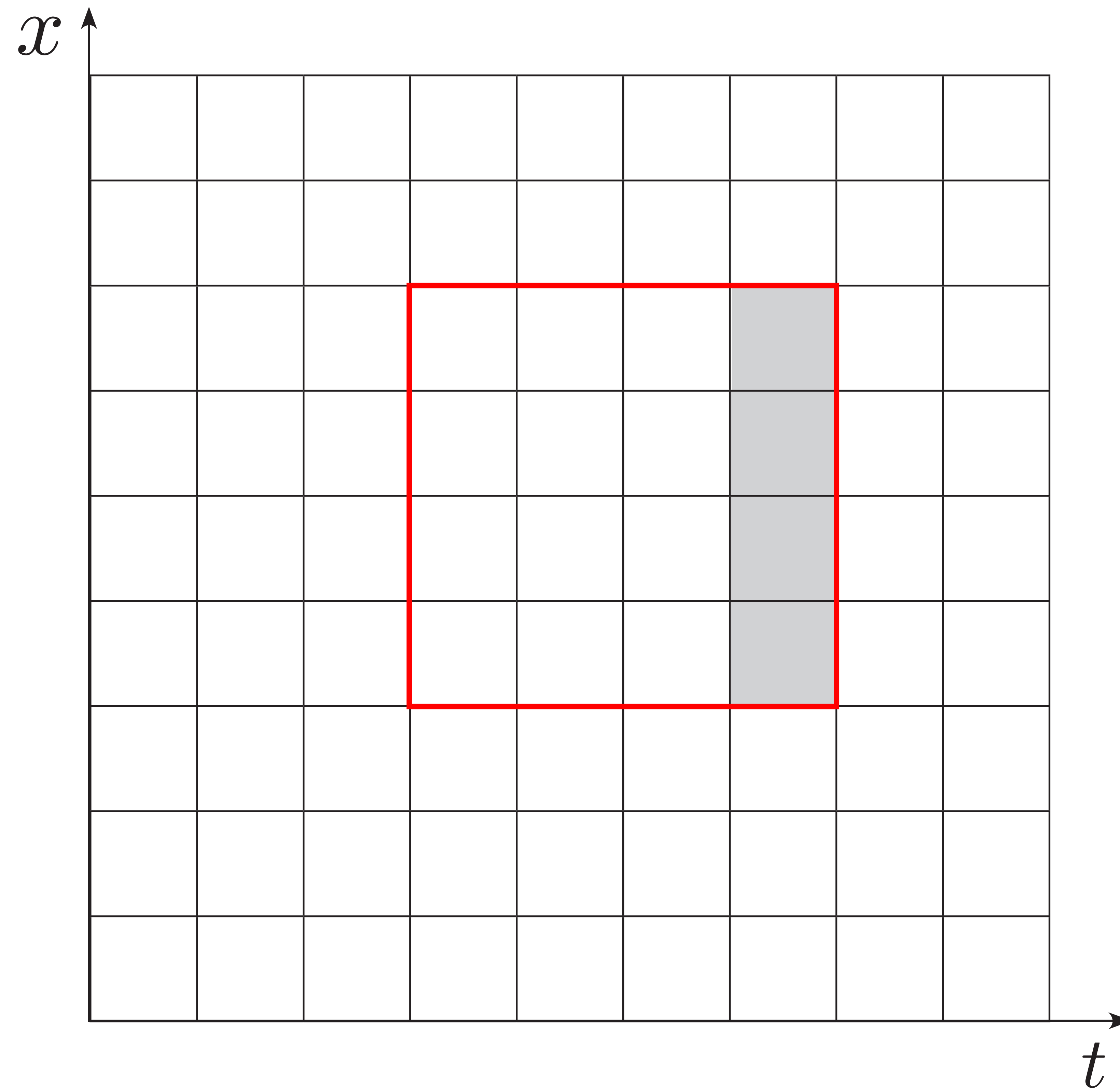


Training task:
predict randomly
masked neighbor-
hoods in space-
time

AtmoRep Training

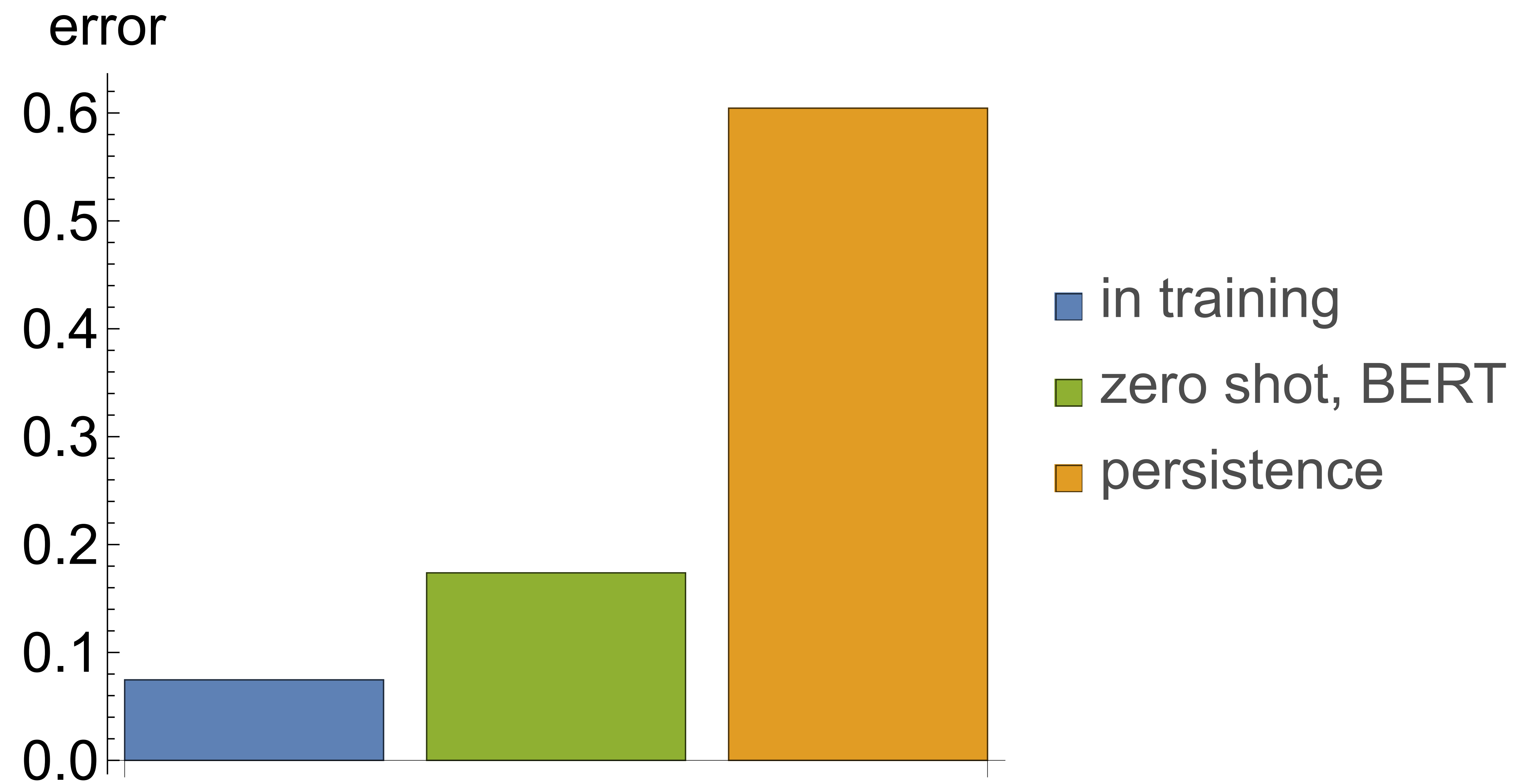
Forecasting

Task is of the form: $p_{\theta}(y|x)$



Training task:
predict randomly
masked neighbor-
hoods in space-
time

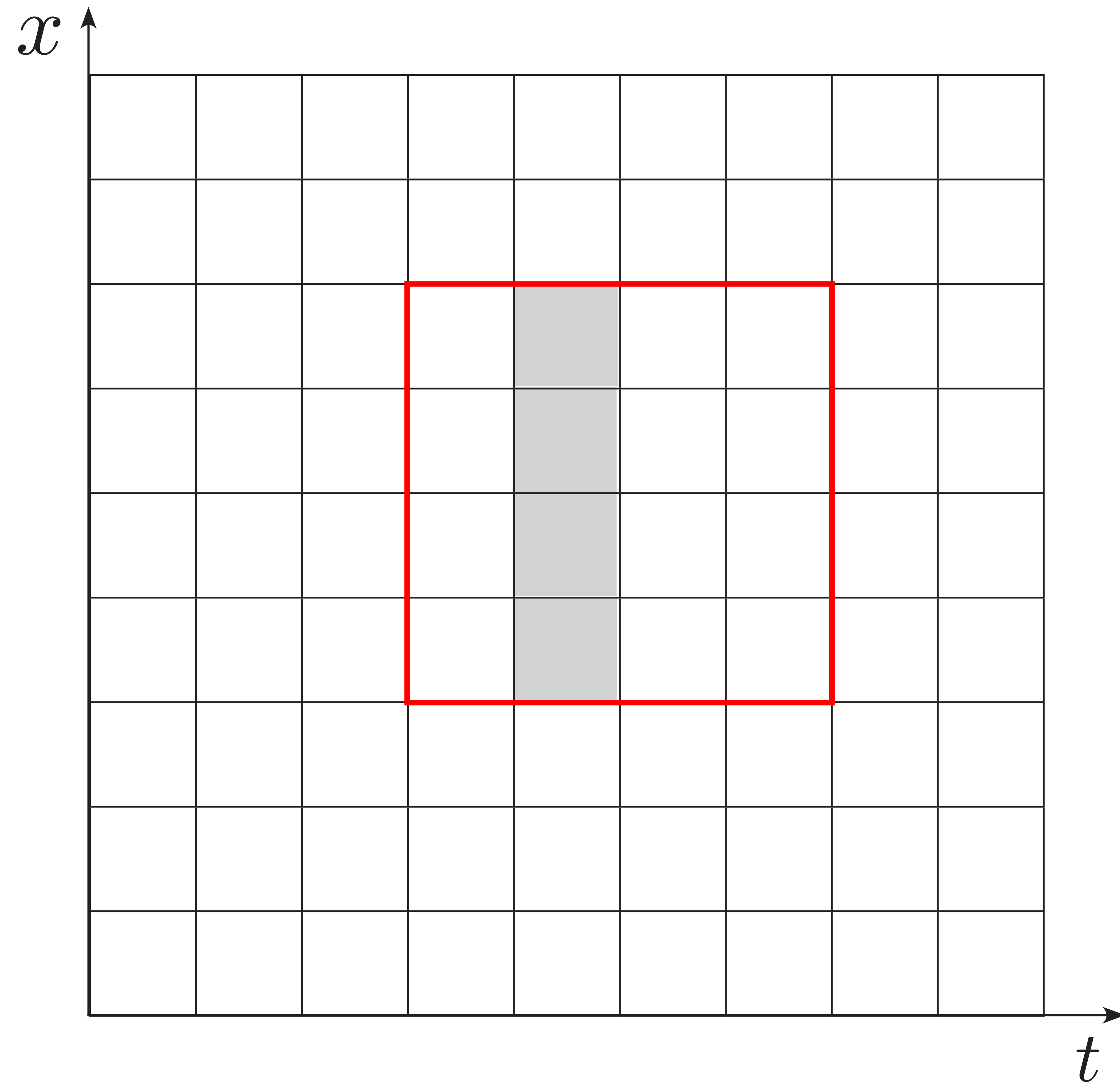
AtmoRep Training



AtmoRep Training

Temporal interpolation

Task is of the form: $p_{\theta}(y|x)$

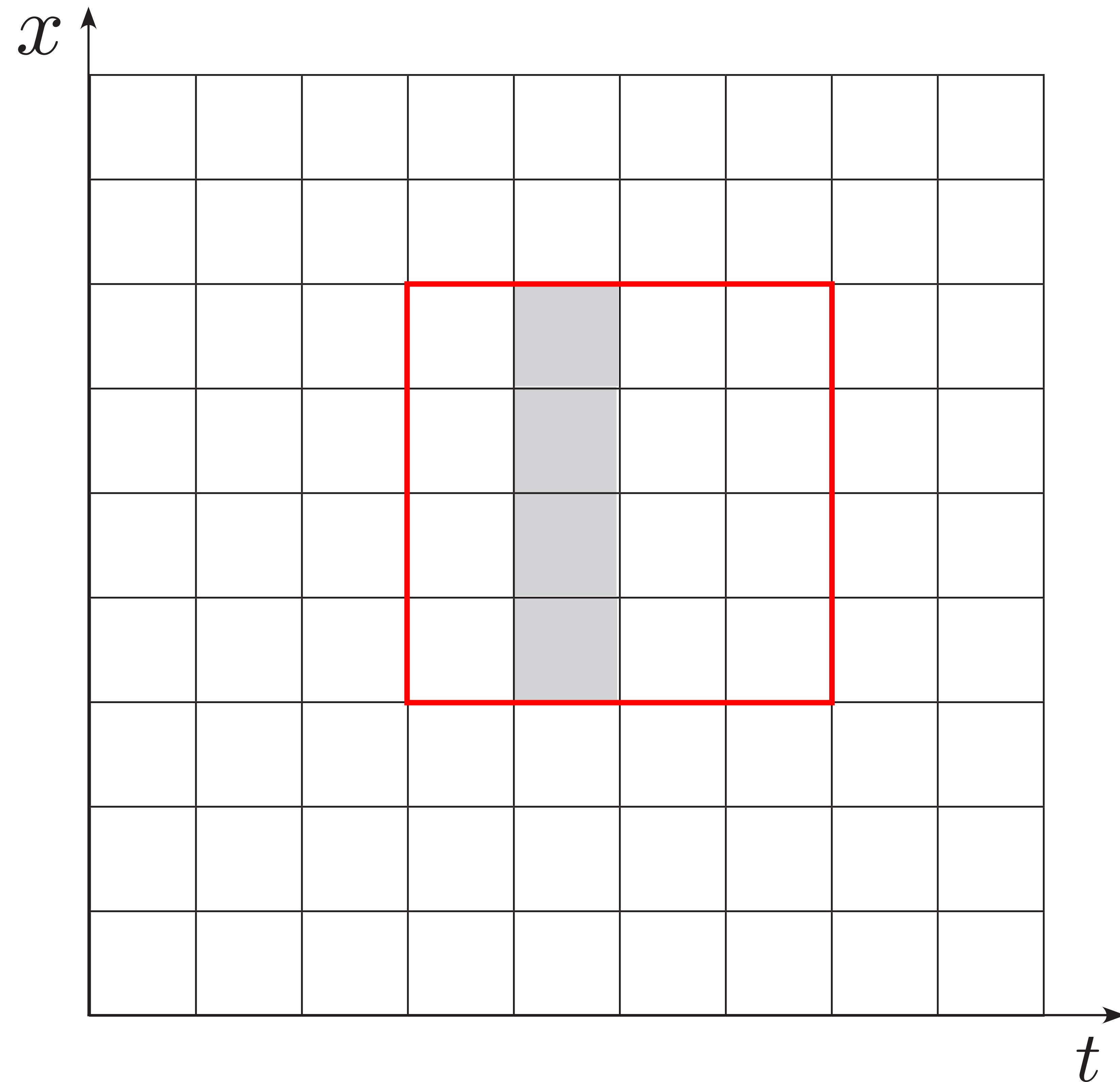


Training task:
predict randomly masked neighborhoods in space-time

AtmoRep Training

Also:

- downscaling
- interpolation in space (missing data)
- ...



Training task:

predict randomly masked neighborhoods in space-time

AtmoRep

- Numerical statistical atmospheric model
 - › Task agnostic but applicable to a wide range of applications and scientific questions

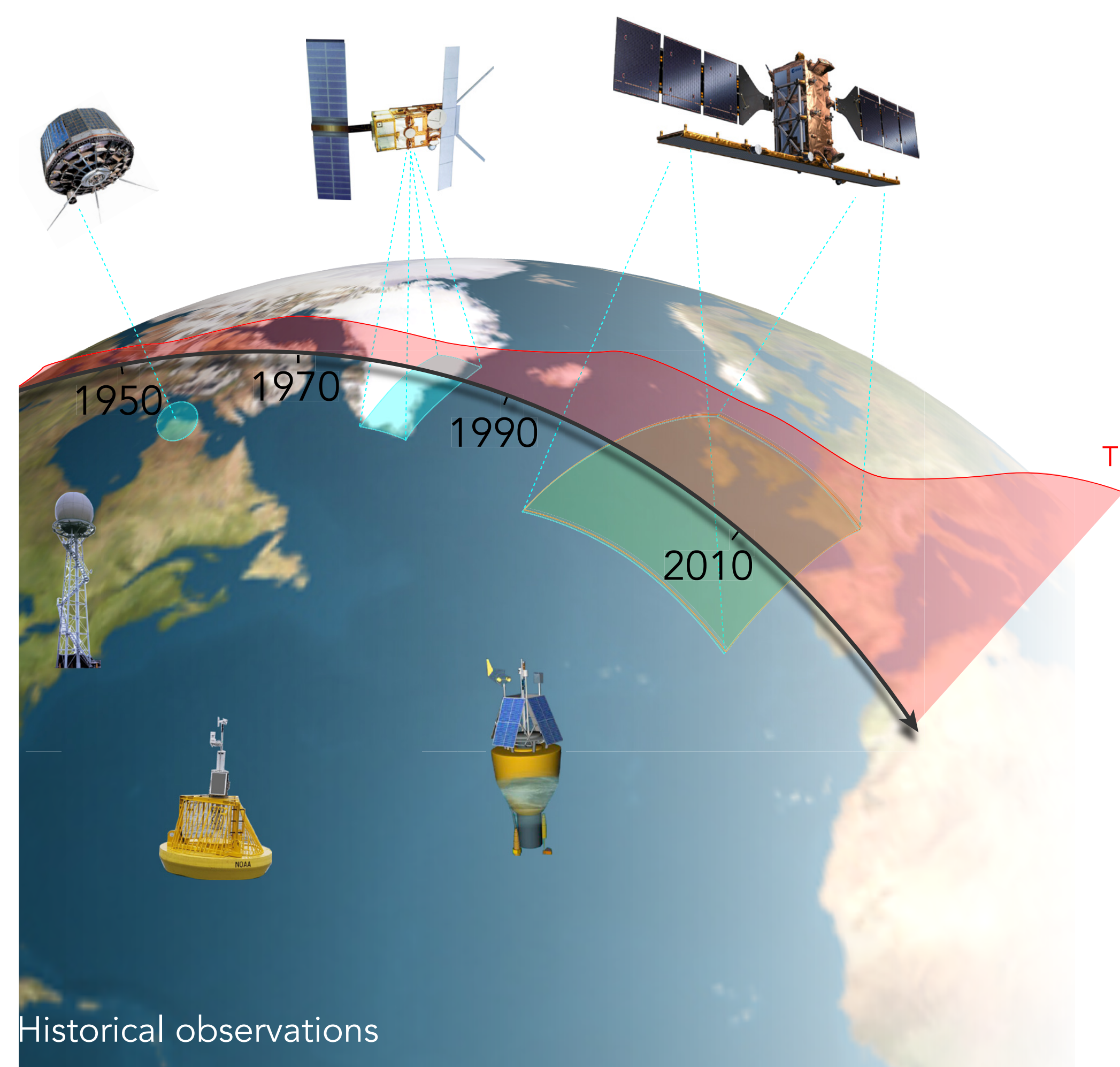
AtmoRep

- Numerical statistical atmospheric model
 - › Task agnostic but applicable to a wide range of applications and scientific questions
 - › Corresponds to large scale representation learning (foundational model) in machine learning

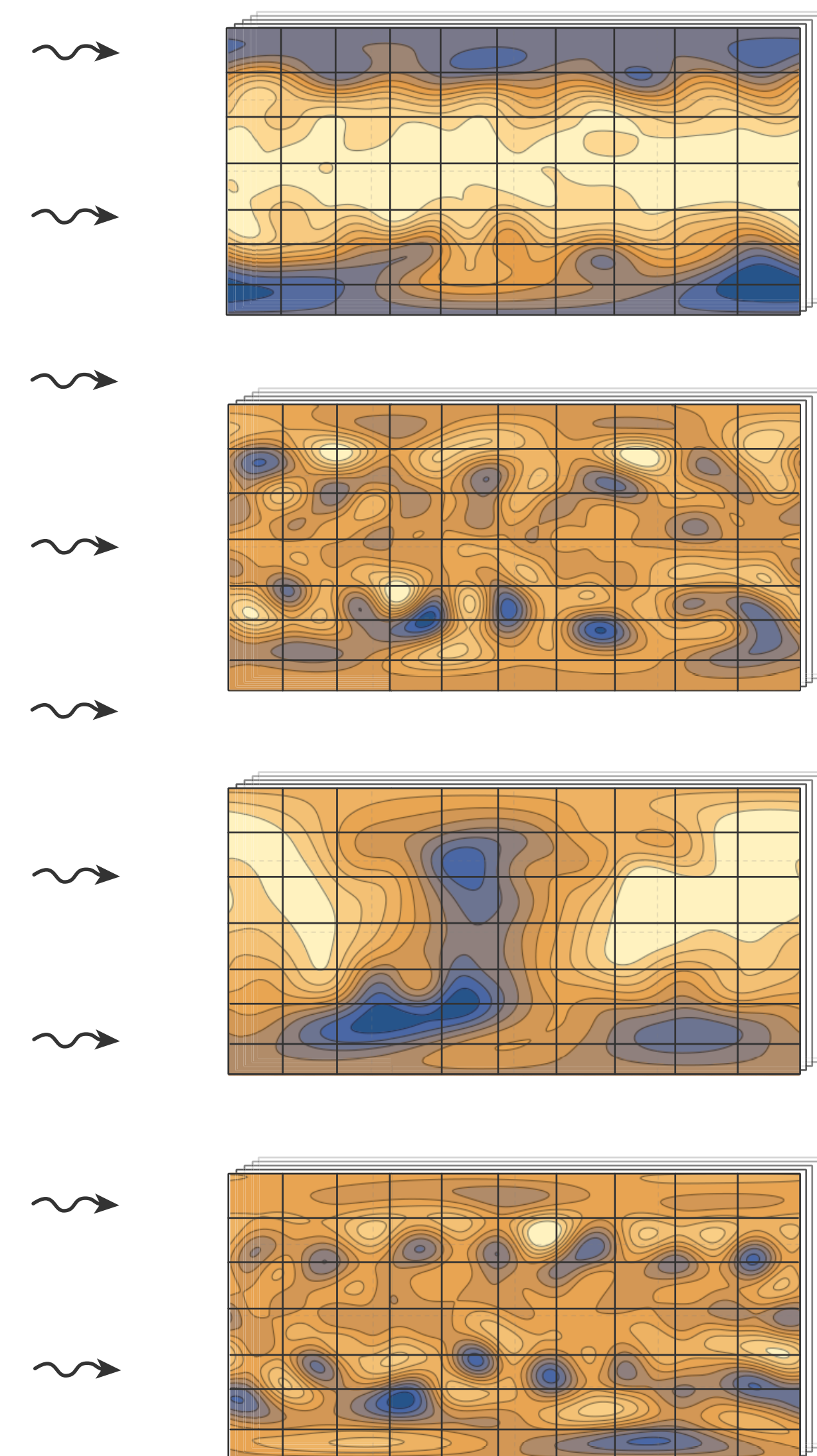
AtmoRep

- Numerical statistical atmospheric model
 - › Task agnostic but applicable to a wide range of applications and scientific questions
 - › Corresponds to large scale representation learning (foundational model) in machine learning
 - › Training on observational data (with reanalysis as stepping stone for proof-of-concept)

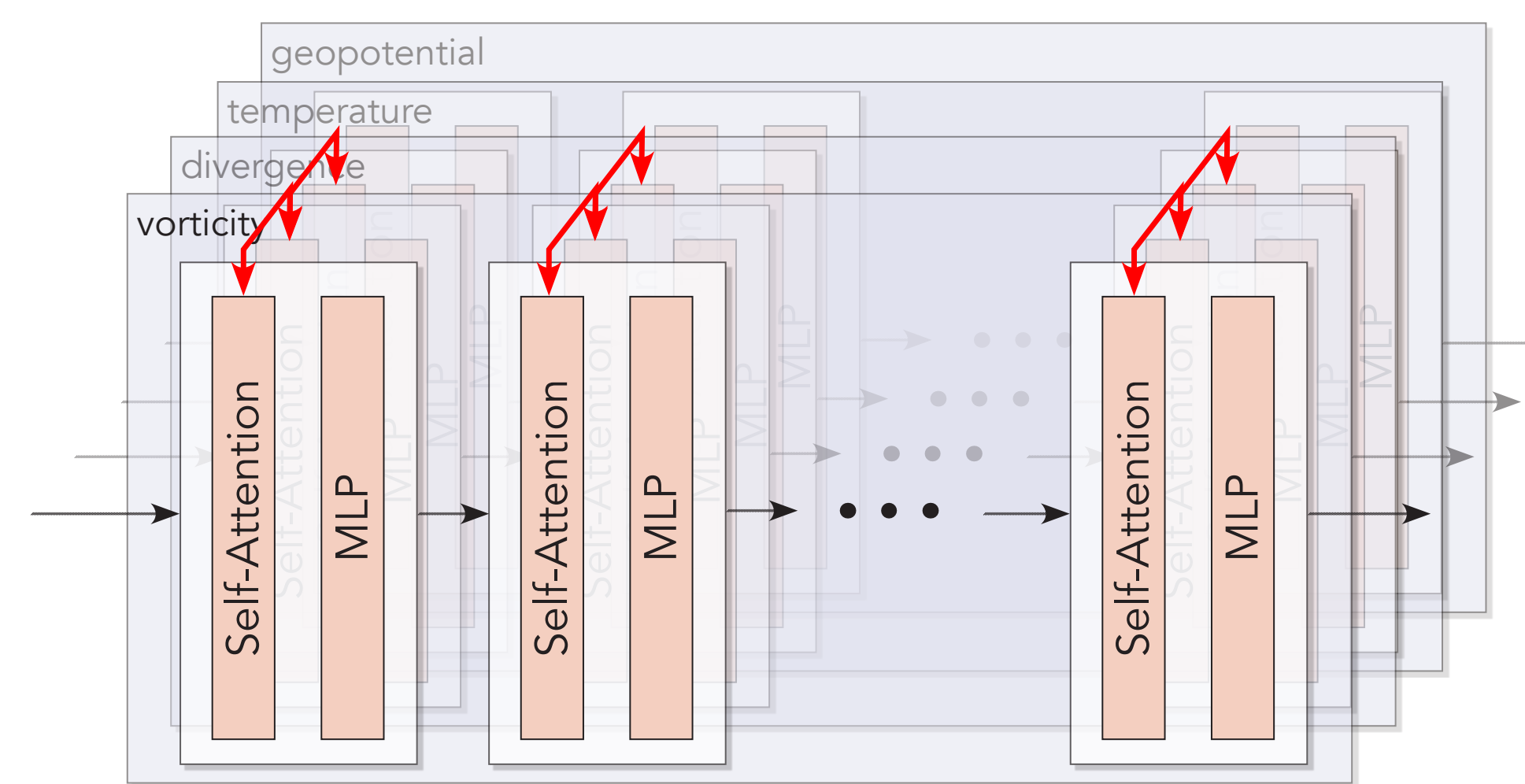
AtmoRep



ERA5 reanalysis

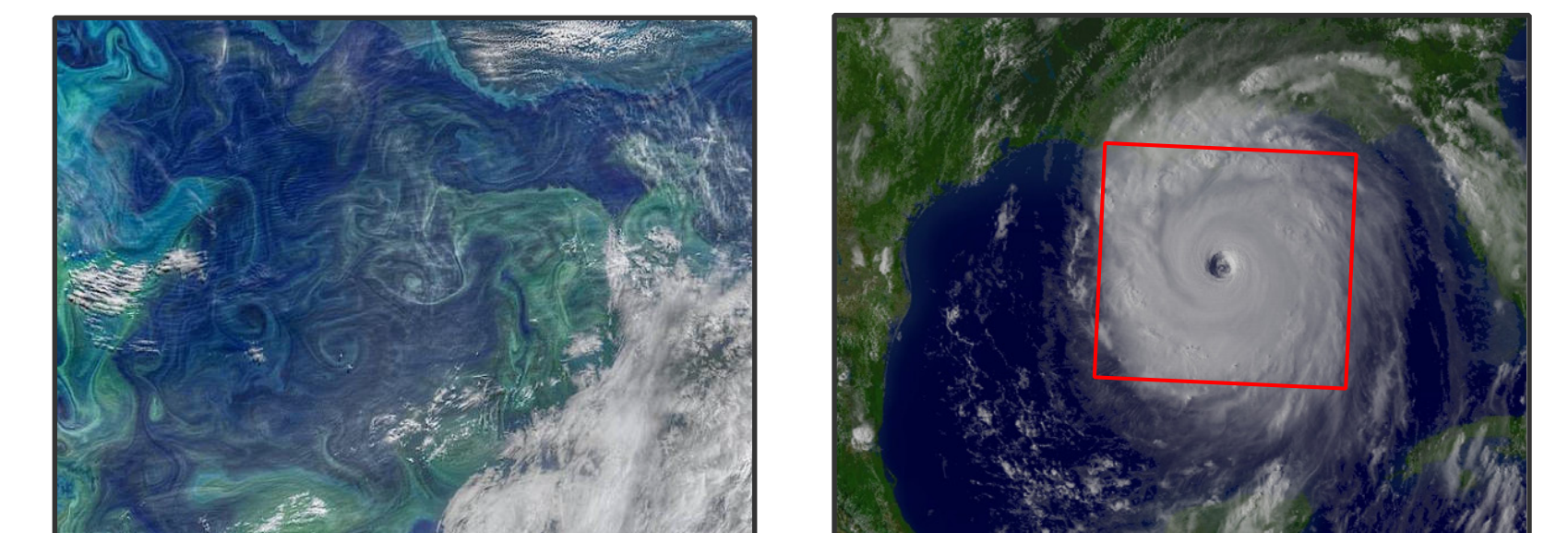


$$p_{\theta}(y|x)$$



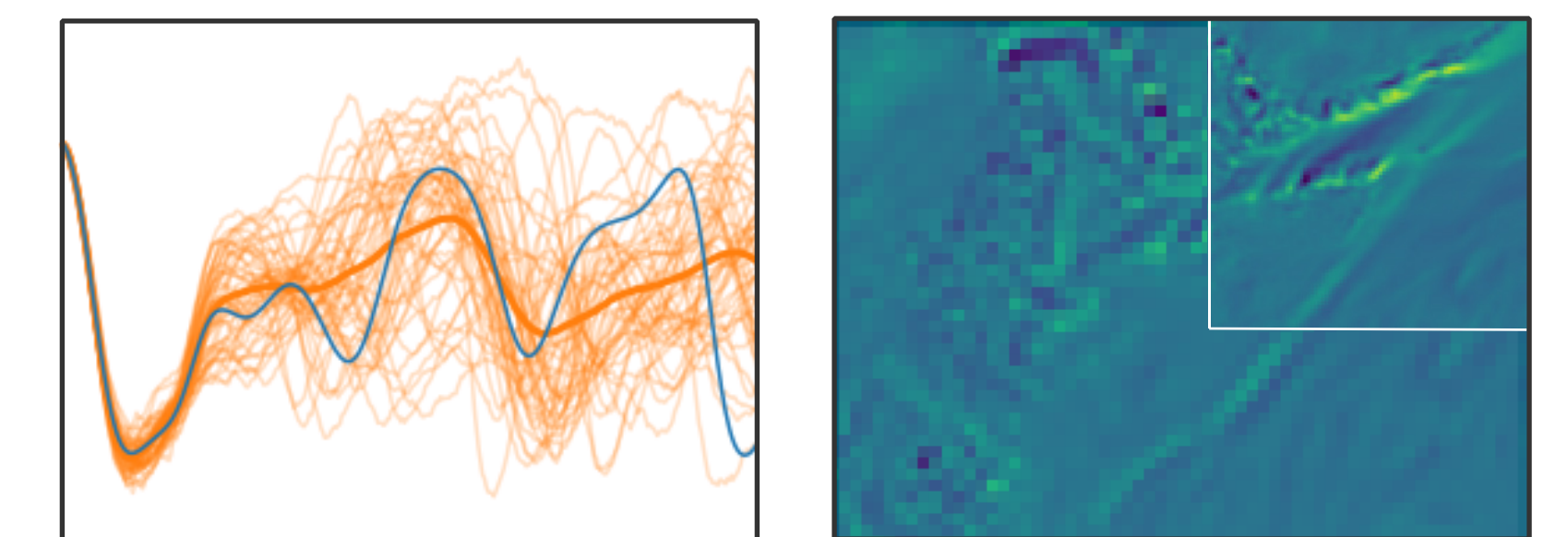
large transformer
with 4×10^9 parameters

applications



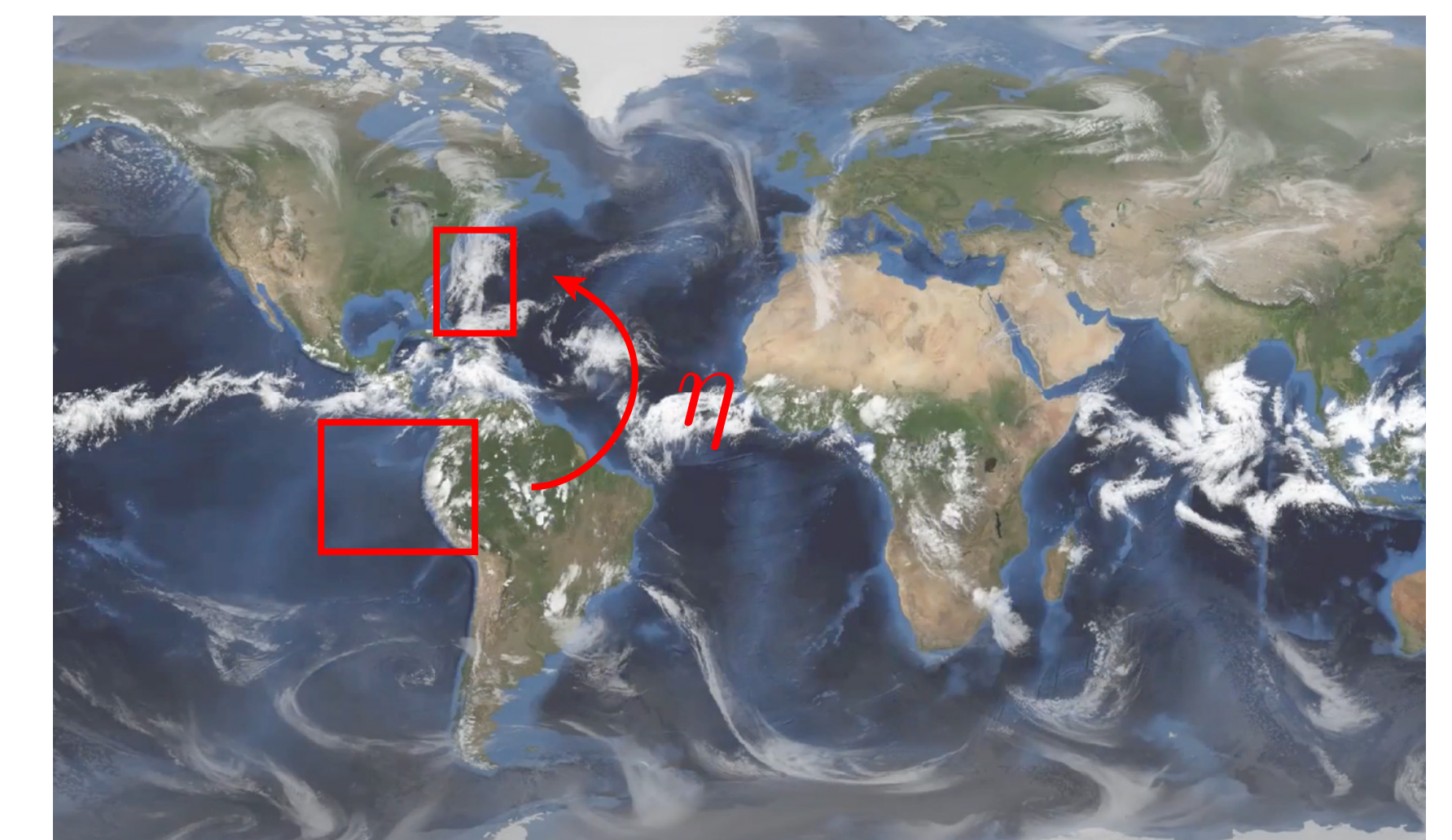
Forecasting

Impact analysis



Climate projections

Downscaling

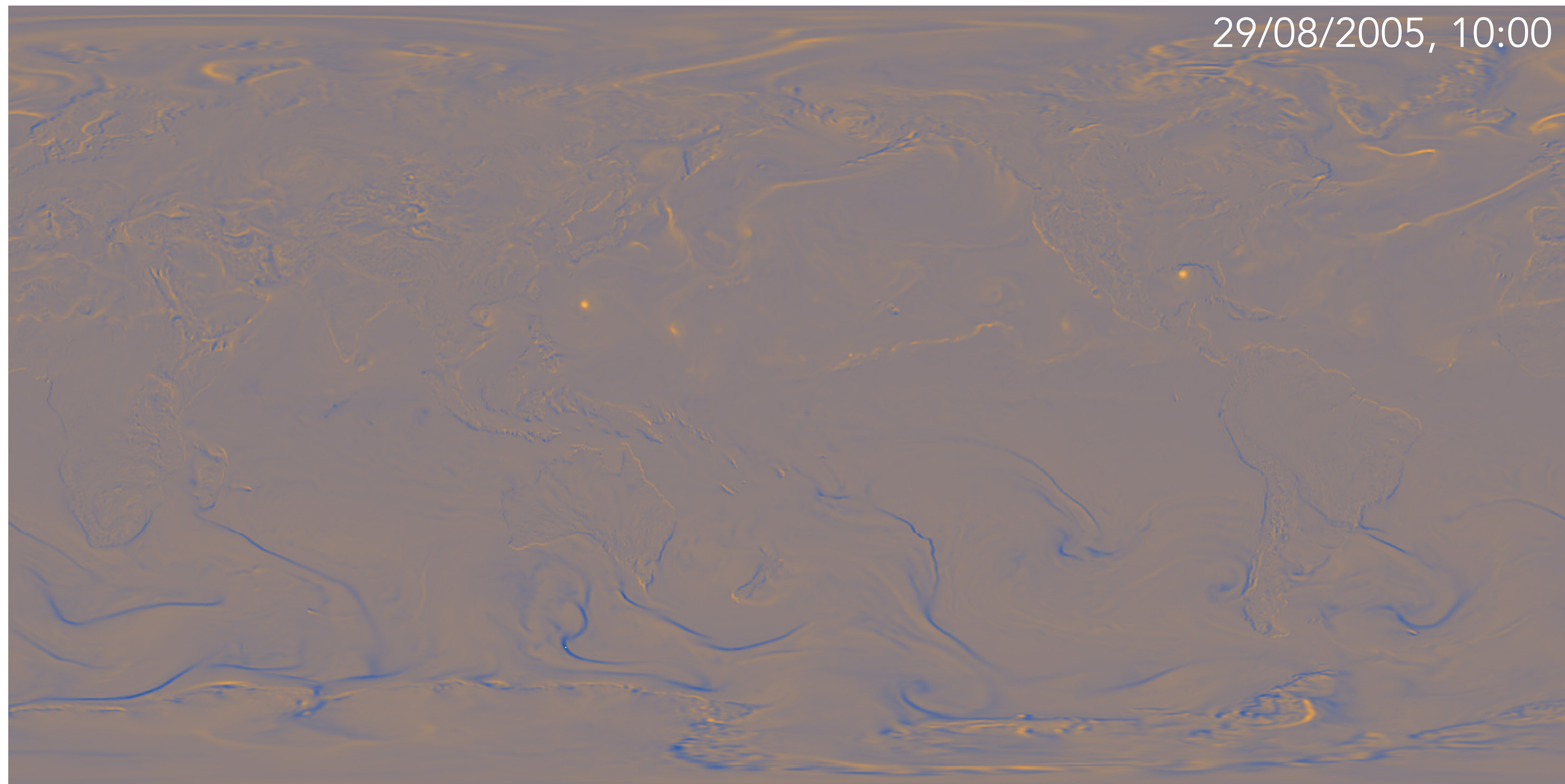


scientific insight

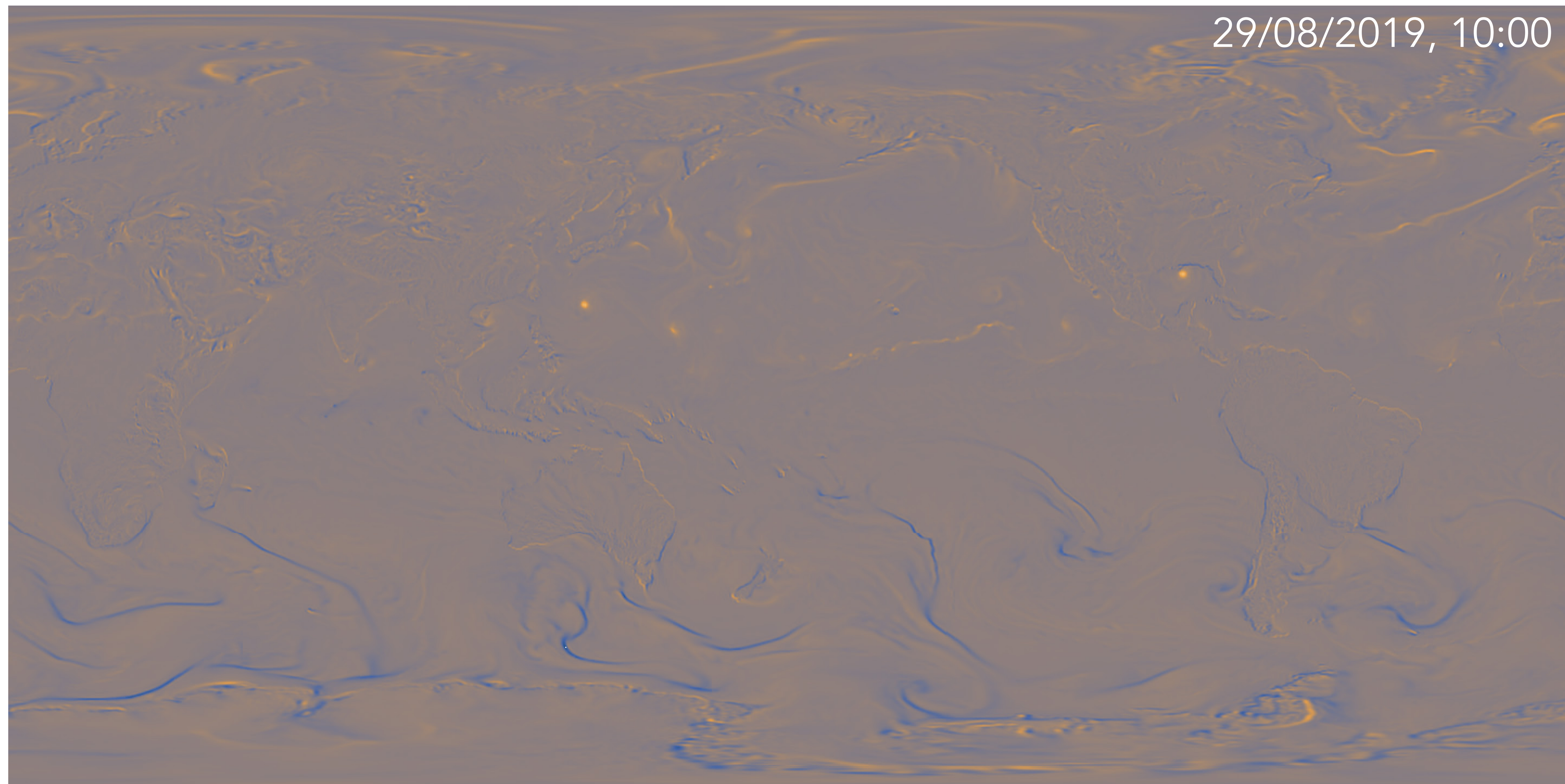
AtmoRep Training

- Statistical atmospheric model $p_{\theta}(y|x)$
 - › Training models spatio-temporal relationship between arbitrary state x and y
 - › Train on local relations in space-time whose iteration gives global dependence
 - › Probabilistic relationship is modeled in training using ensemble prediction (generalizing discrete probability distribution used in large language models)

How would Katrina have looked in 2019?



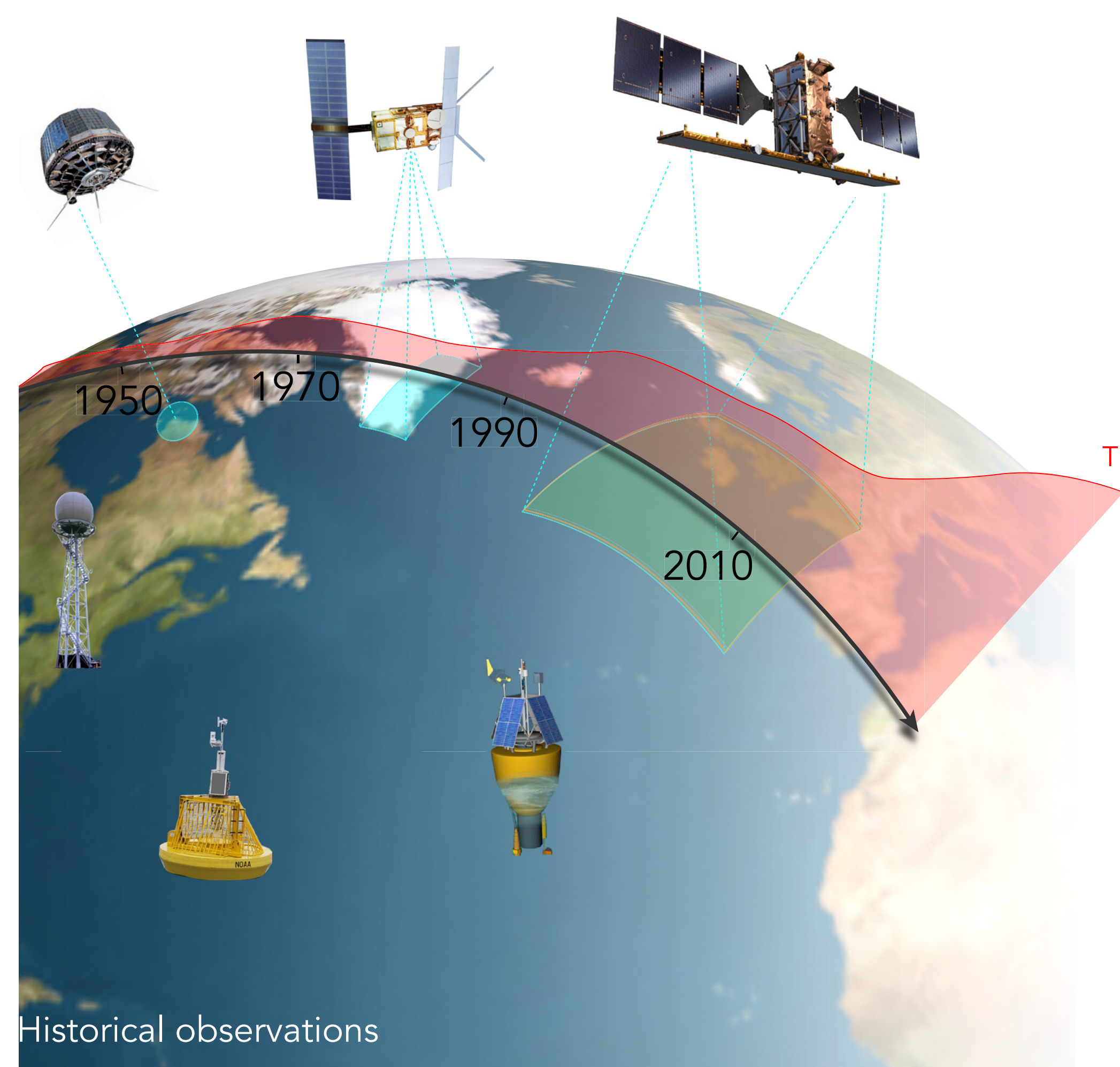
How would Katrina have looked in 2019?



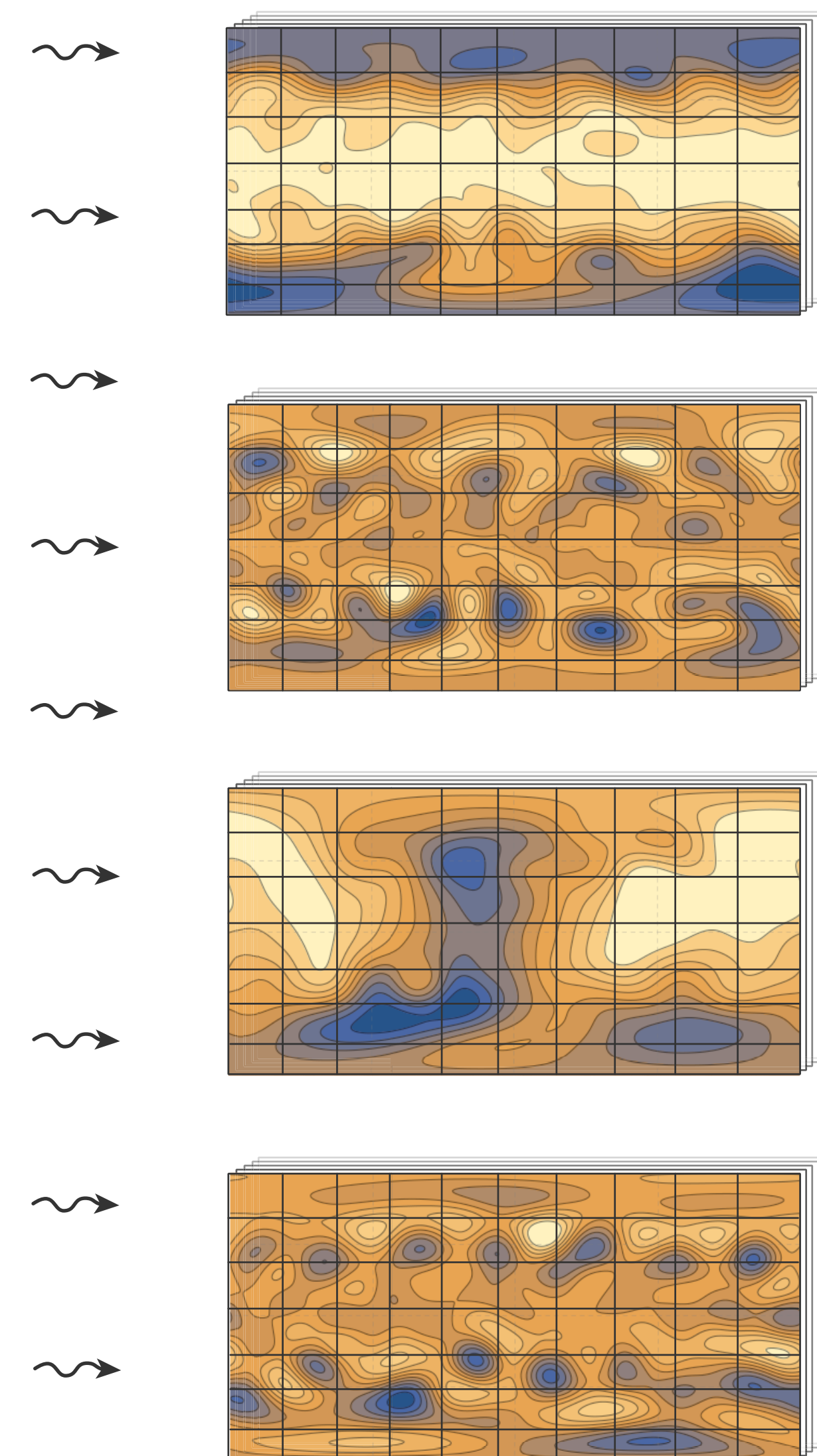
How would Katrina have looked in 2019?



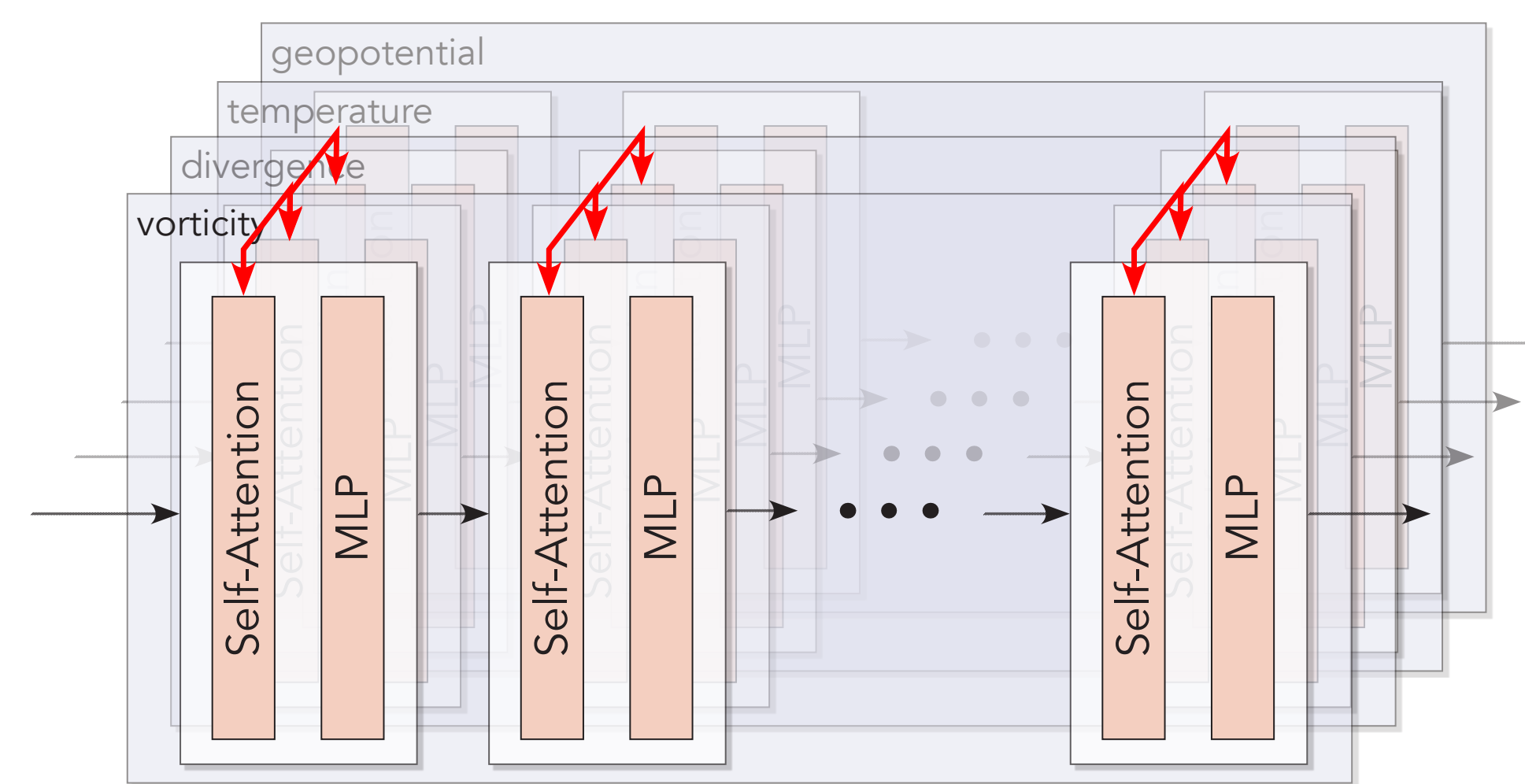
AtmoRep



ERA5 reanalysis

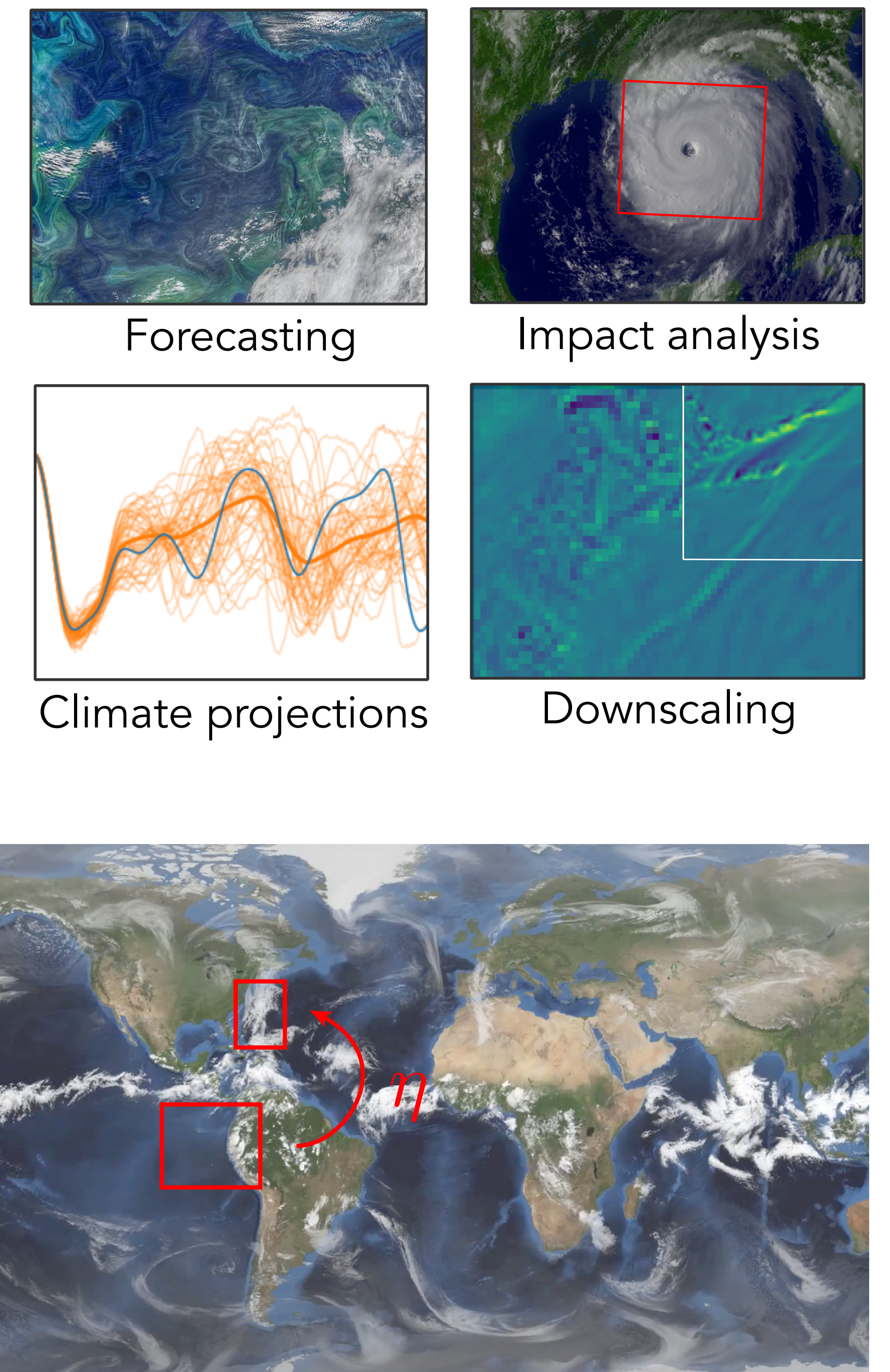


$$p_{\theta}(y|x)$$



large transformer
with 4×10^9 parameters

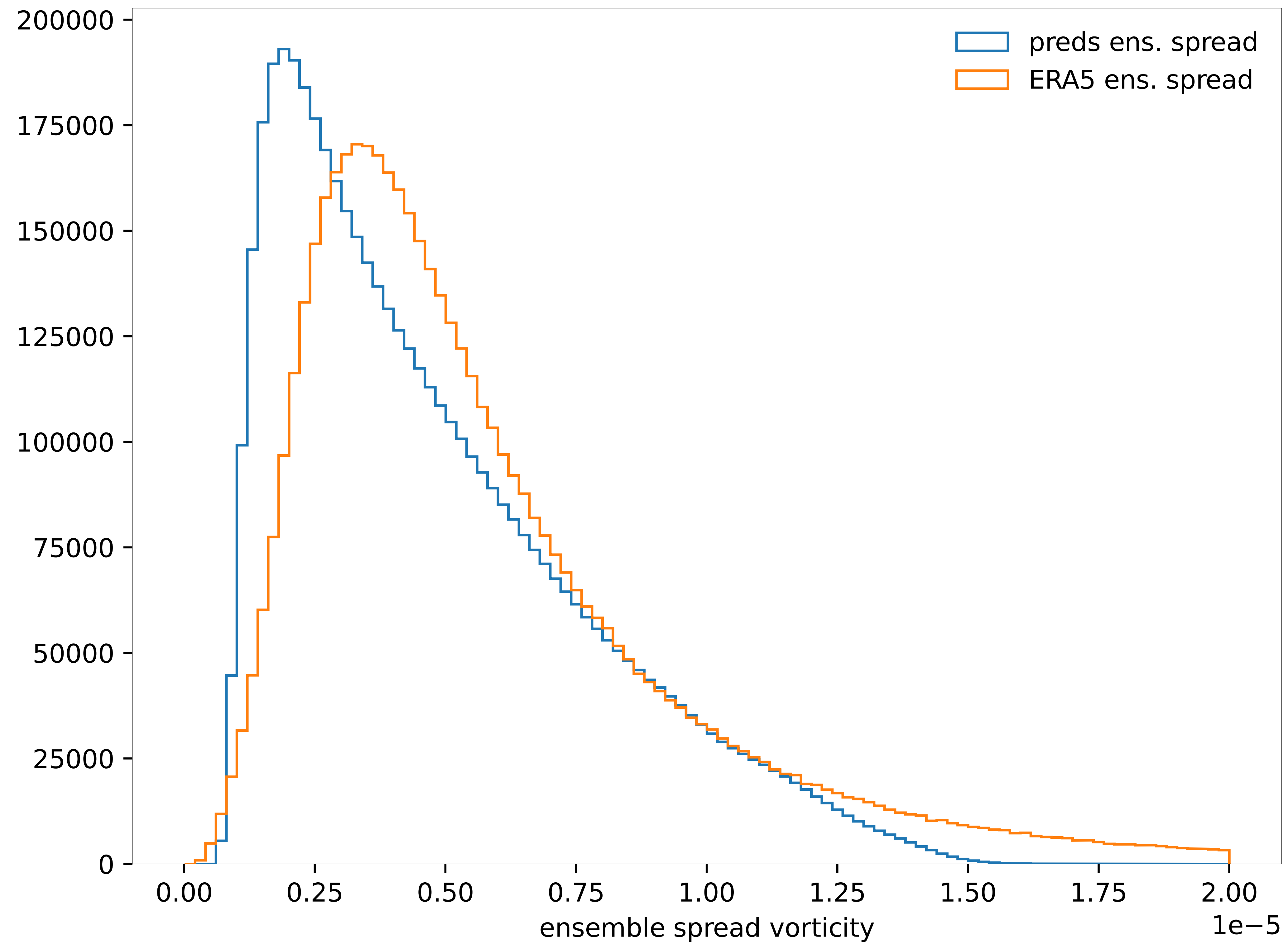
applications



scientific insight

More details ...

AtmoRep Training



Large scale representation learning

- Think of large language models: ChatGPT, GPT-X, PaLM, ...
 - › Useful for scientists (reviews, grant applications, thesis appraisals, ...) but not directly for science

Large scale representation learning

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

Large scale 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 with very large networks
 - › Useful for downstream applications using tail network, fine-tuning, or in-context learning

Large scale 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 with very large networks parameters
 - › Useful for downstream applications using tail network, fine-tuning, or in-context learning

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
 - ERA5 reanalysis: 6+ PB
 - ESA's MetOp-SG satellites: 8 x 864 GB/day
 - Data essentially completely unlabelled
- GPT-3: 10^{11} tokens
ERA5: 5^{14} tokens

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

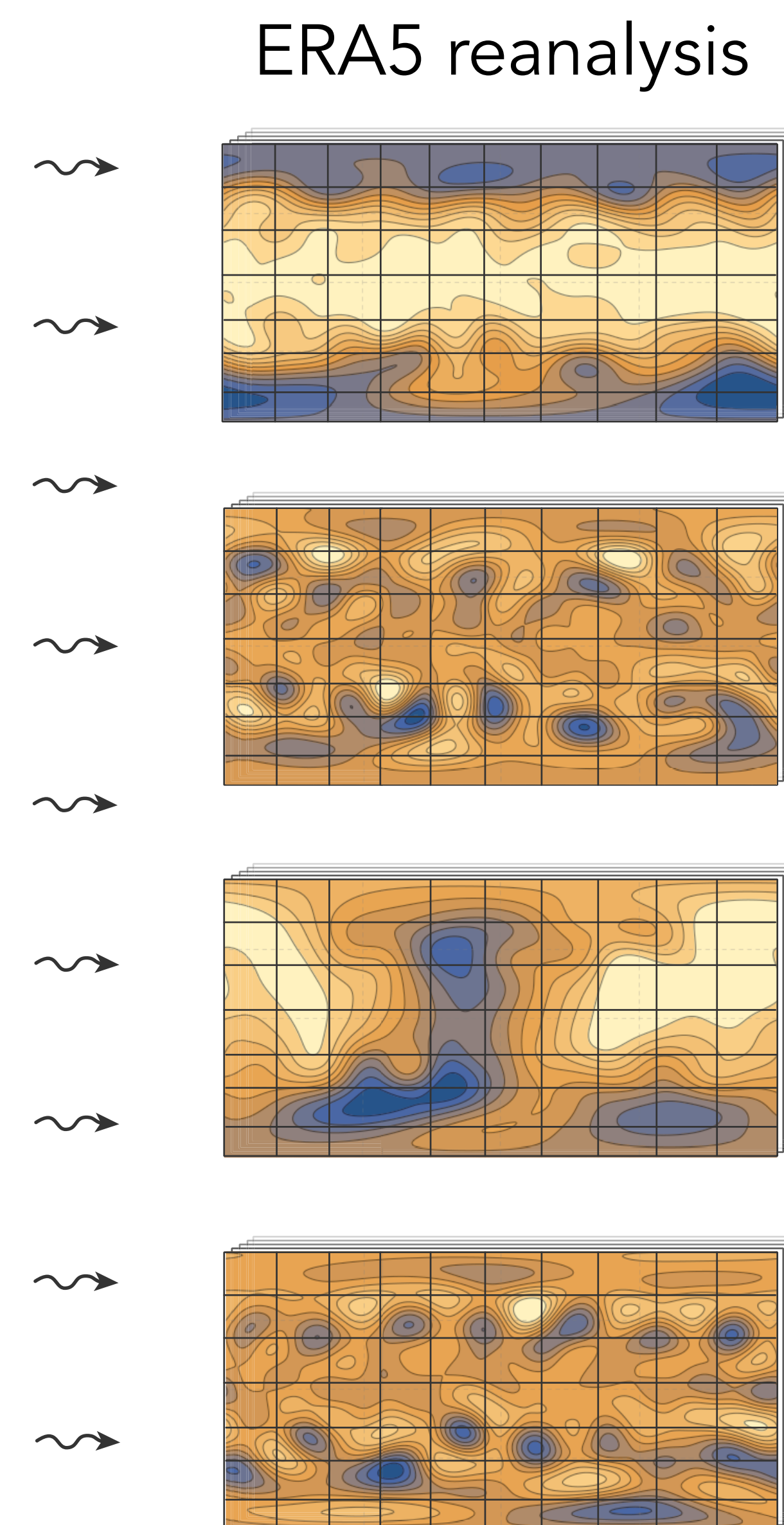
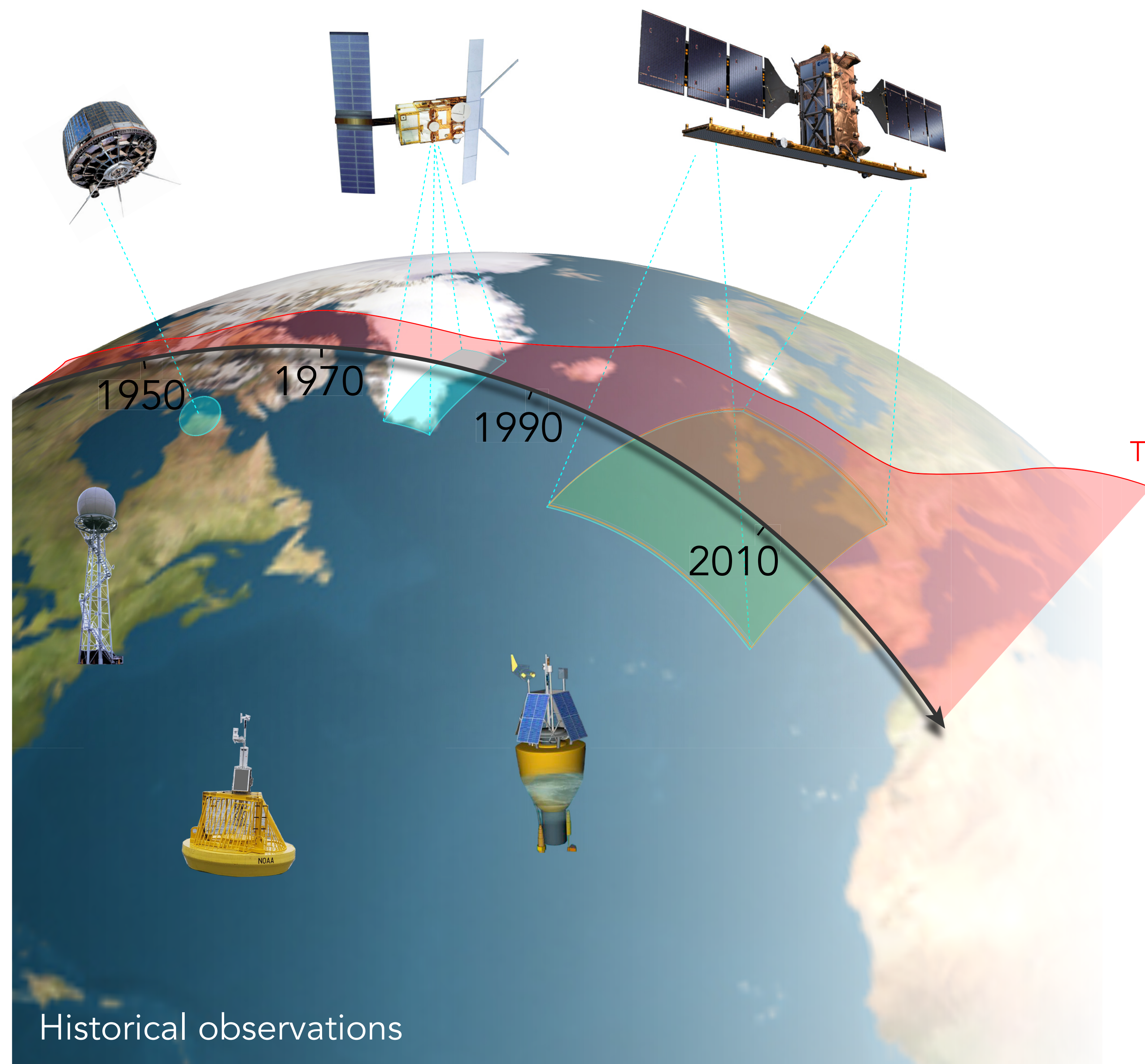
Representation learning for the Earth sciences?

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

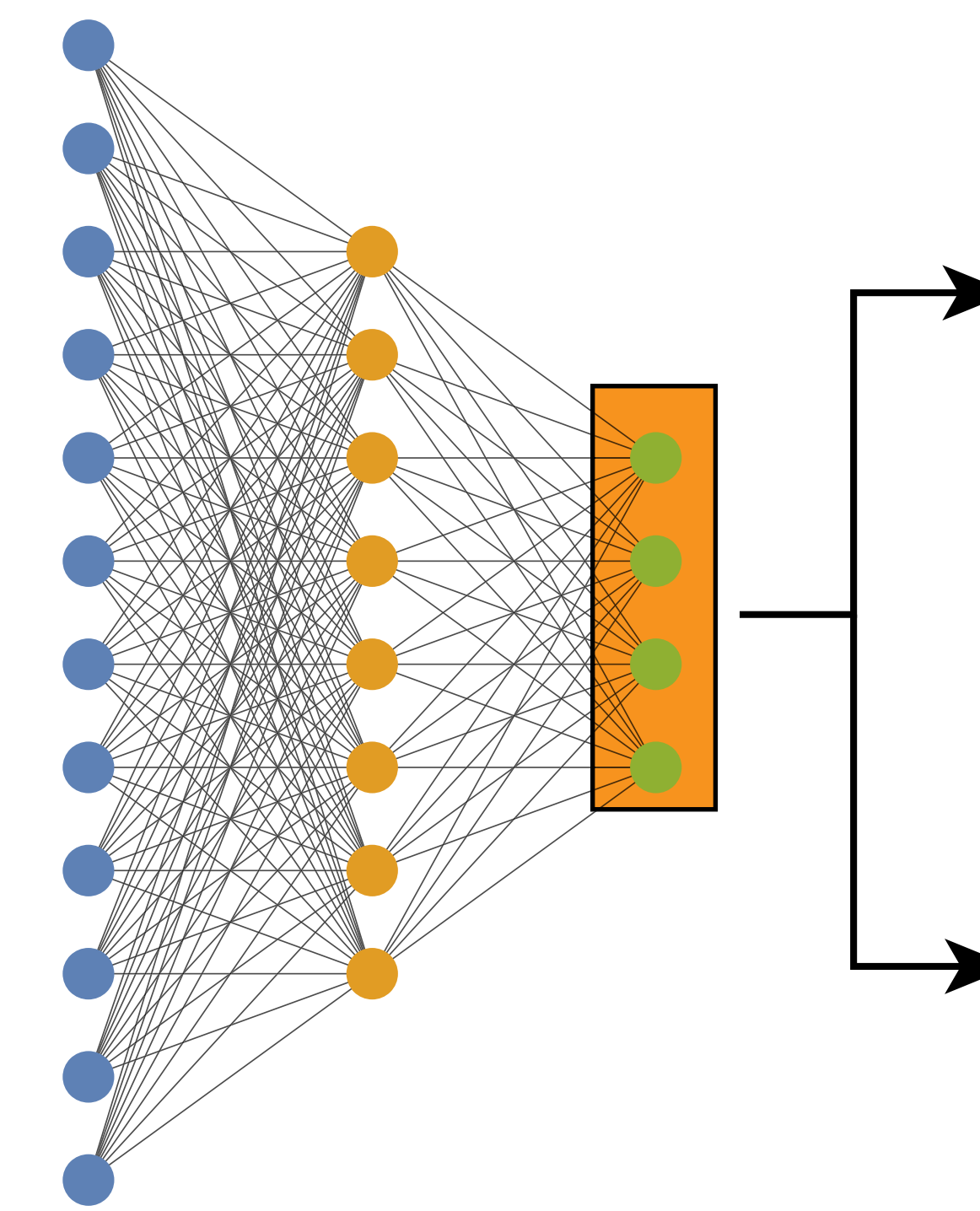
AtmoRep

Large scale representation learning of
atmospheric dynamics

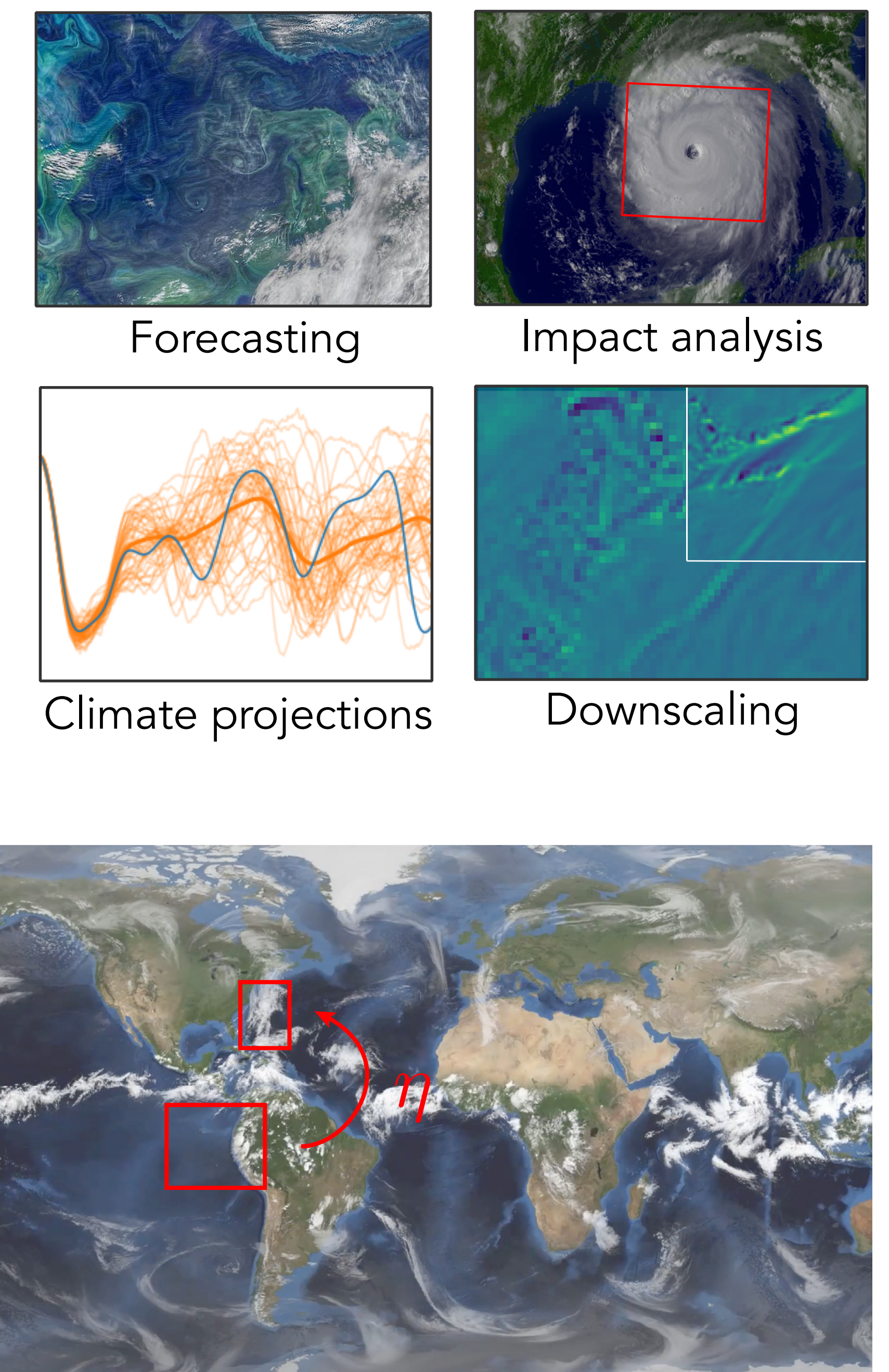
AtmoRep



large scale
machine learning

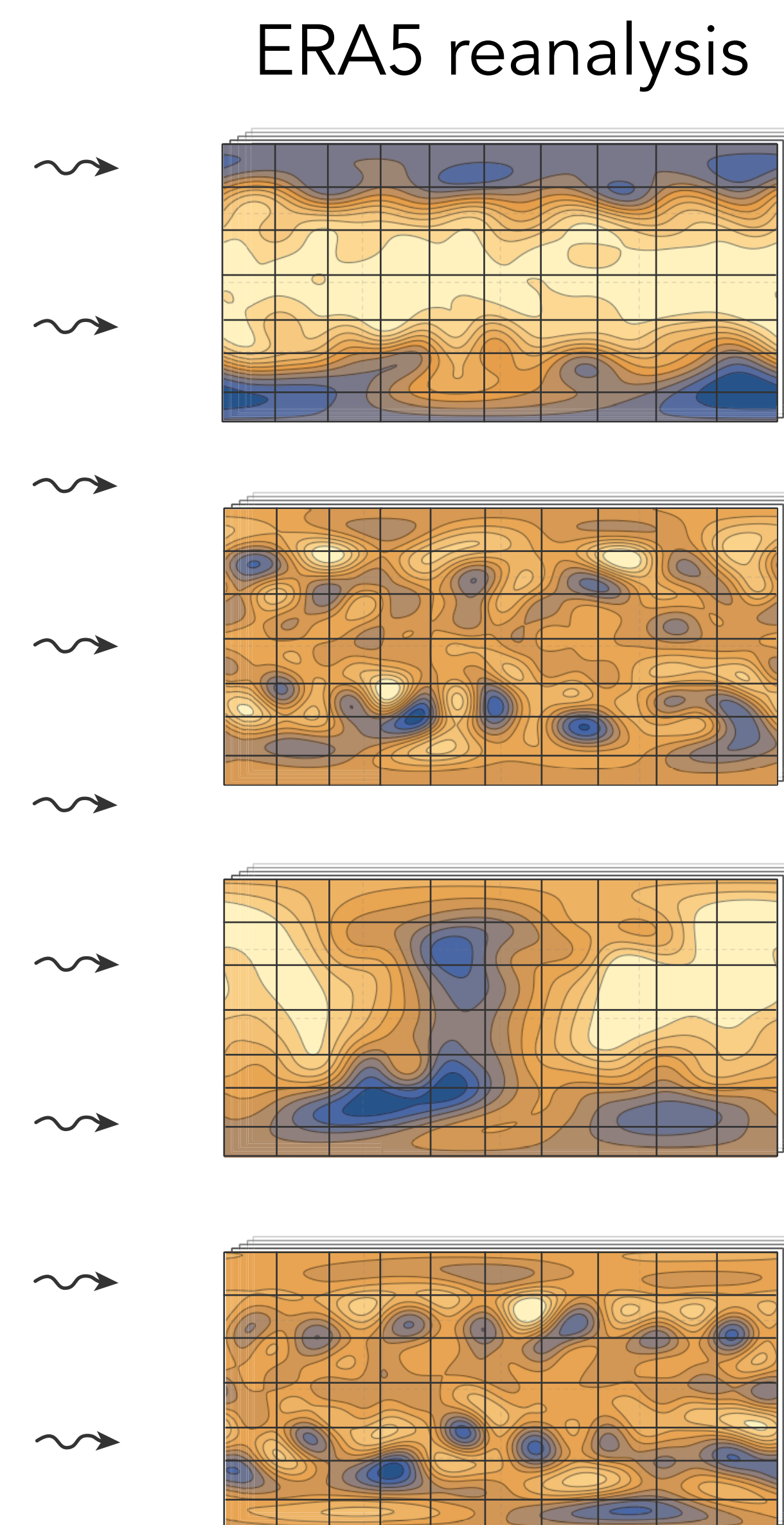
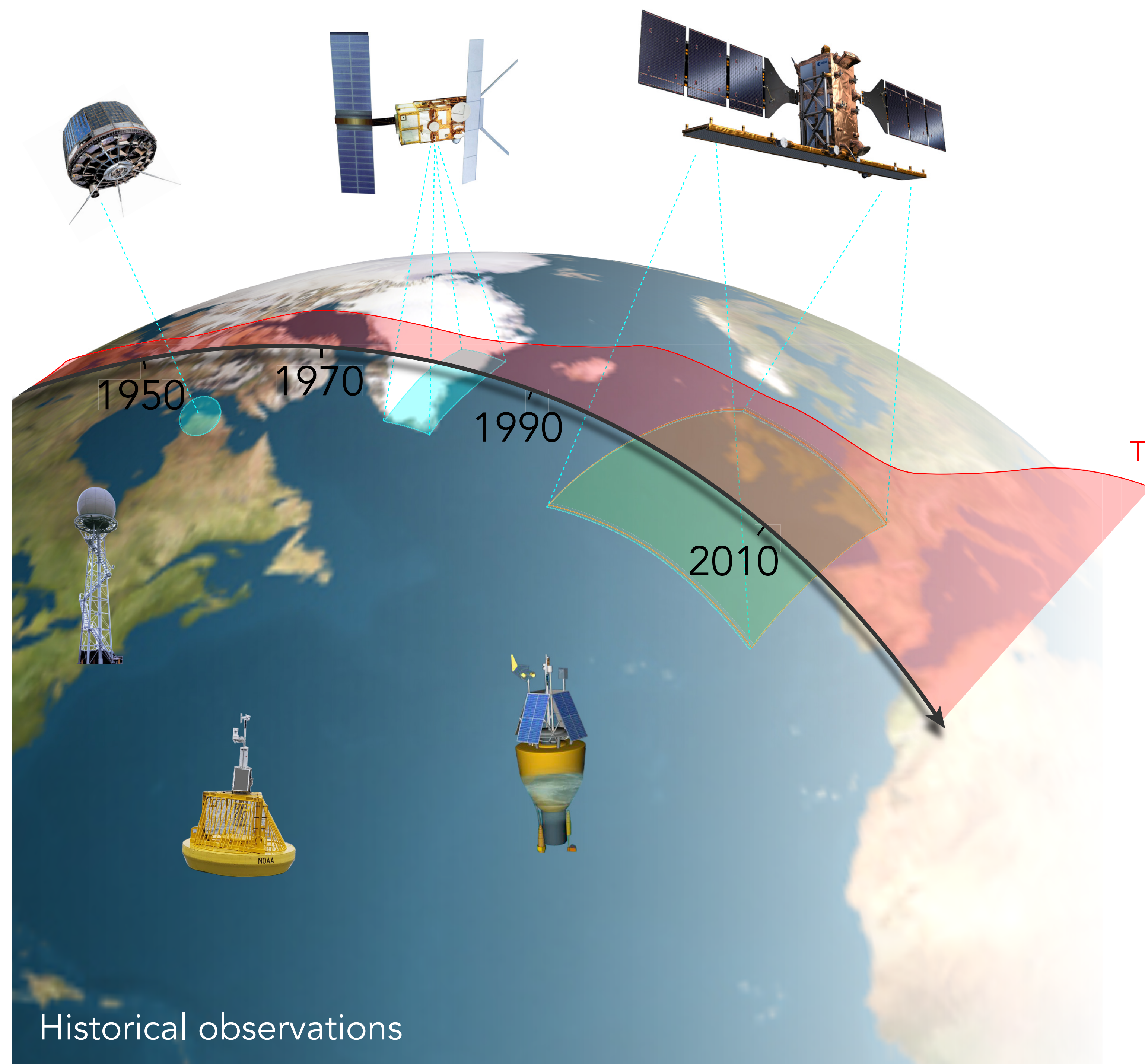


applications

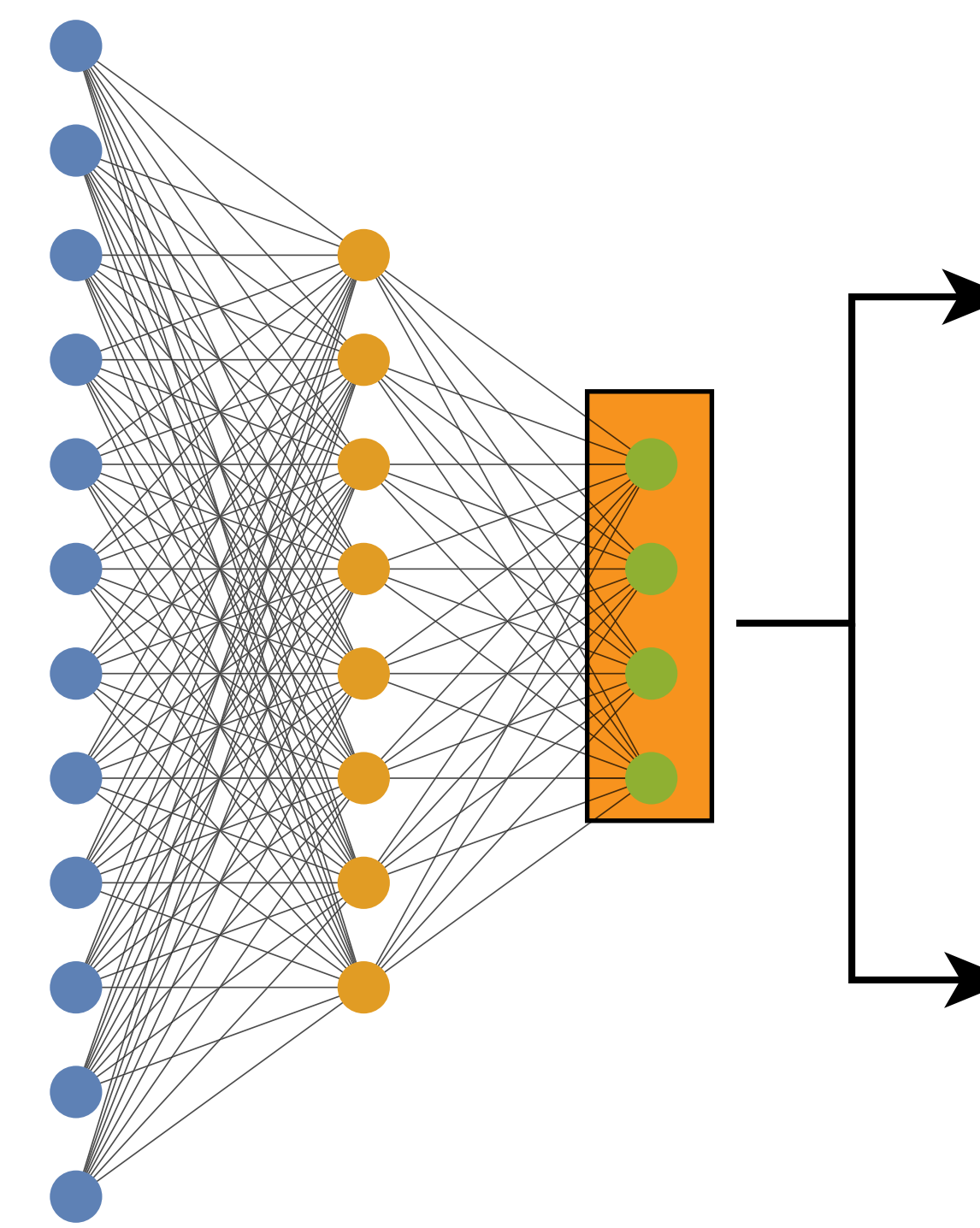


scientific insight

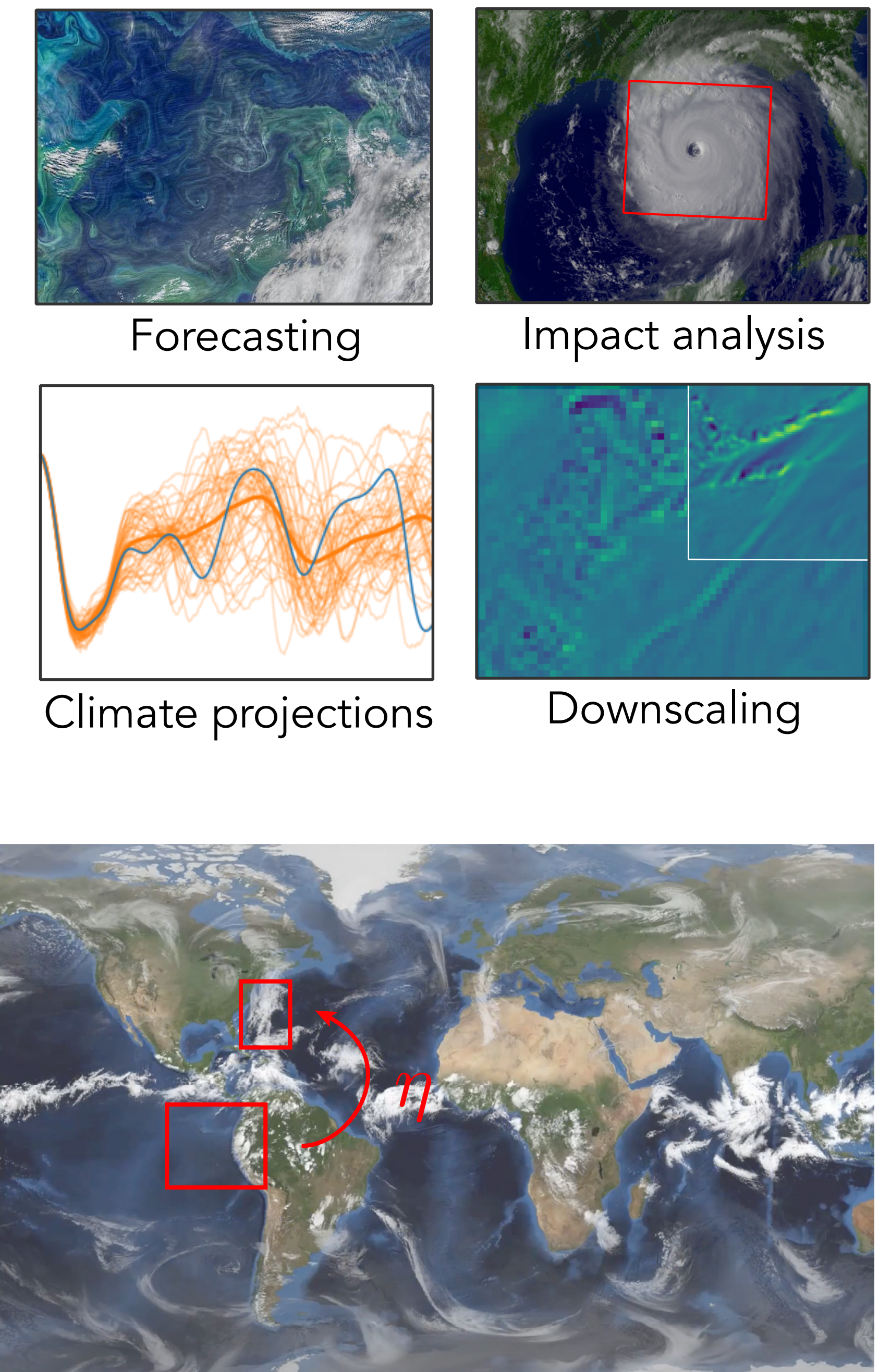
AtmoRep



$$p_{\theta}(y|x, \alpha)$$



$$p_{\theta}(y|x, \alpha, \tau)$$

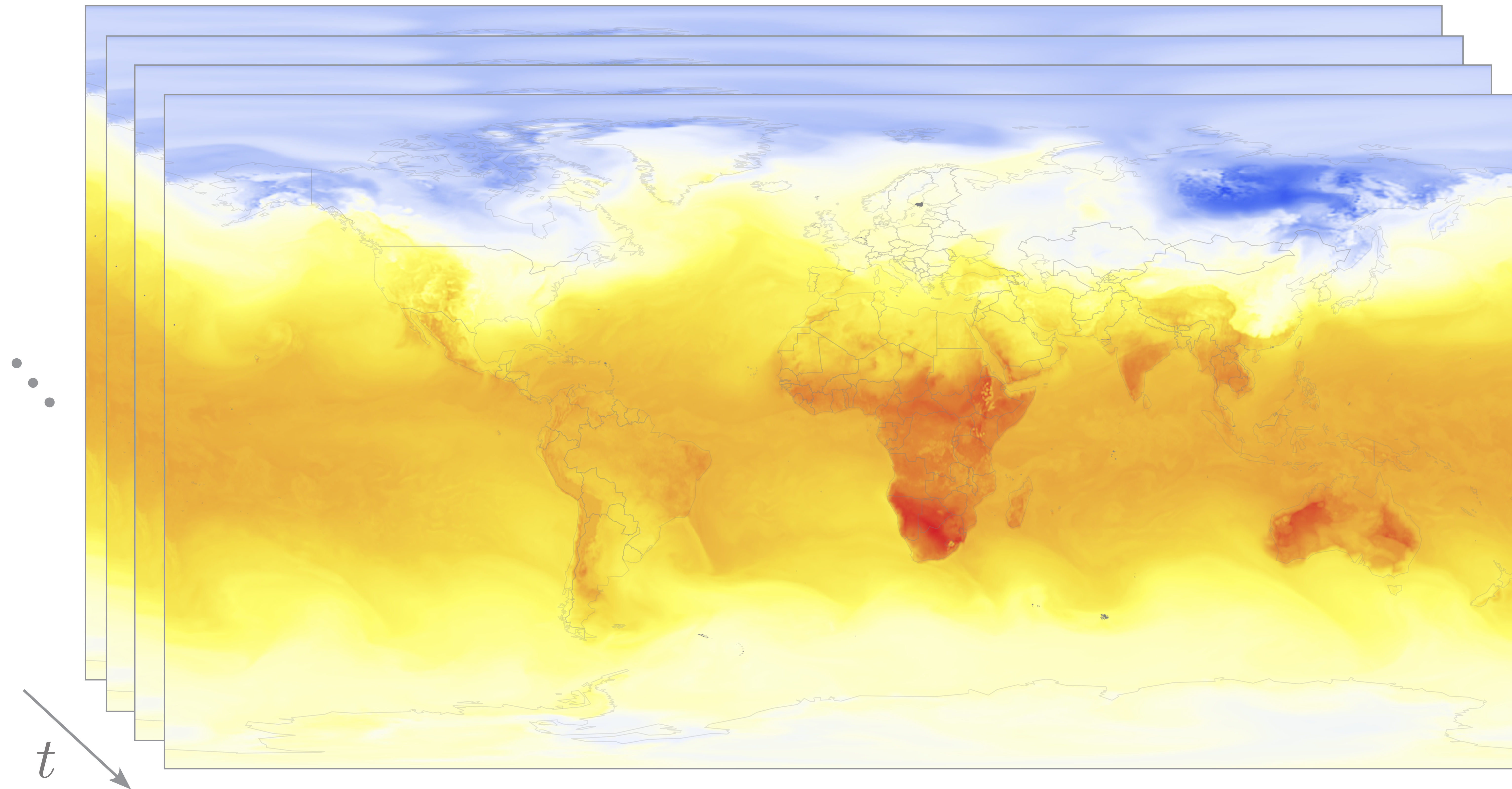


$$p_{\theta}(y|x, \alpha) \approx \bar{p}(\bar{y}|\bar{x})$$

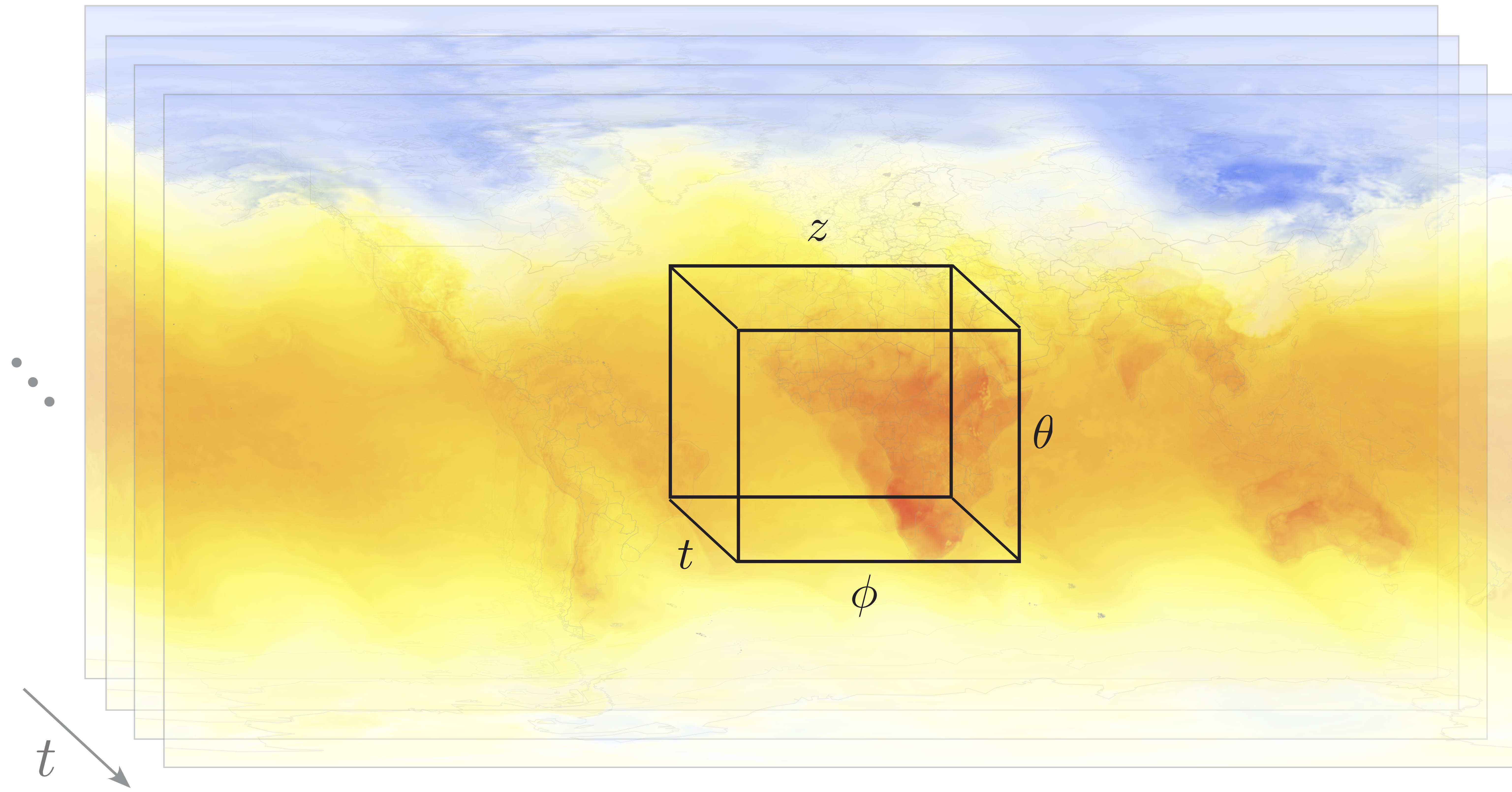
AtmoRep network architecture

- Network is local in space-time

AtmoRep network architecture



AtmoRep network architecture

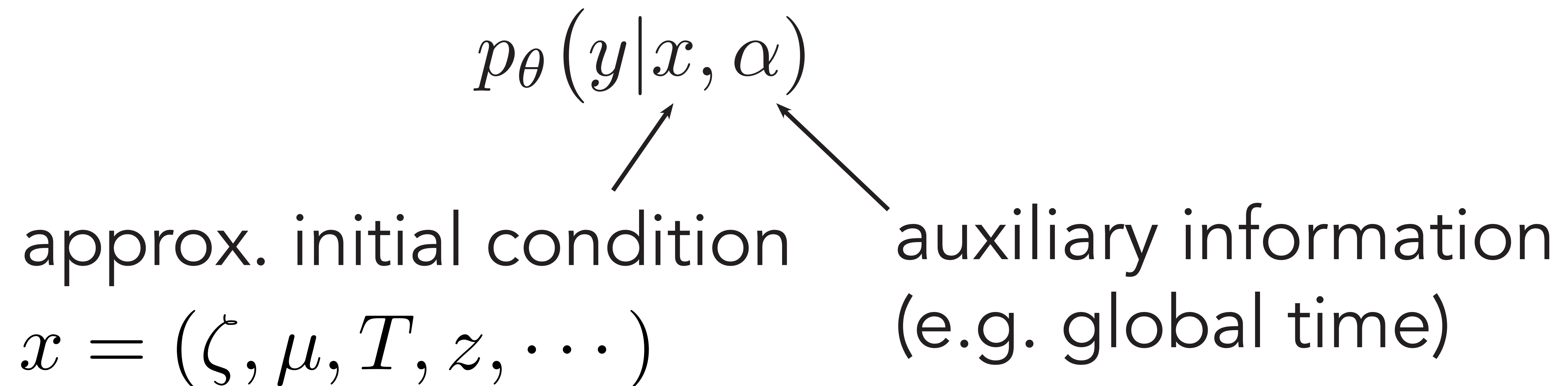


AtmoRep network architecture

- Network is local in space-time
 - › Physics are universally valid
 - › Local particularities can be learned by providing time + space position as auxiliary information

AtmoRep network architecture

- Network is local in space-time
 - › Physics are universally valid
 - › Local particularities can be learned by providing time + space position as auxiliary information



AtmoRep network architecture

- Network is local in space-time
 - › Physics are universally valid
 - › Local particularities can be learned by providing time + space position as auxiliary information
 - › Machine learning model can be leaner and learn faster

AtmoRep network architecture

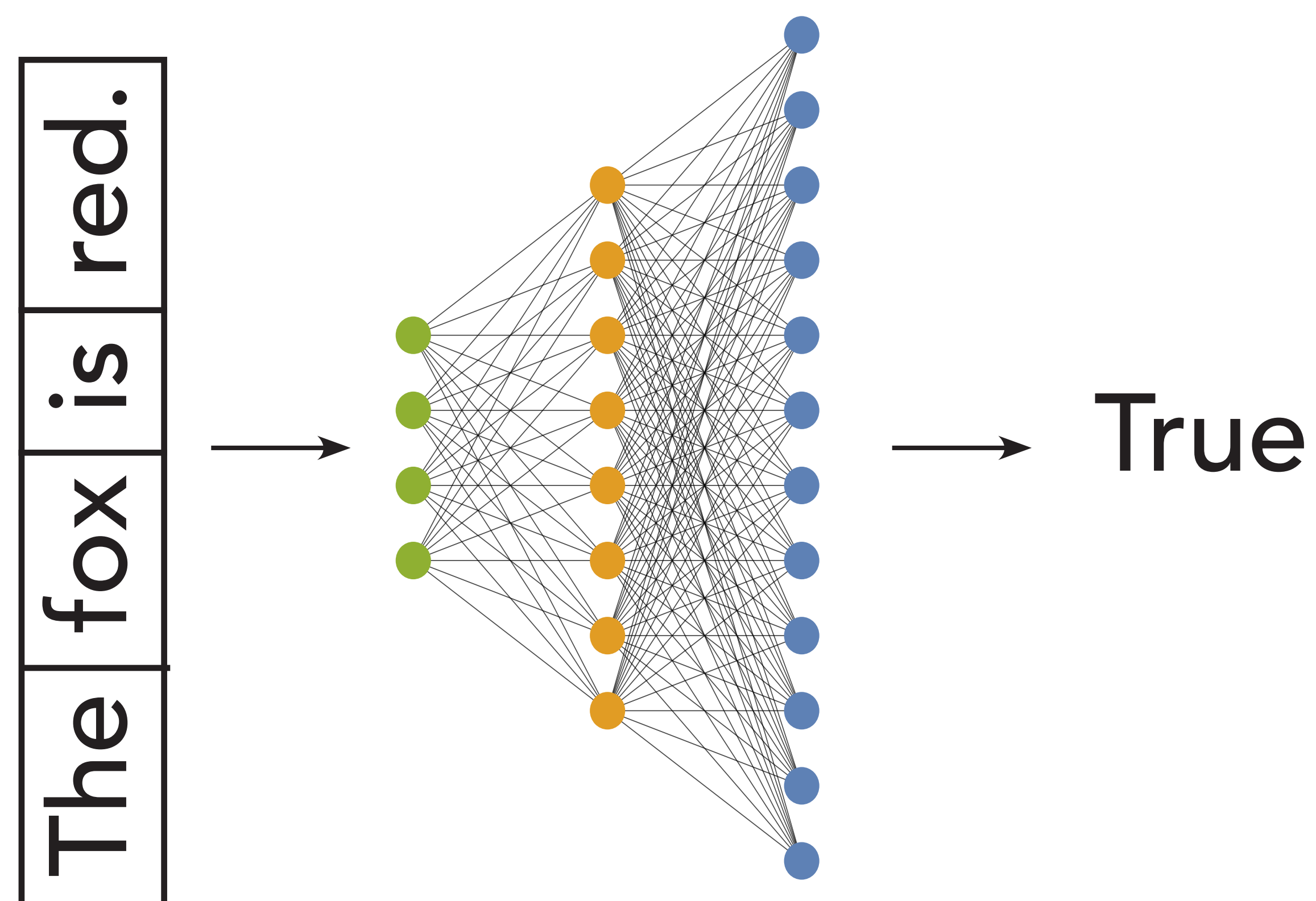
- Transformer as network architecture
 - › Scales well to very large data-sets
 - › Generative model (with decoder)
 - › Attention maps provide (physical) interpretability

What is a token?

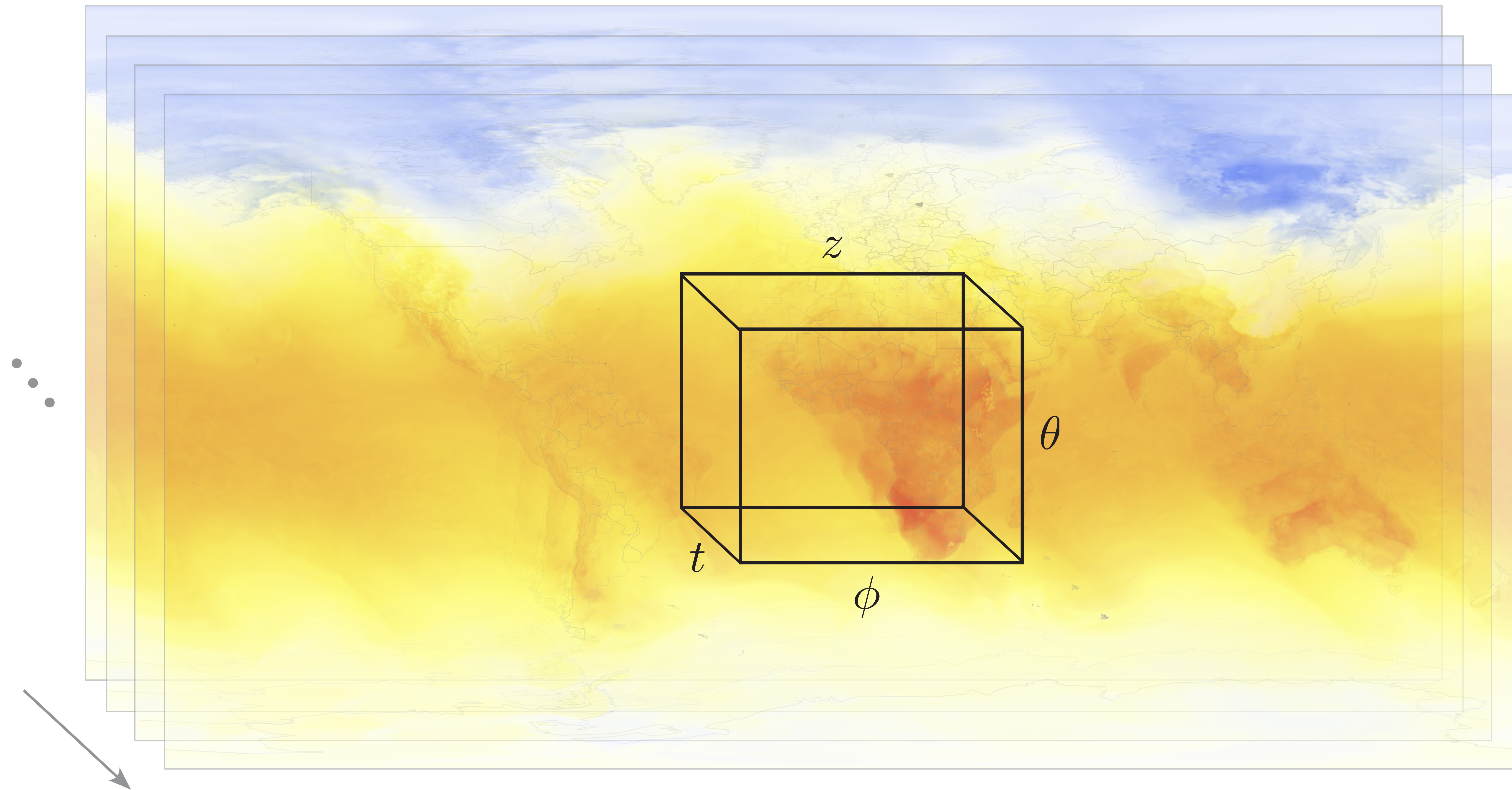
- Key property of transformers: sequence of inputs is processed simultaneously
 - › Language: input is sentence with words being tokens
 - › Images: small image patches

What is a token?

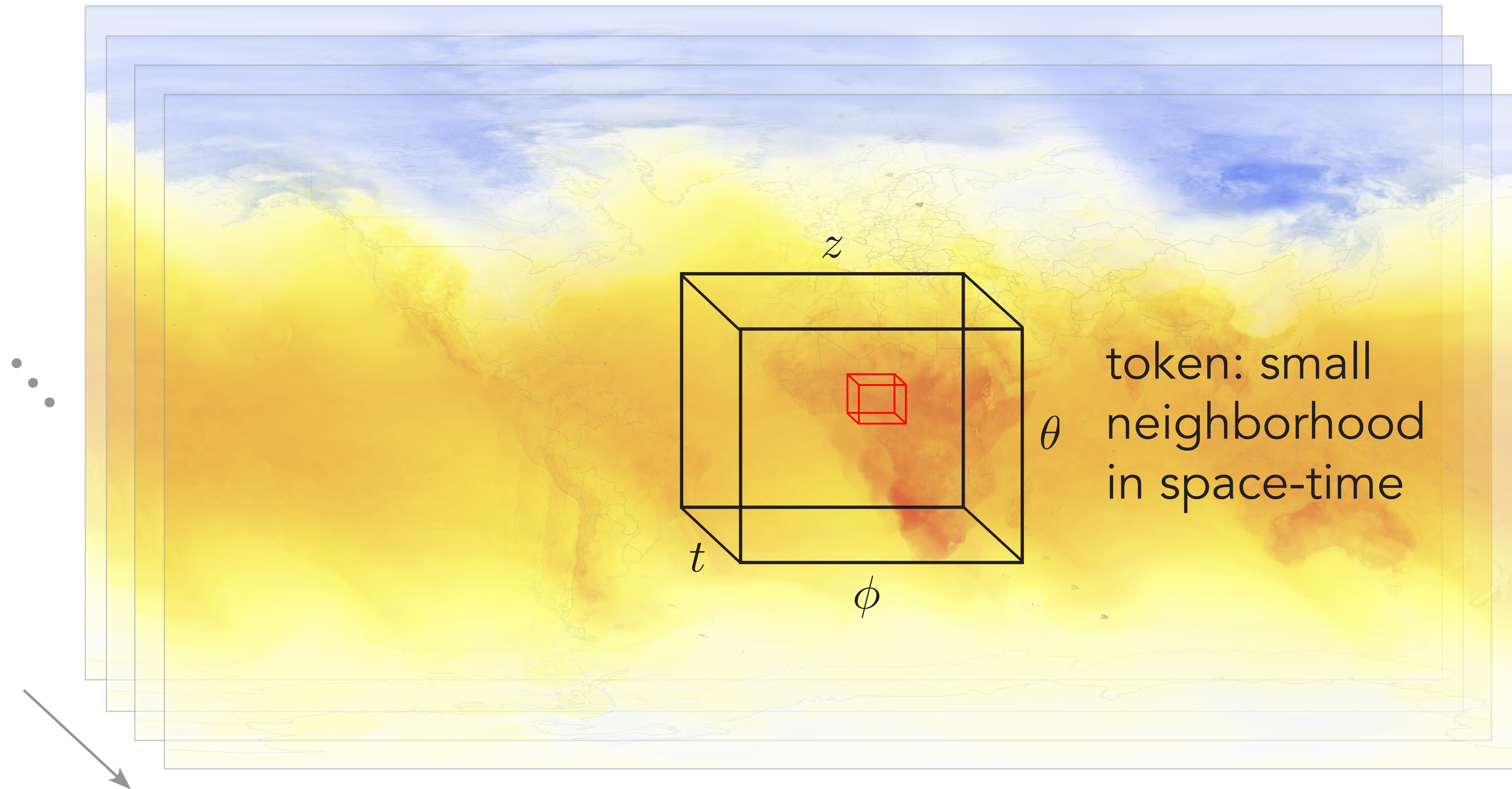
- Key property of transformers: sequence of inputs is processed simultaneously
 - › Language: input is sentence with words being tokens
 - › Images: small image patches



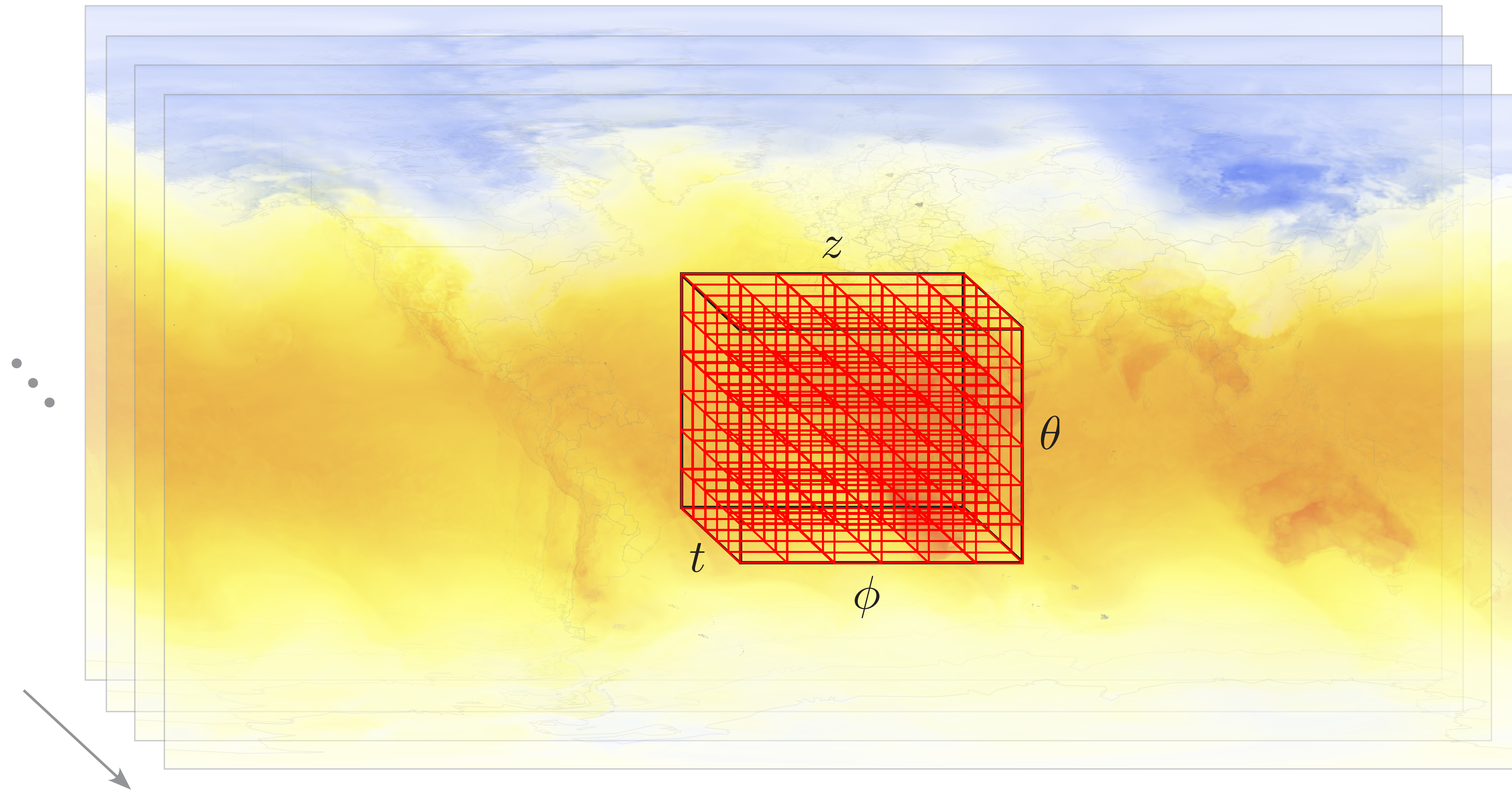
What is a token?



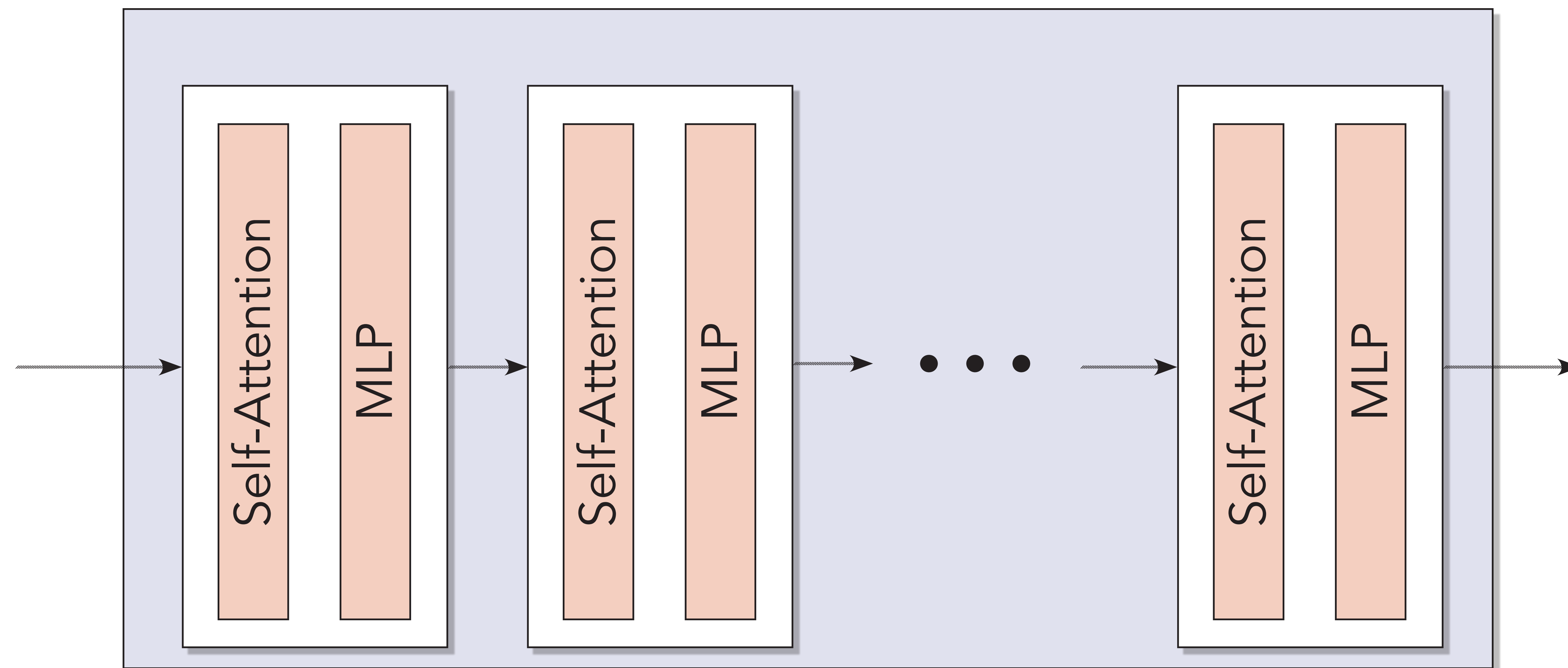
What is a token?



What is a token?



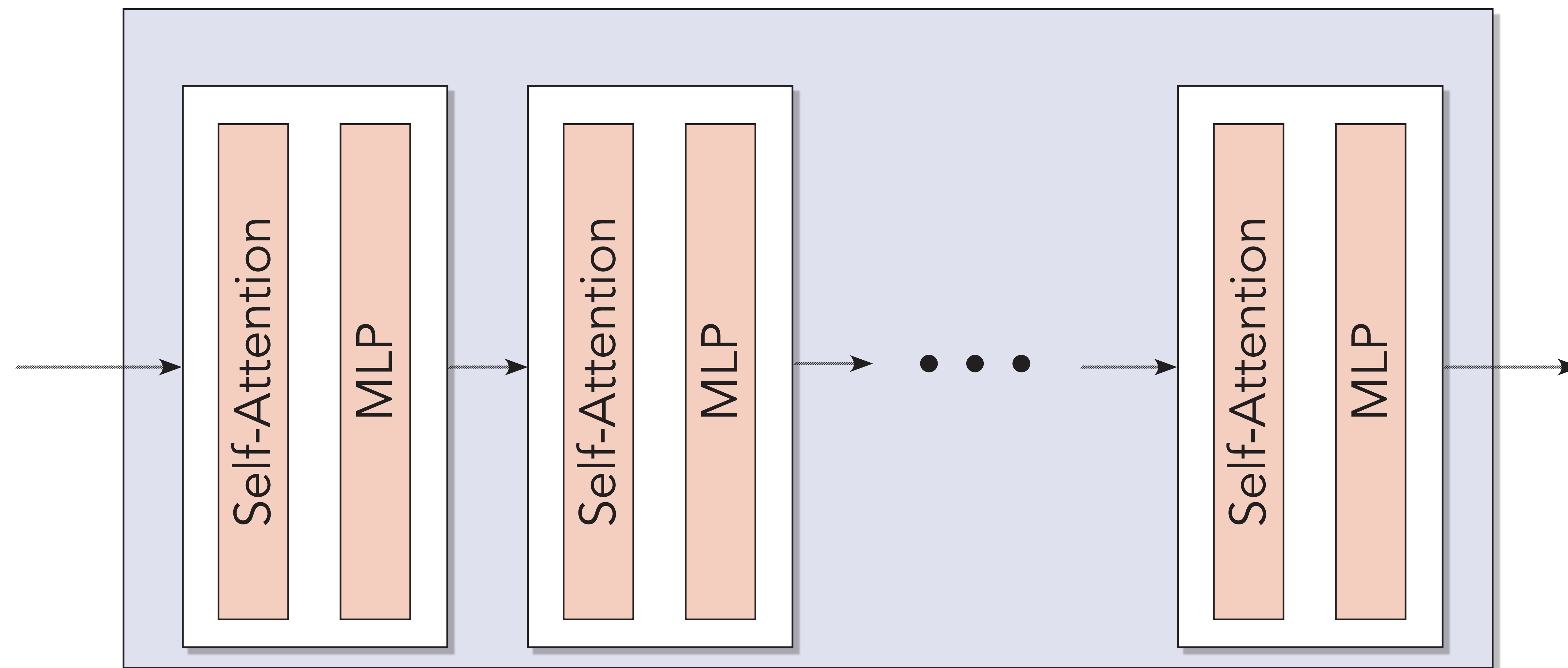
Multiformer



Multiformer

Self
attention

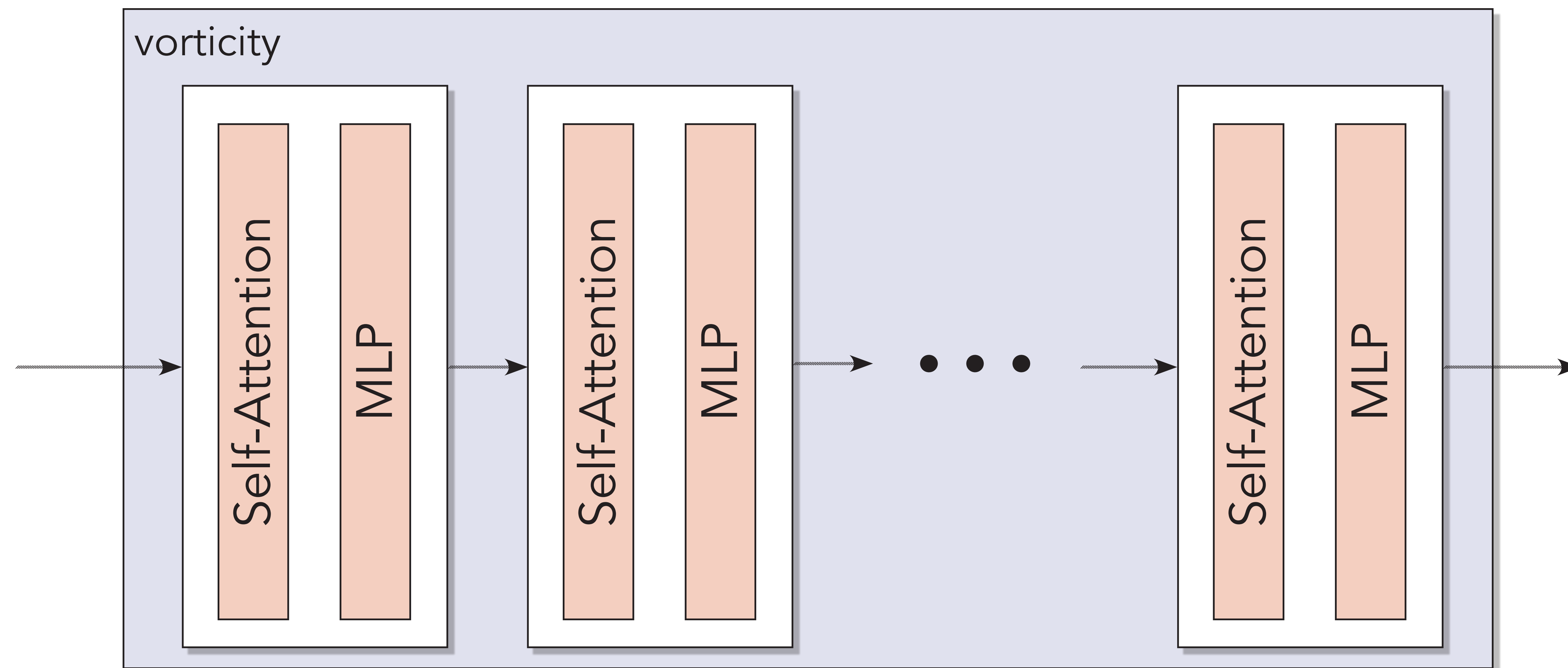
$$\sigma(Q K^T) V$$



Multiformer

Self
attention

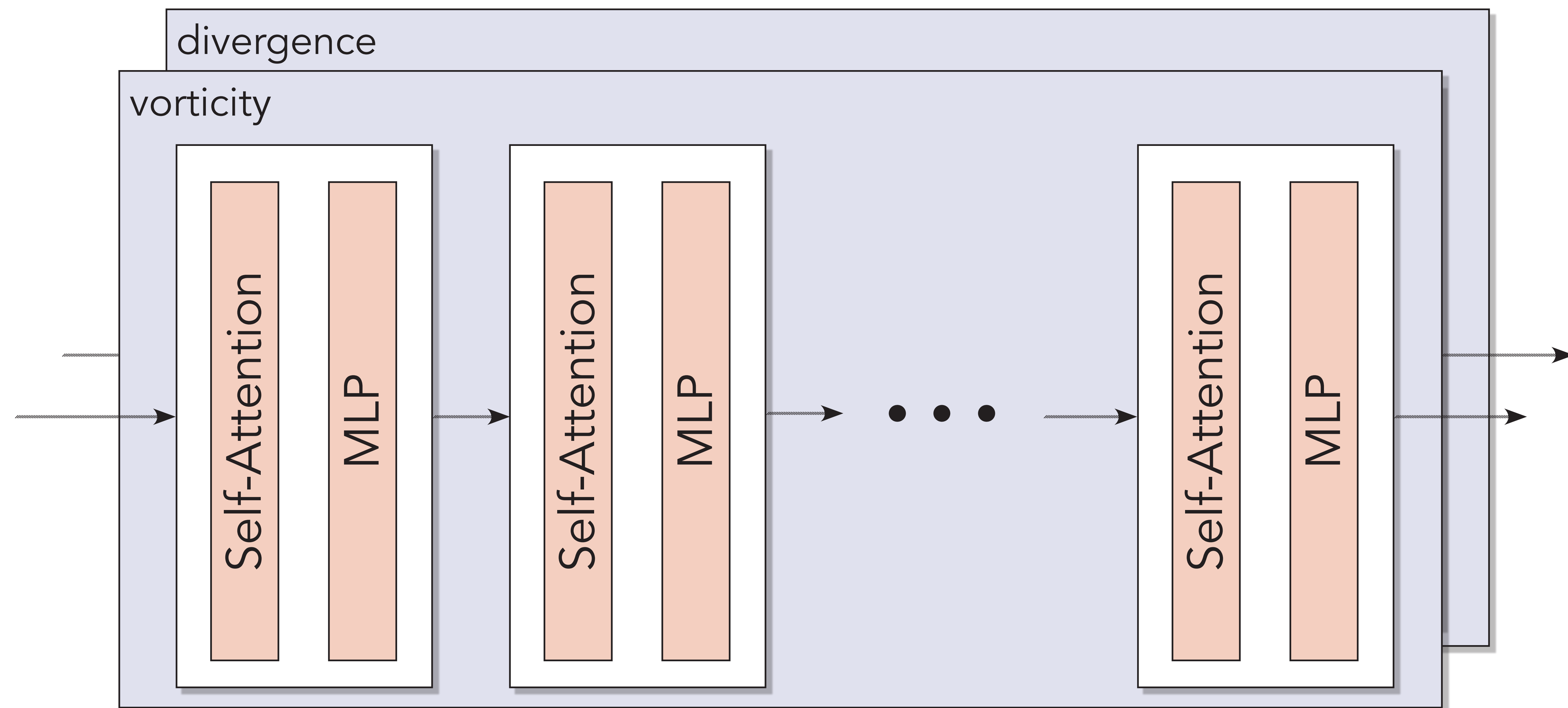
$$\sigma(Q K^T) V$$



Multiformer

Self
attention

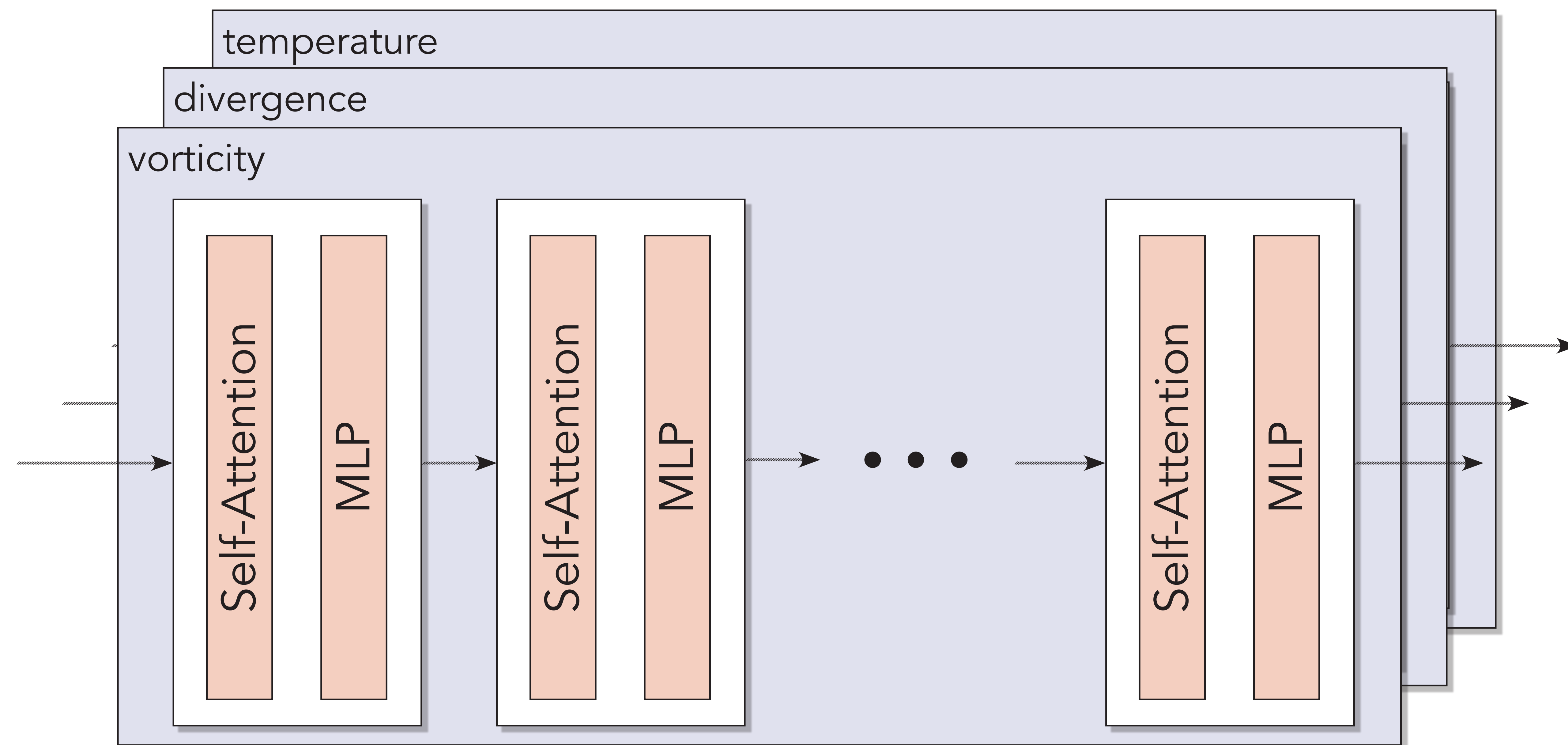
$$\sigma(Q K^T) V$$



Multiformer

Self
attention

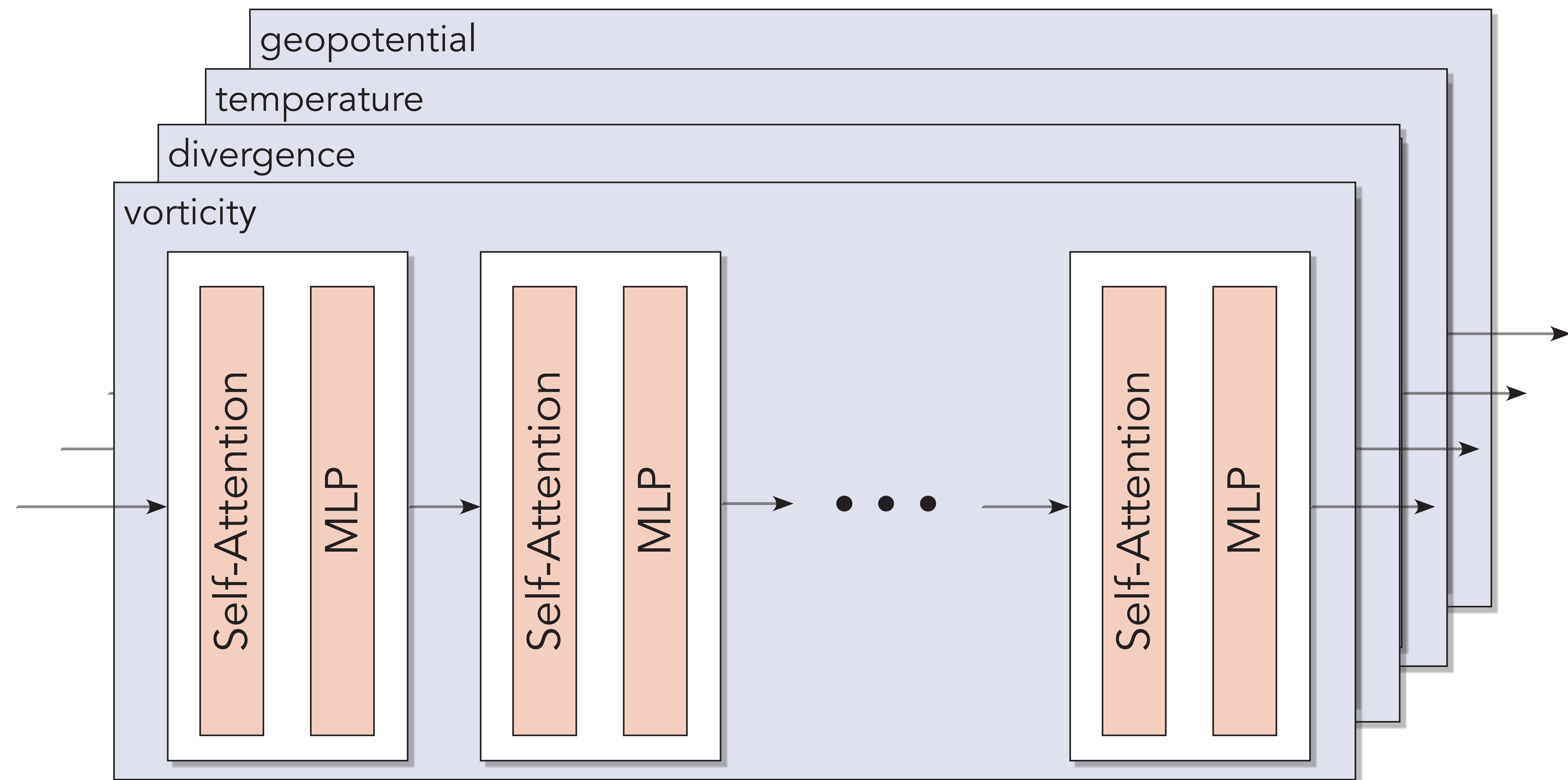
$$\sigma(Q K^T) V$$



Multiformer

Self
attention

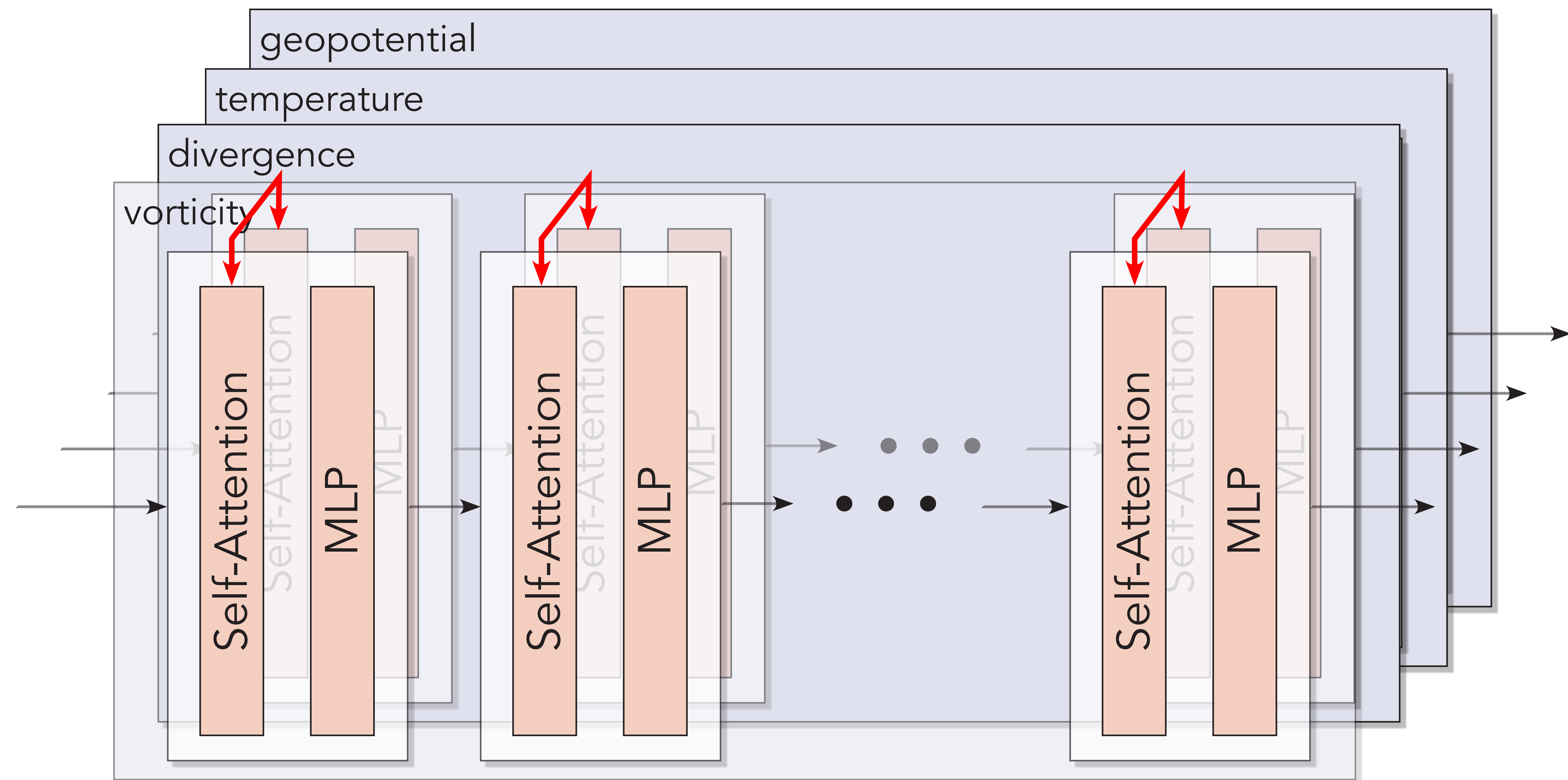
$$\sigma(Q K^T) V$$



Multiformer

Self
attention

$$\sigma(Q K^T) V$$



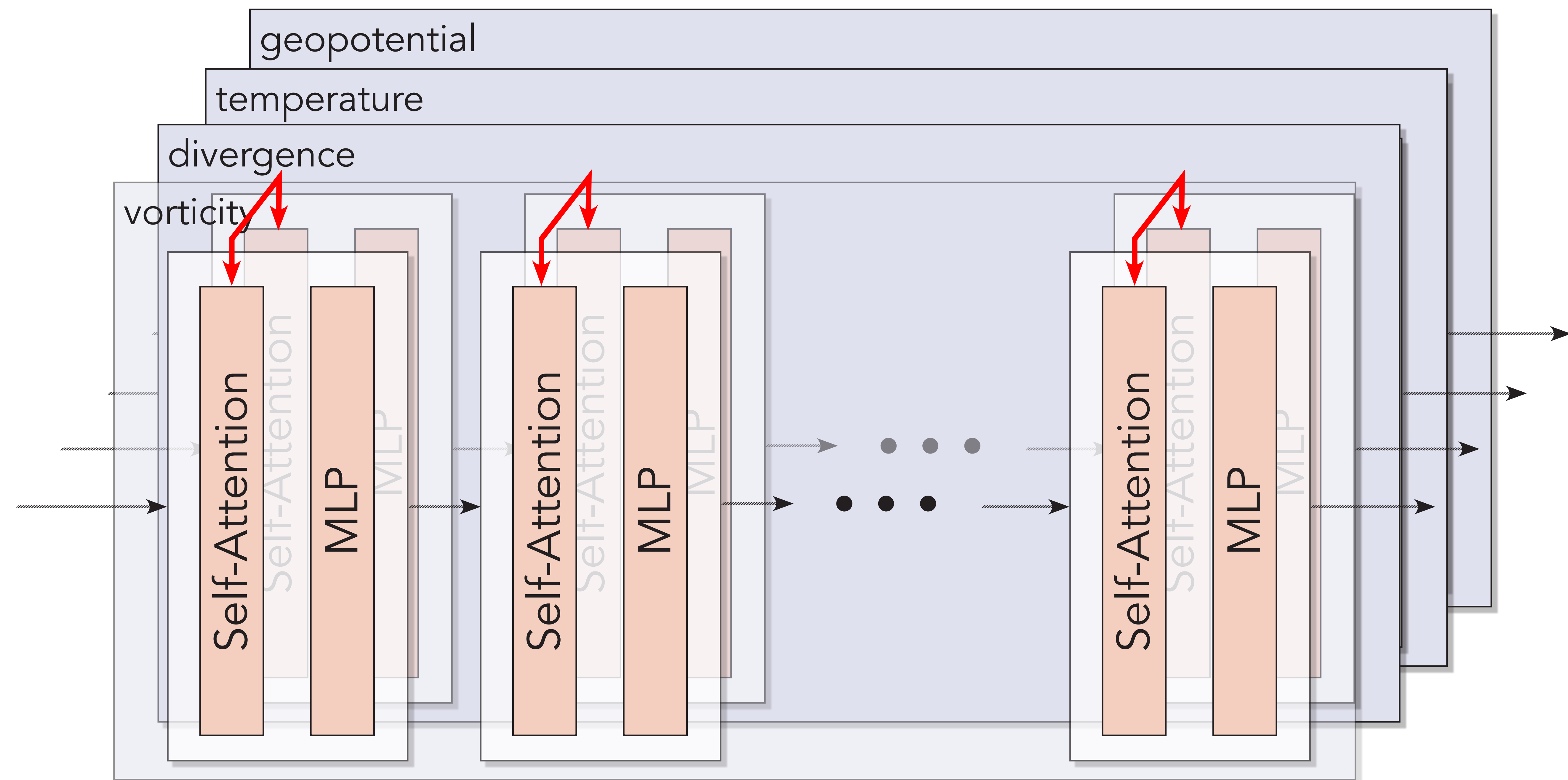
Multiformer

Self
attention

$$\sigma(Q K^T) V$$

Cross
attention

$$\sigma(Q_{\zeta} K_{\mu}^T) V_{\mu}$$



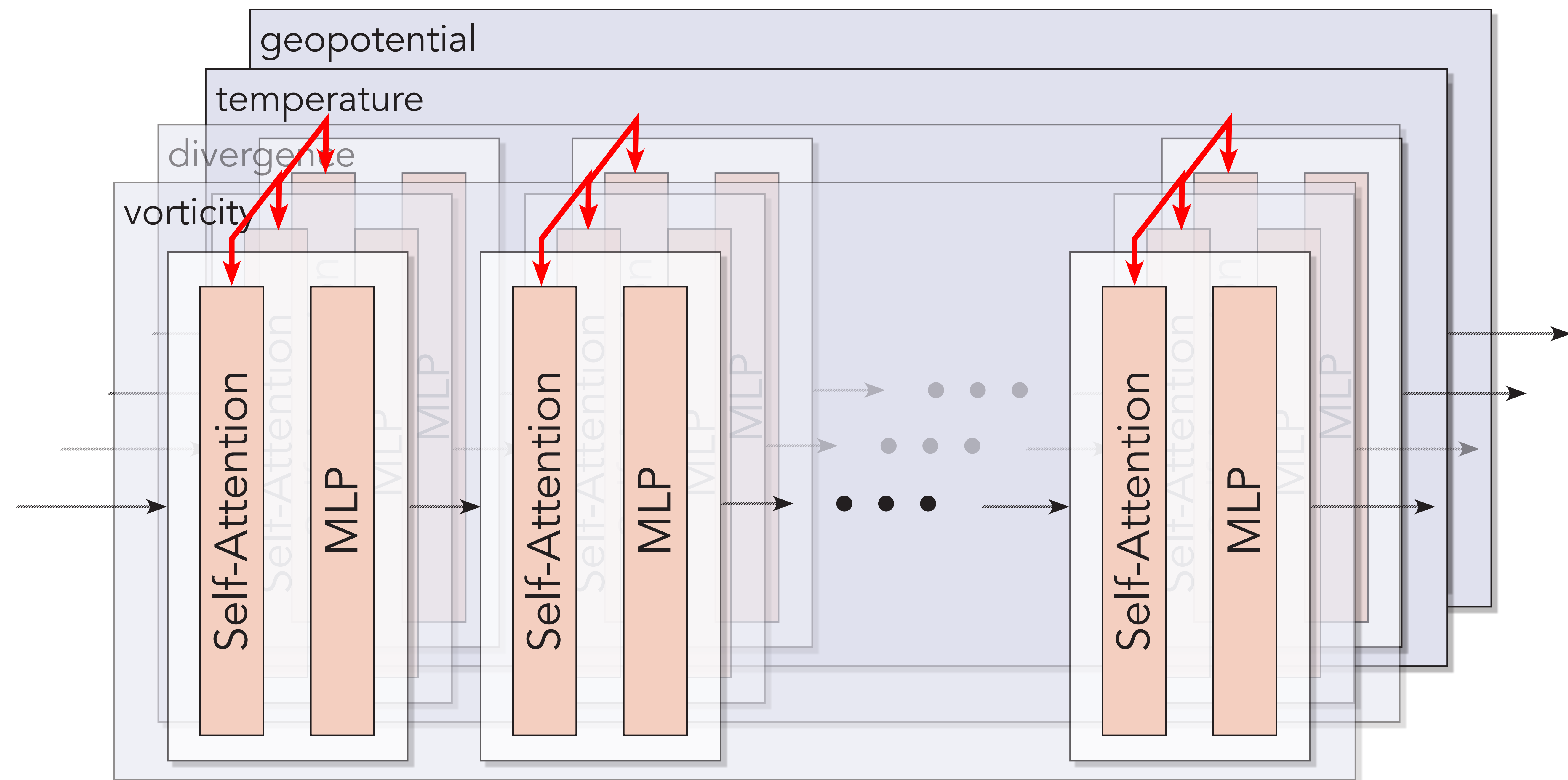
Multiformer

Self
attention

$$\sigma(Q K^T) V$$

Cross
attention

$$\sigma(Q_{\zeta} K_{\mu}^T) V_{\mu}$$



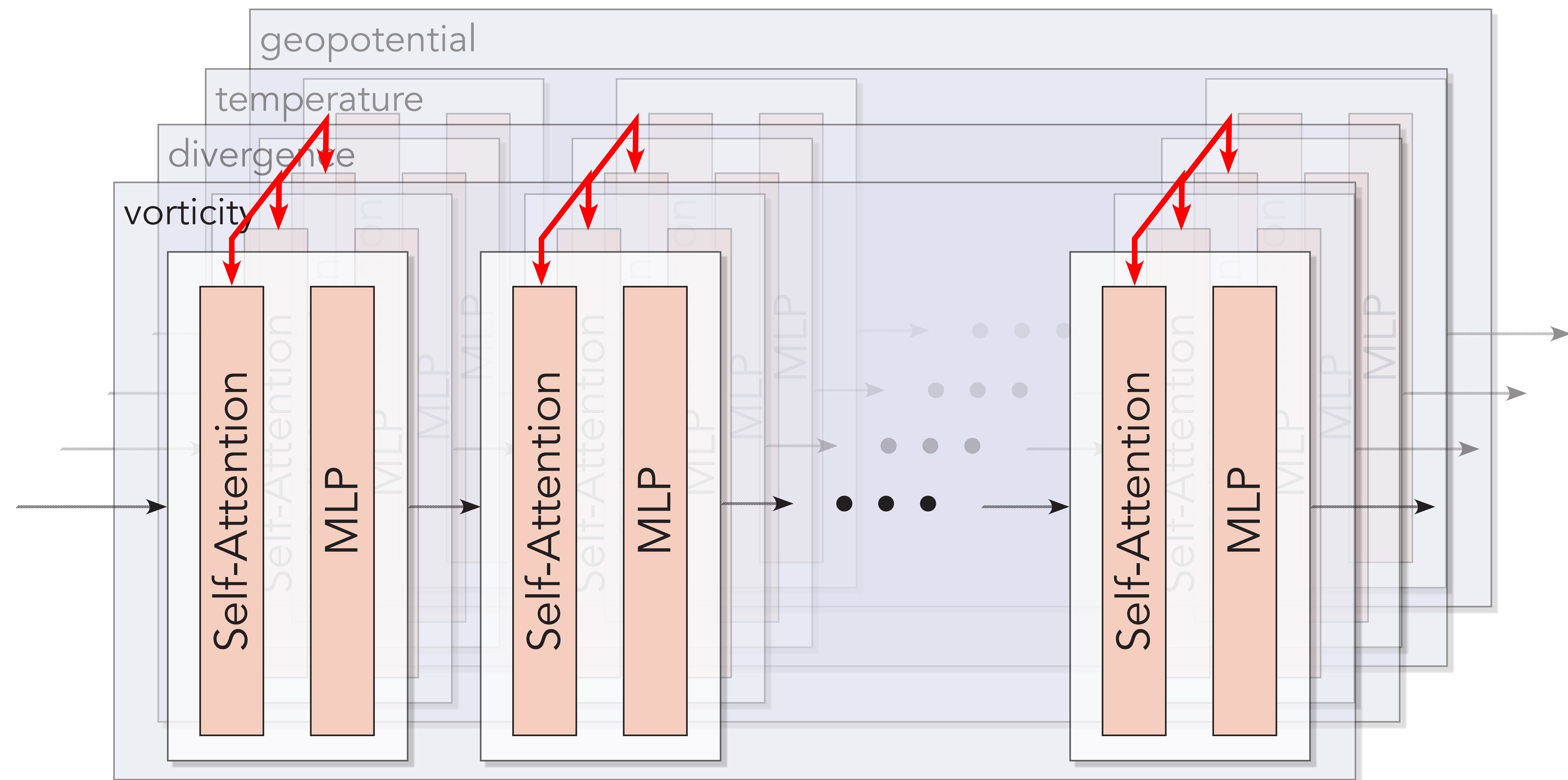
Multiformer

Self
attention

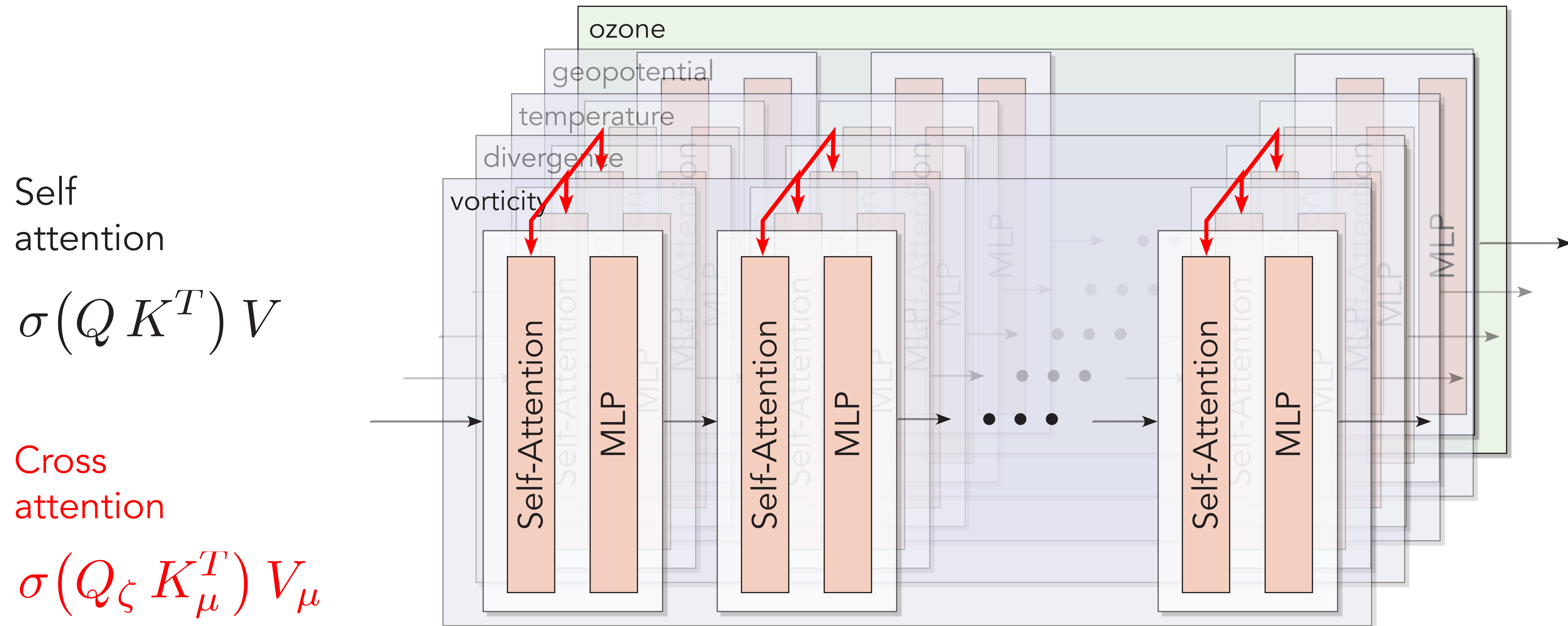
$$\sigma(Q K^T) V$$

Cross
attention

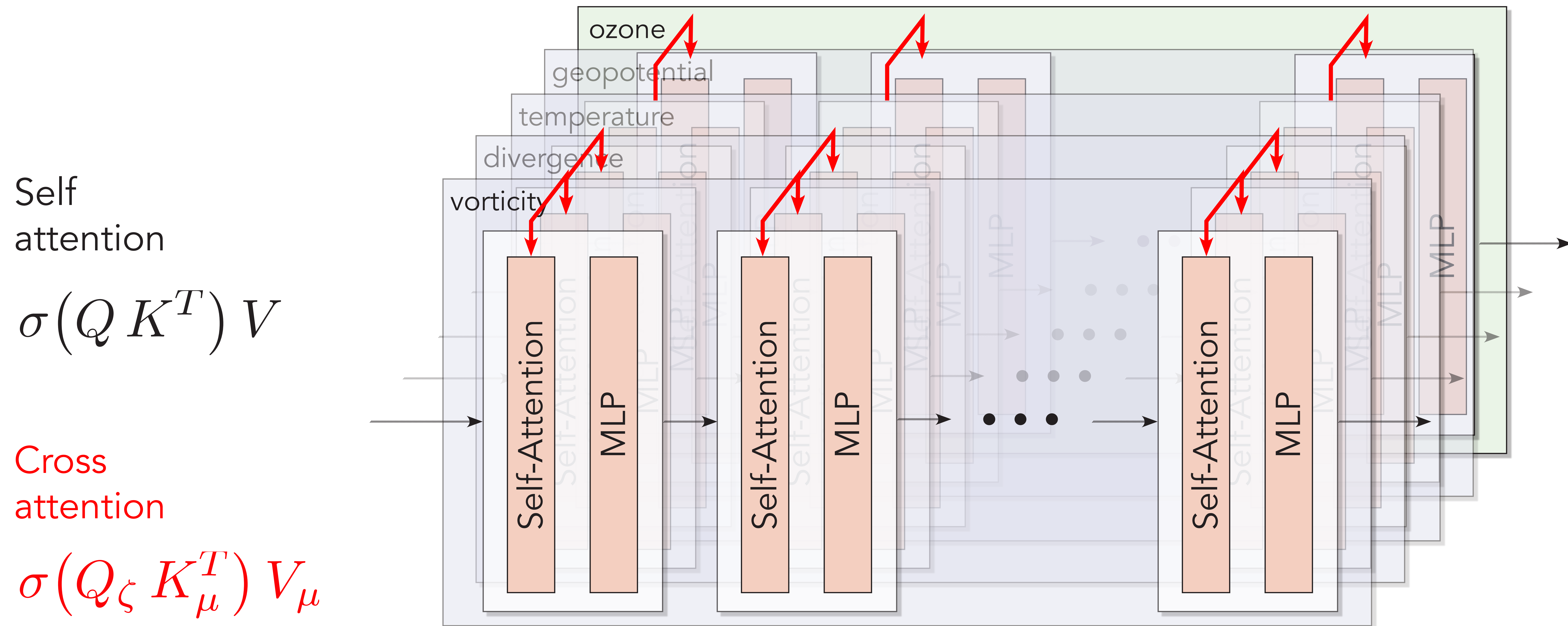
$$\sigma(Q_{\zeta} K_{\mu}^T) V_{\mu}$$



Multiformer



Multiformer



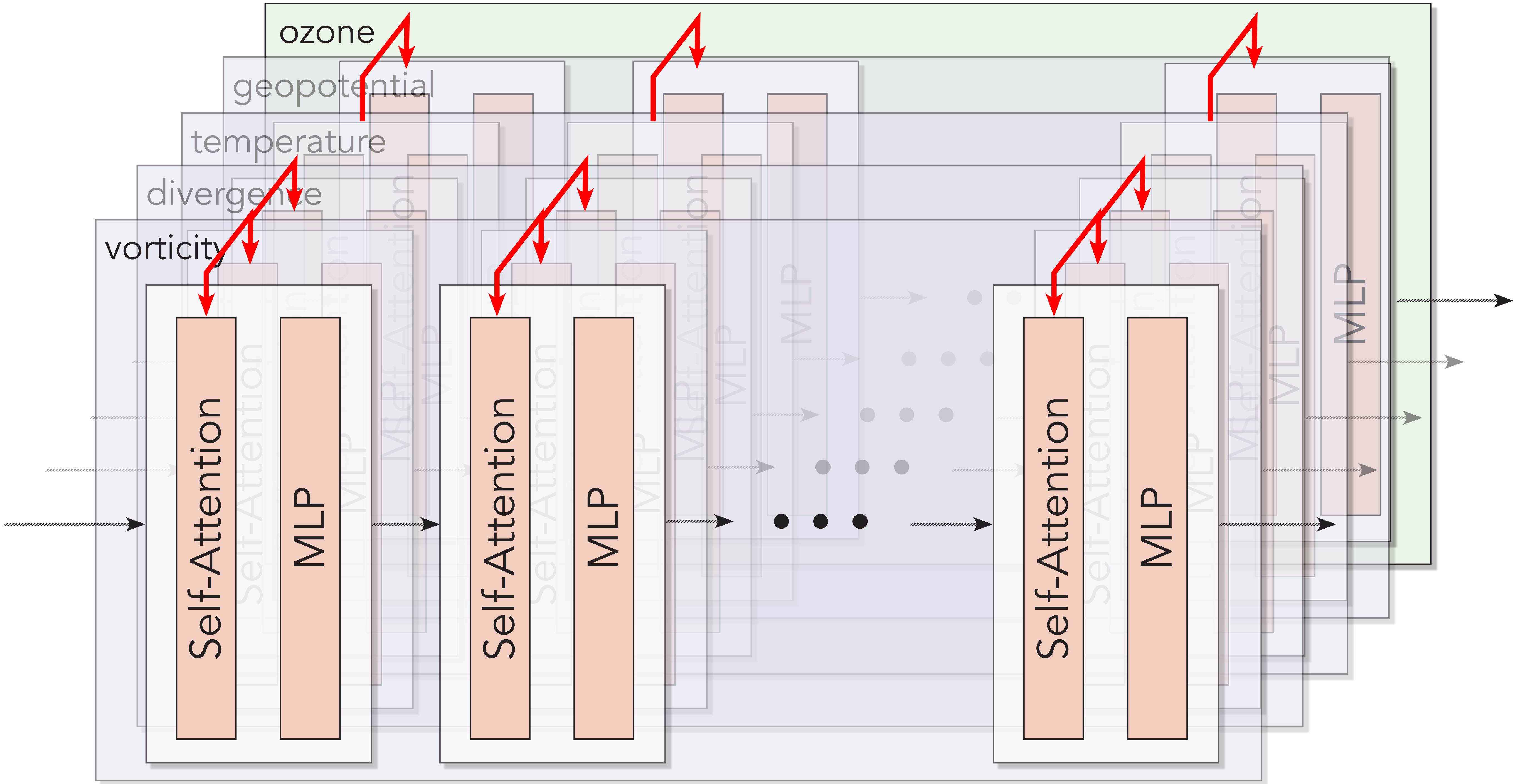
Multiformer

Self attention

$$\sigma(Q K^T) V$$

Cross attention

$$\sigma(Q_{\zeta} K_{\mu}^T) V_{\mu}$$



Multiformer

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

Data: ERA5 reanalysis

721x1440 horizontal grid (0.25 degree)

137 vertical layers

over 6 PB of data readily amenable to machine learning

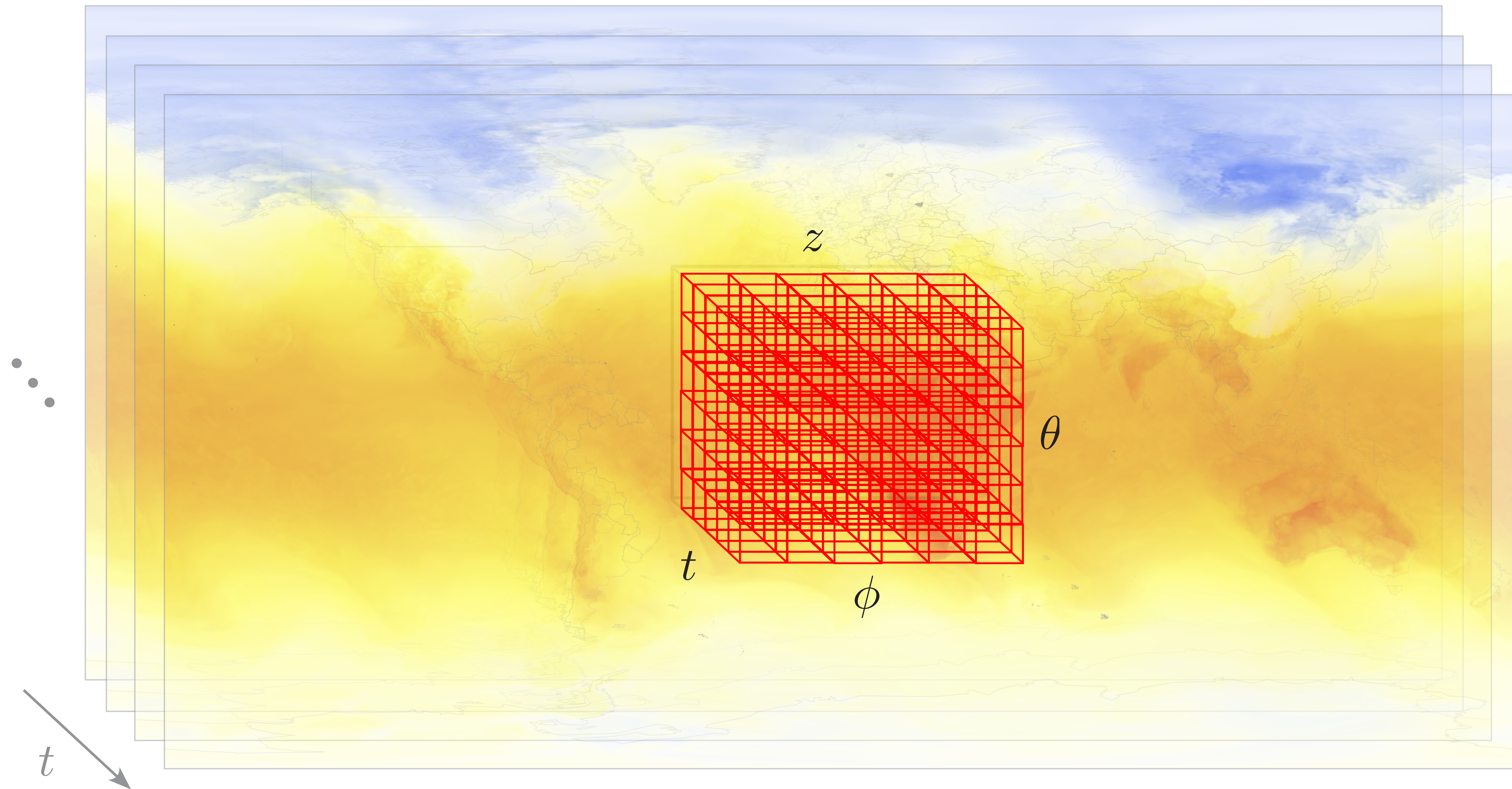
hourly for 70 years

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

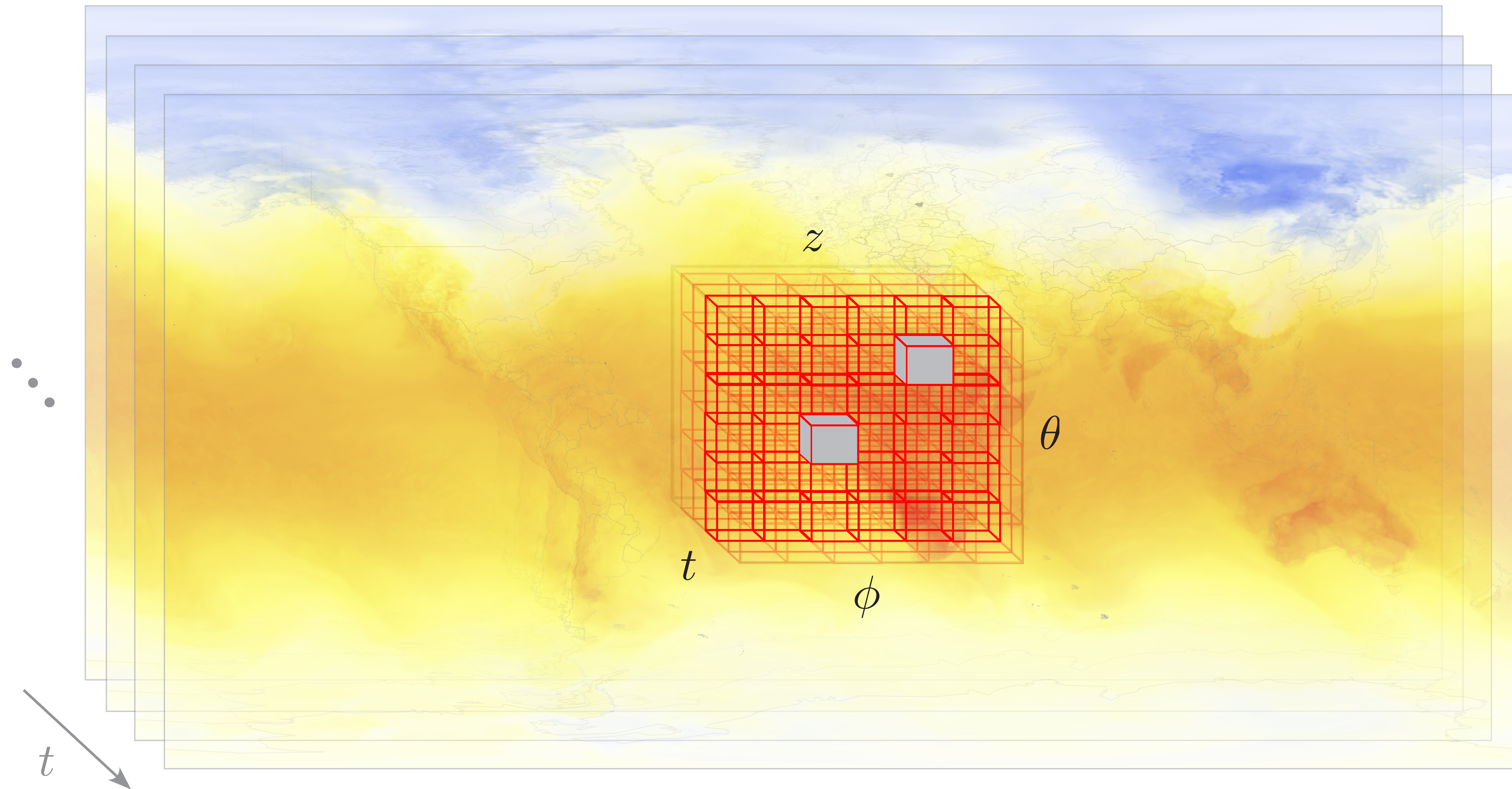
Spatio-temporal BERT

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

Spatio-temporal BERT

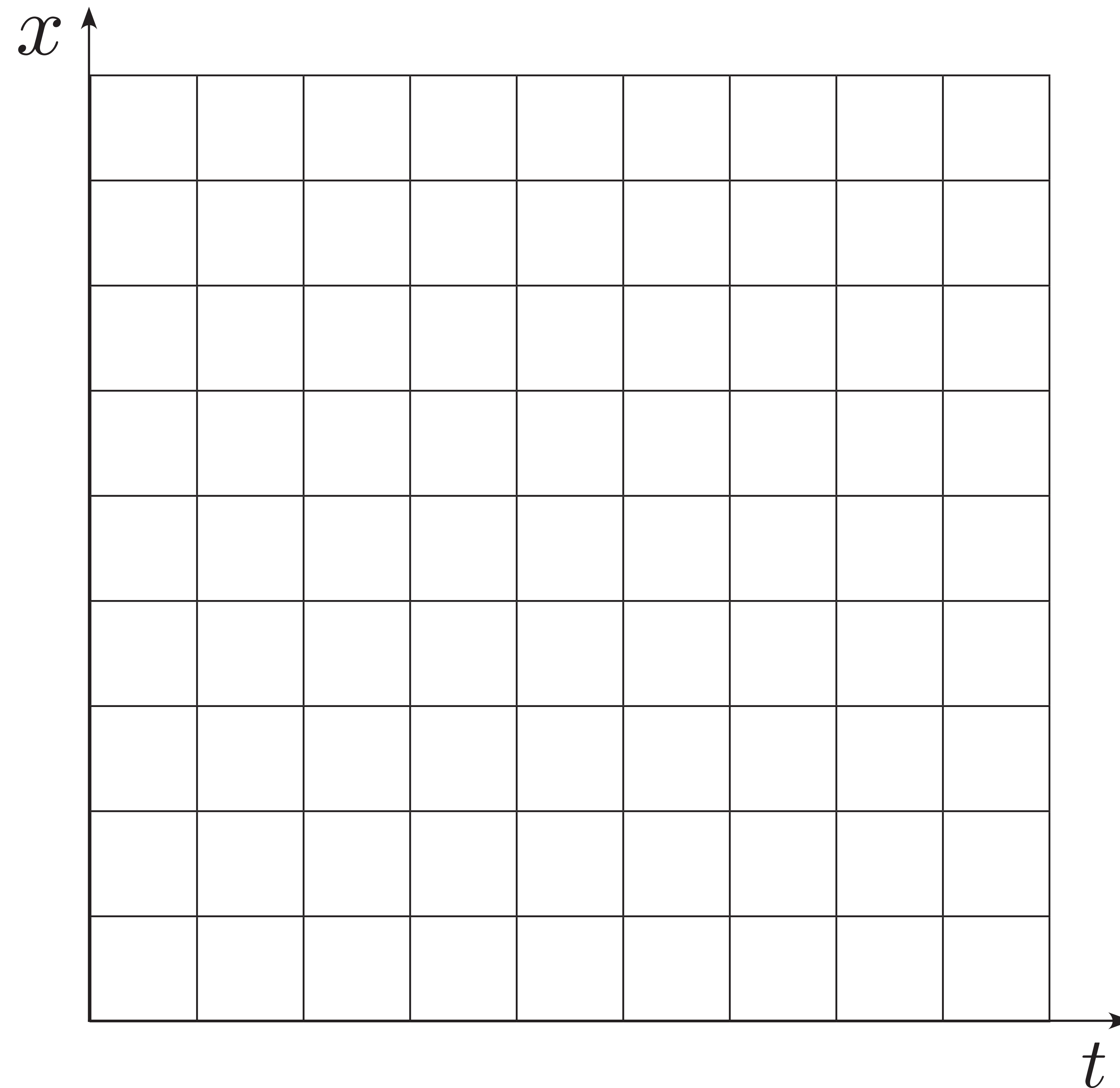


Spatio-temporal BERT



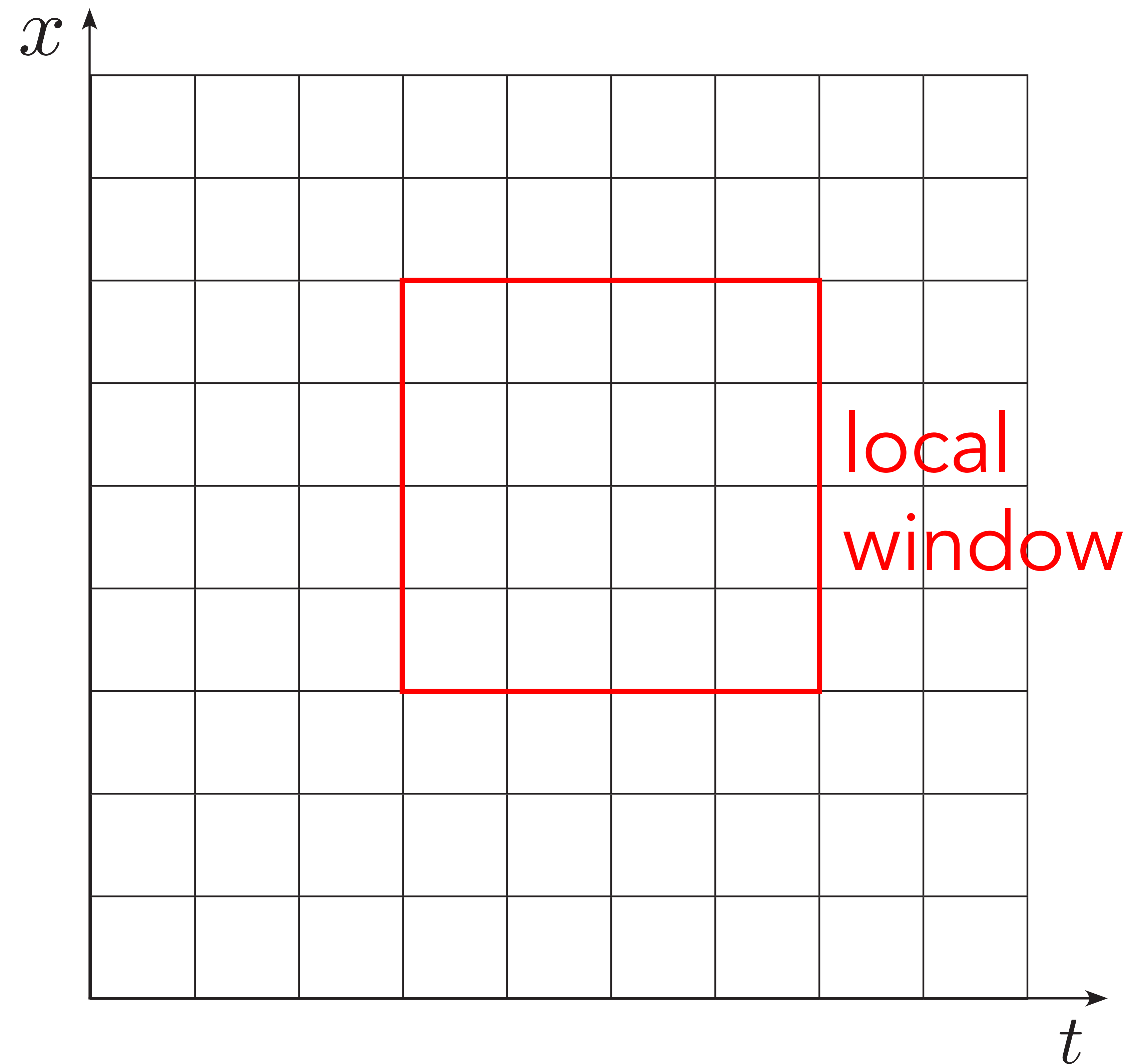
Spatio-temporal BERT

Flatland
view



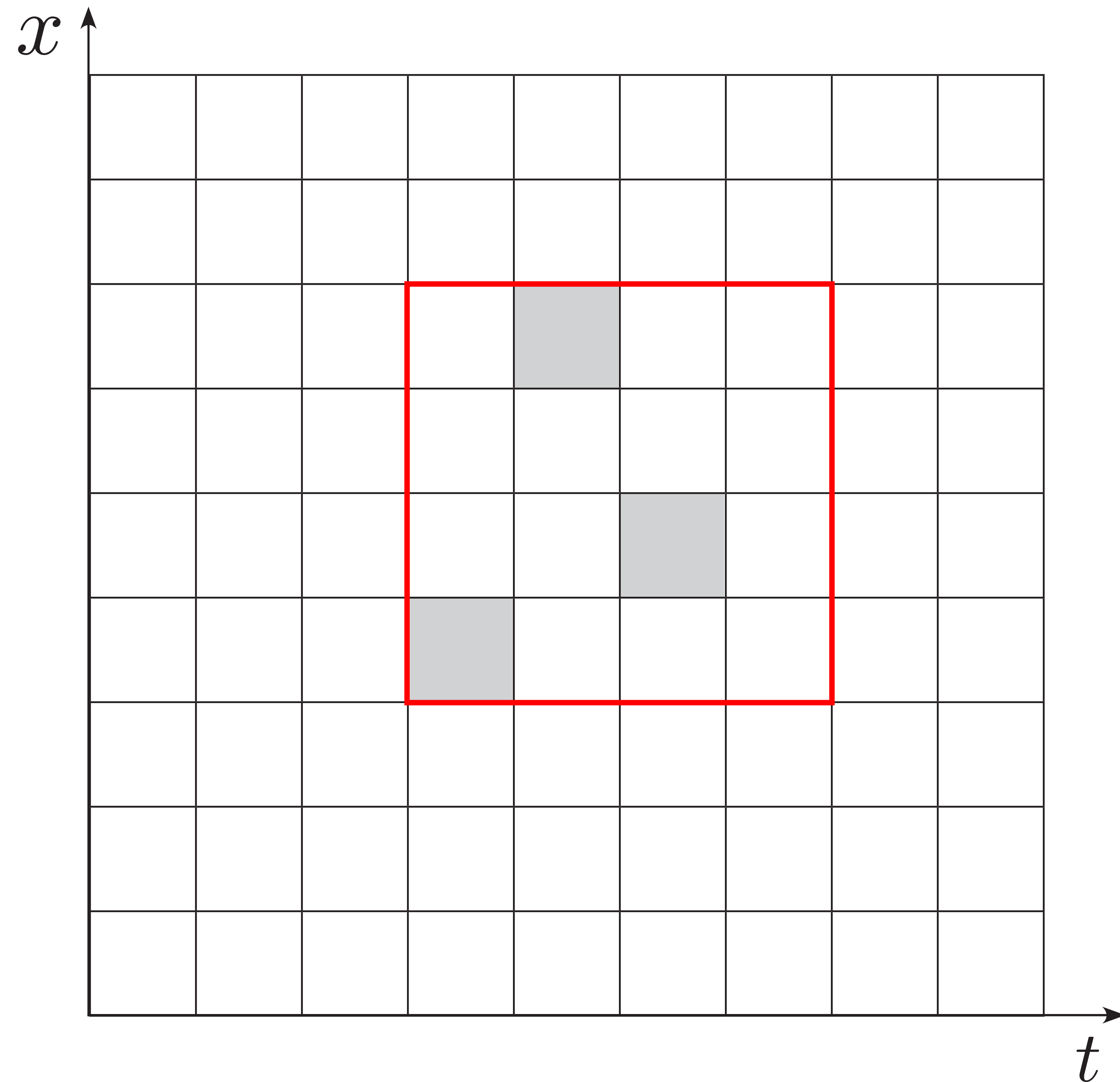
Spatio-temporal BERT

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
 - › Random masking and distortions (noising, coarsening) ensures that a probabilistic model is learned

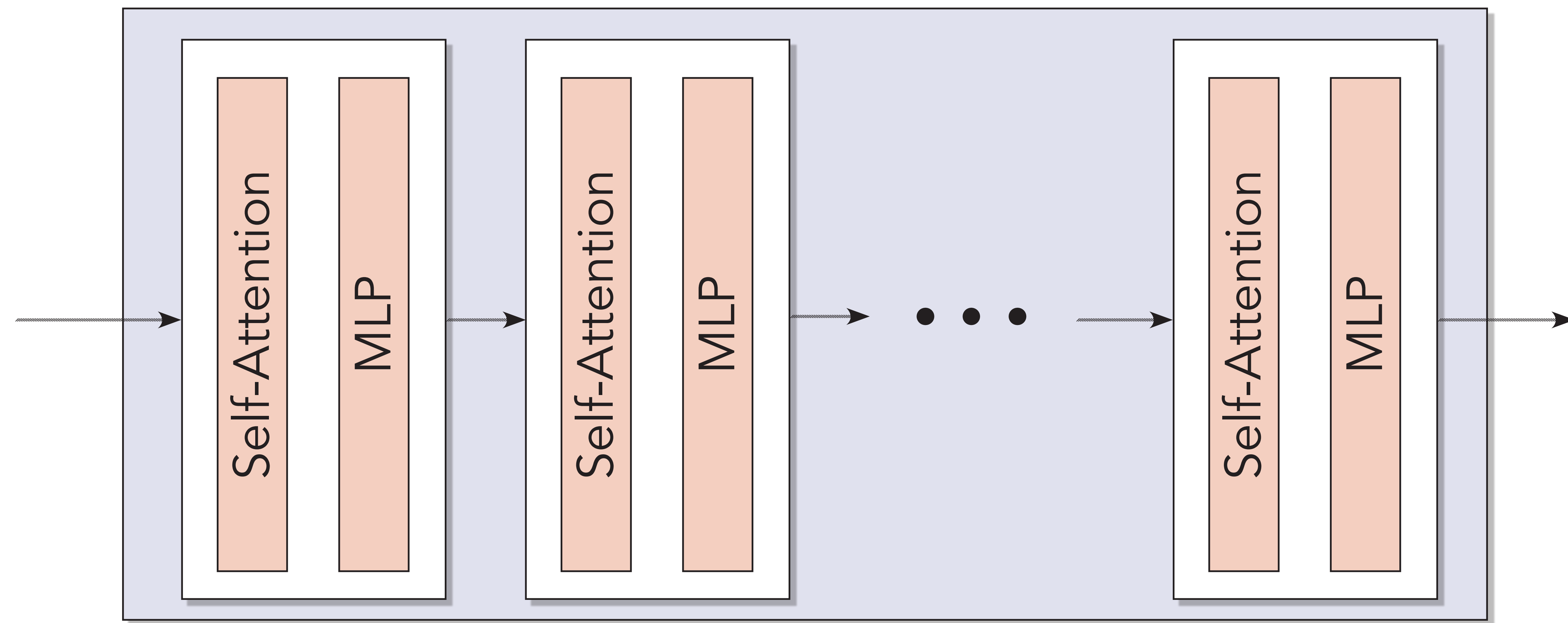
Statistical loss

- Machine learning: Training on MSE loss is problematic in terms of training dynamics
 - › One reason for overly smooth predictions

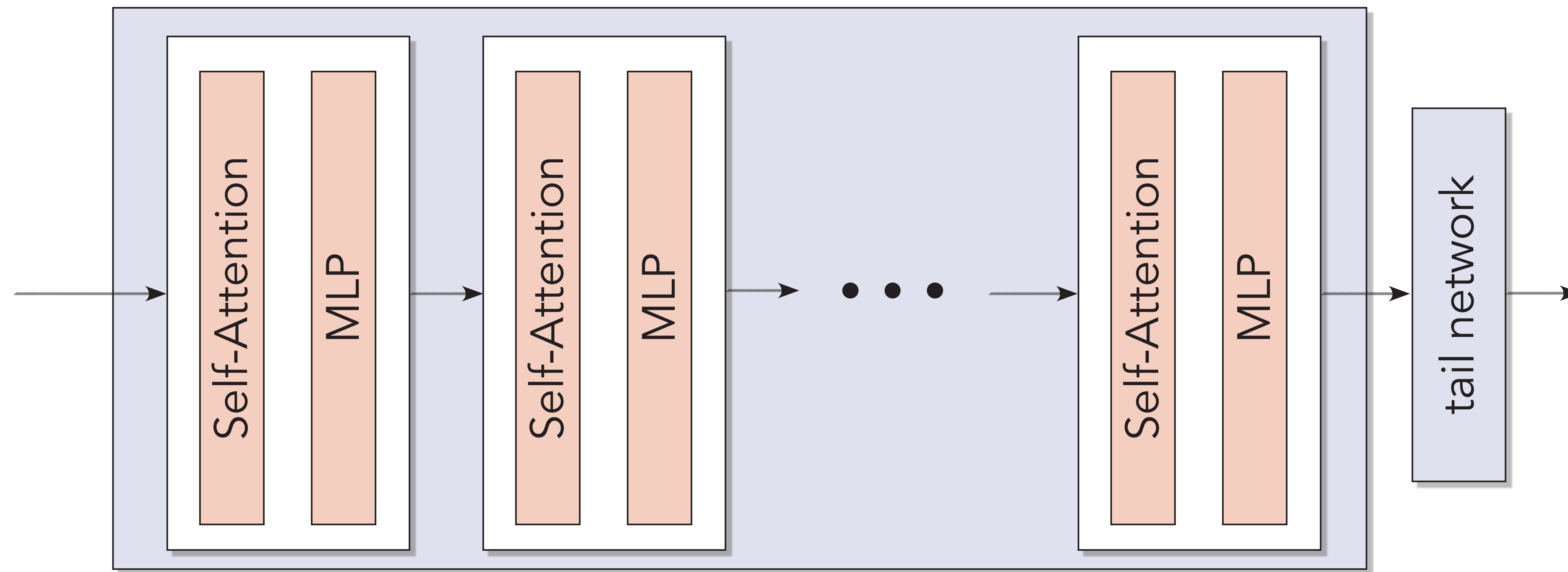
Statistical loss

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

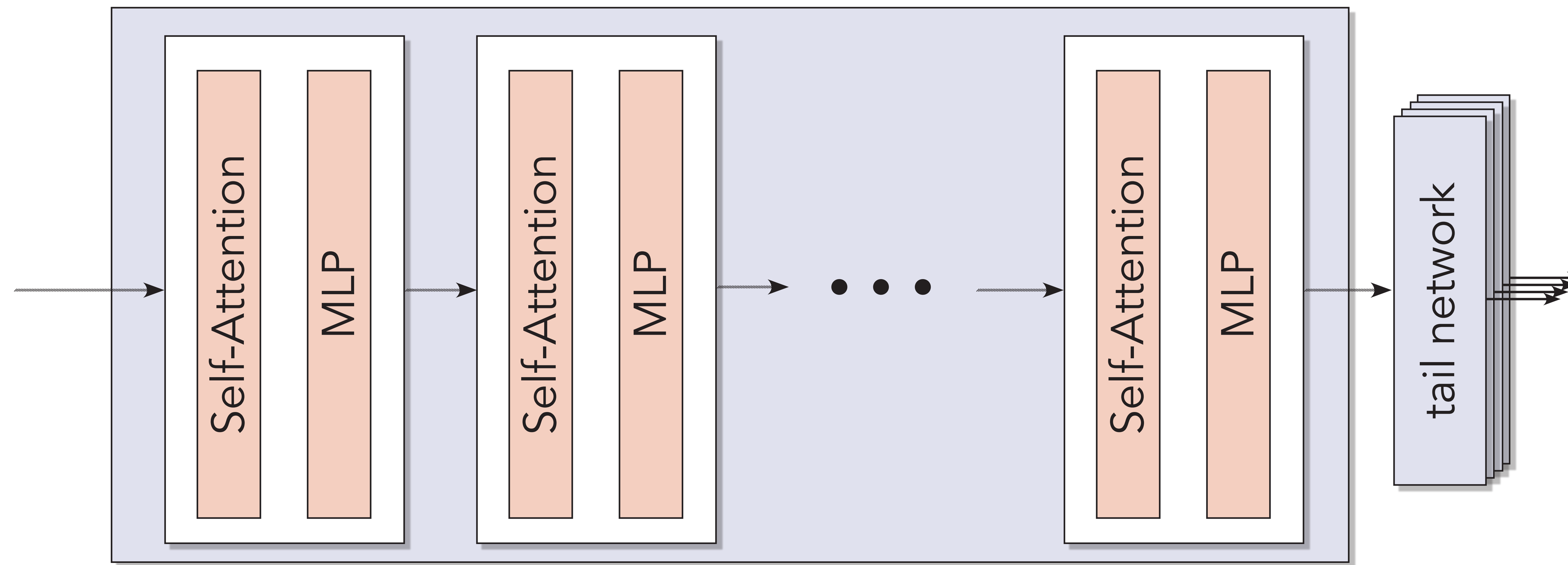
Statistical loss



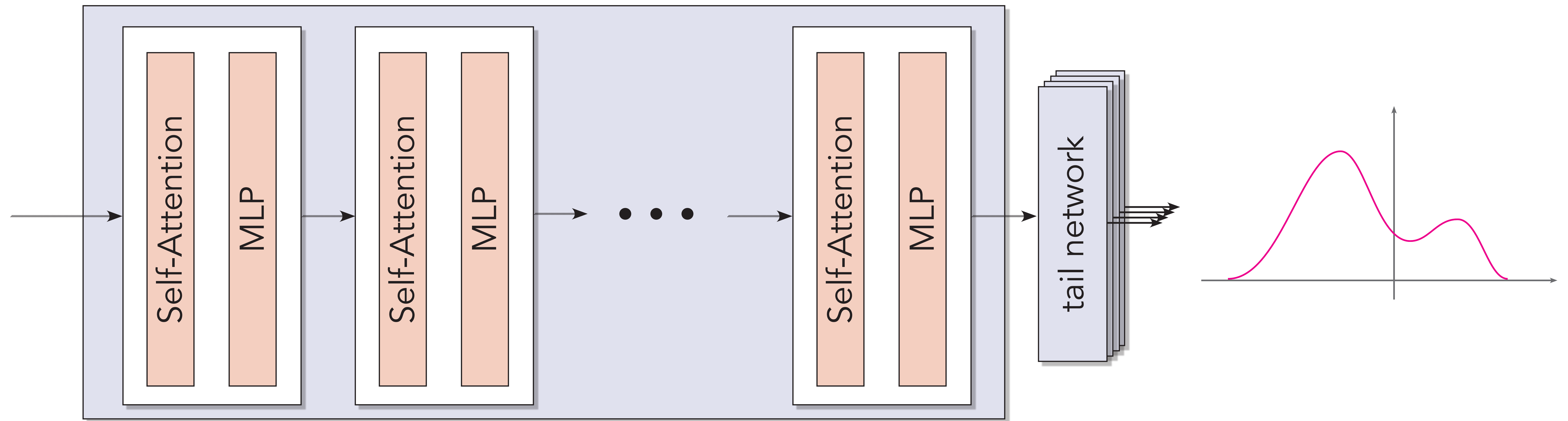
Statistical loss



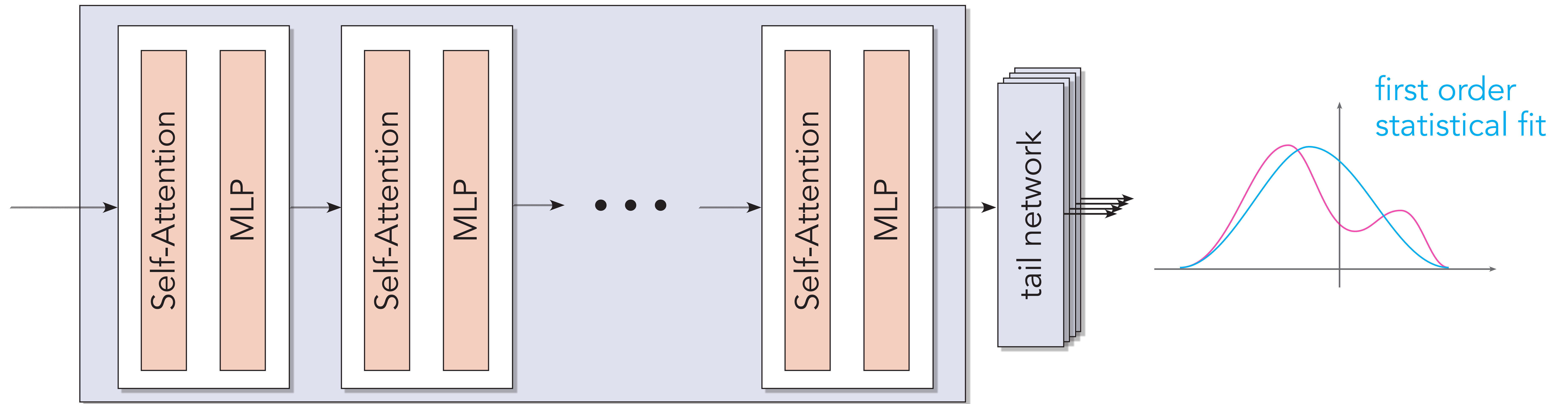
Statistical loss



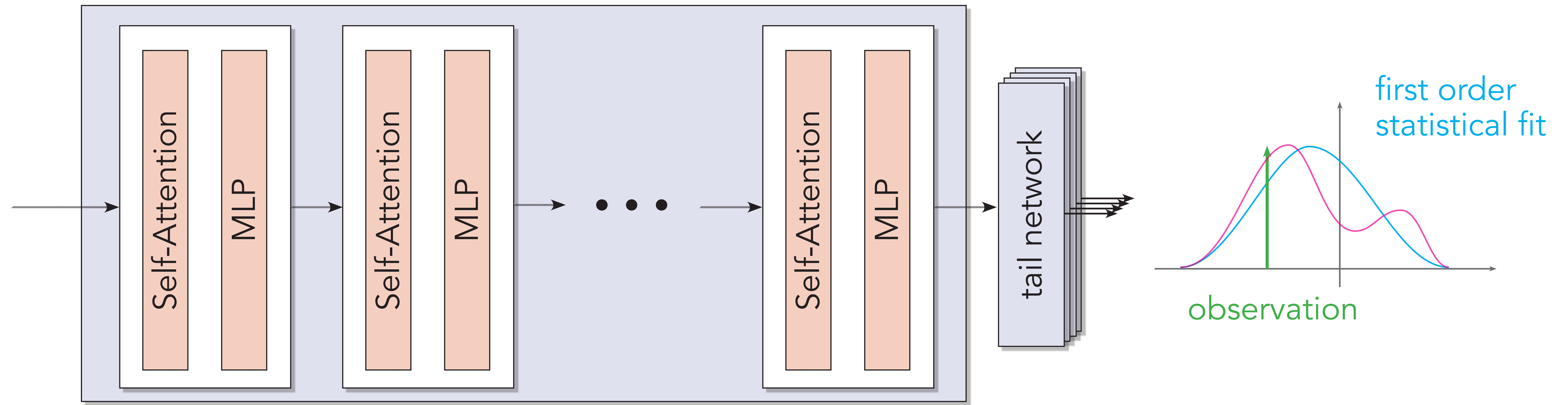
Statistical loss



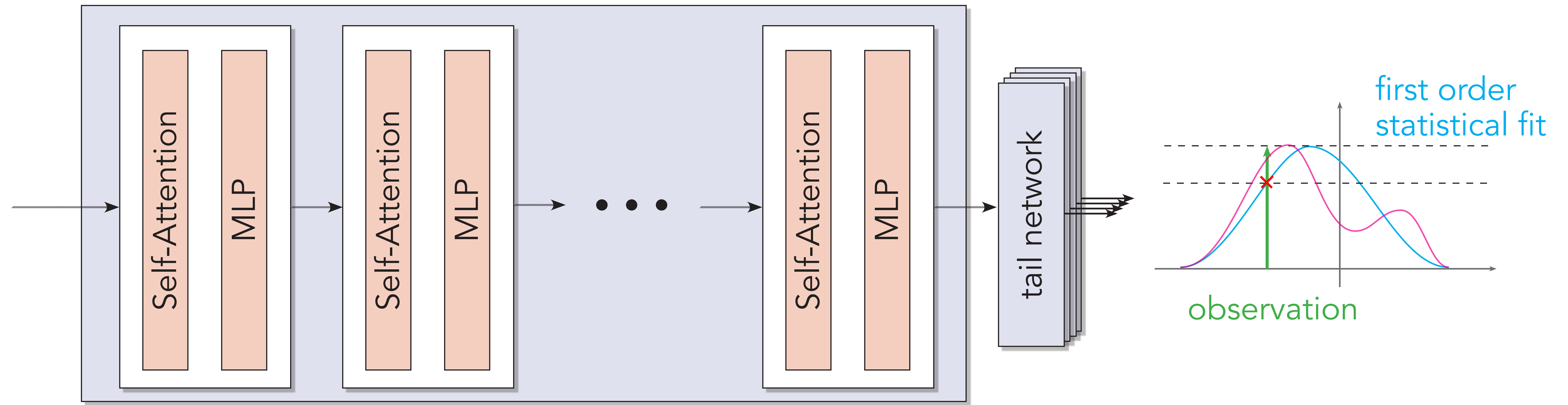
Statistical loss



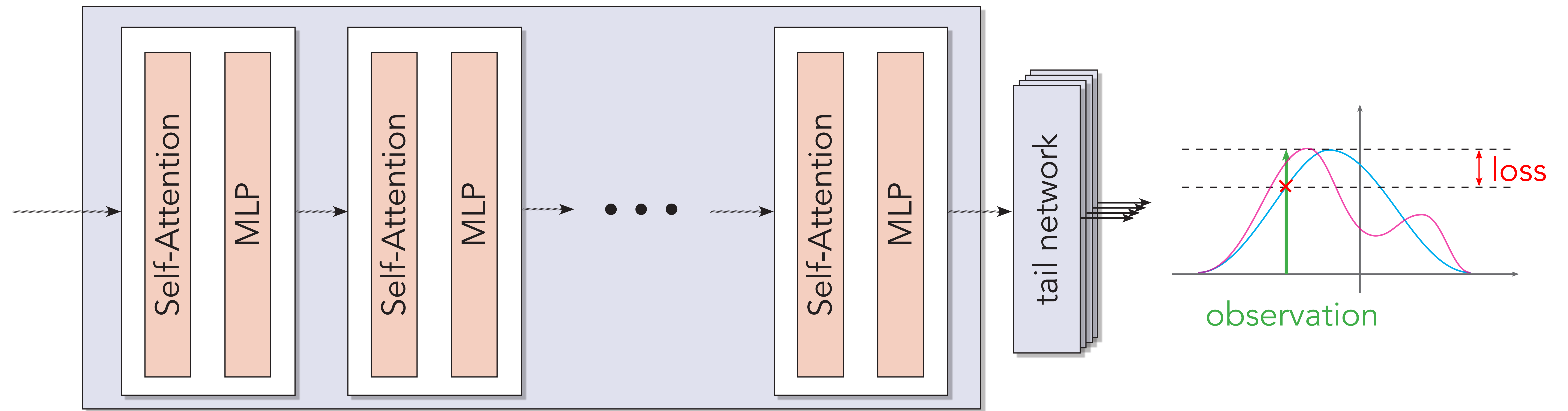
Statistical loss



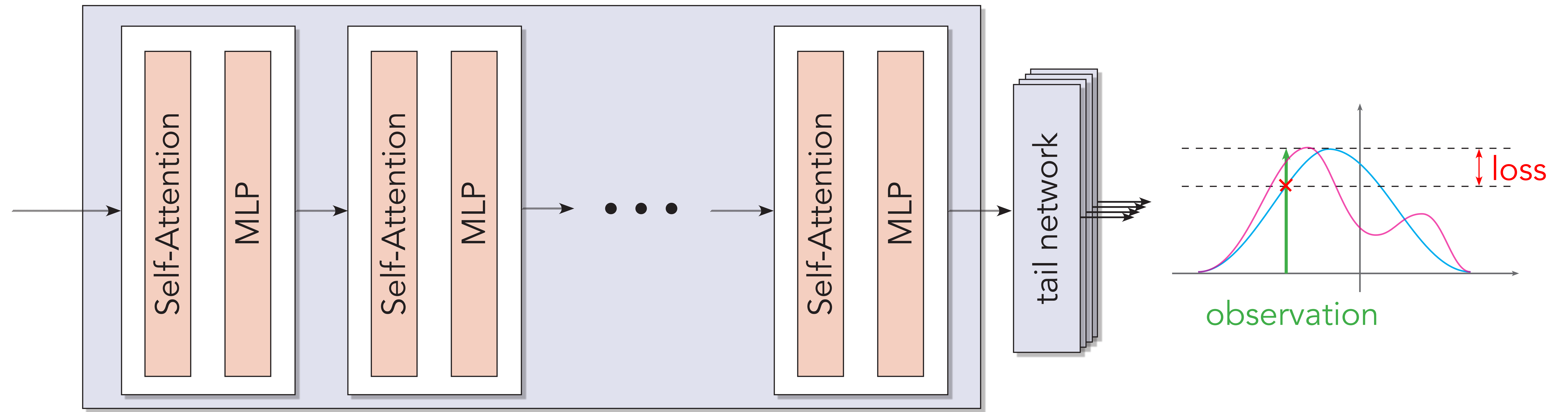
Statistical loss



Statistical loss



Statistical loss

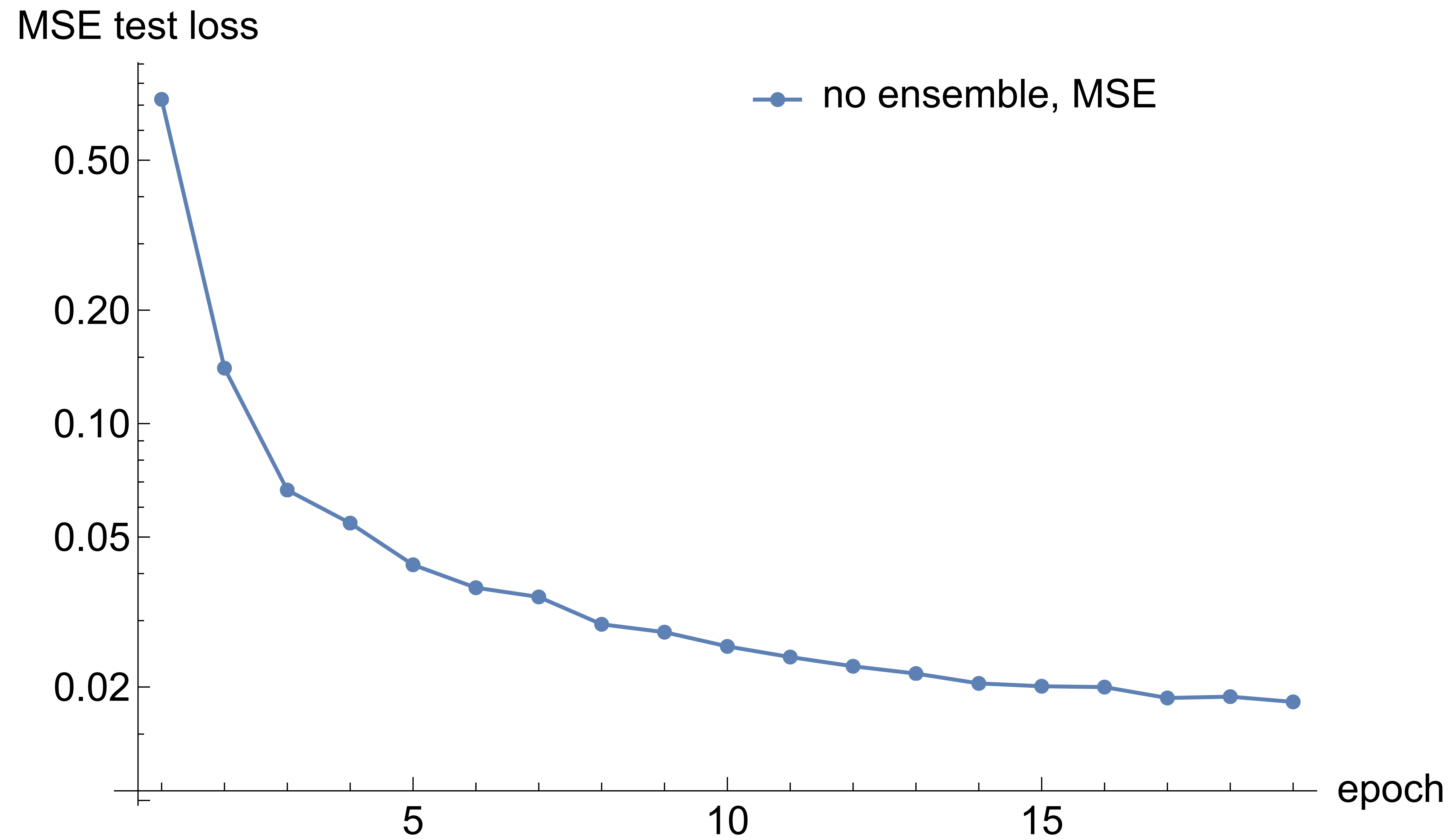


end-to-end training encourages statistical representation

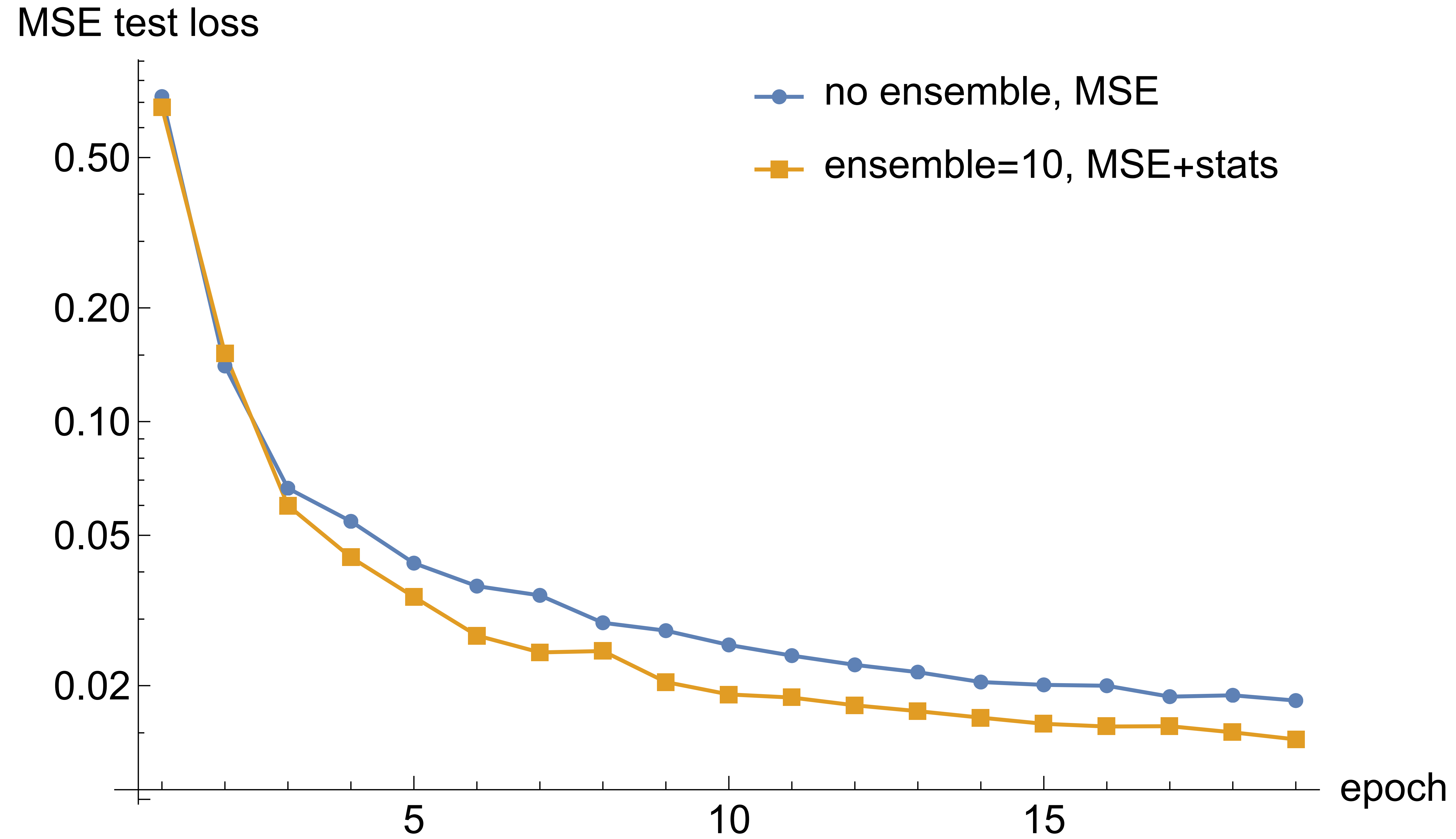
Statistical loss: experiments

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

Statistical loss

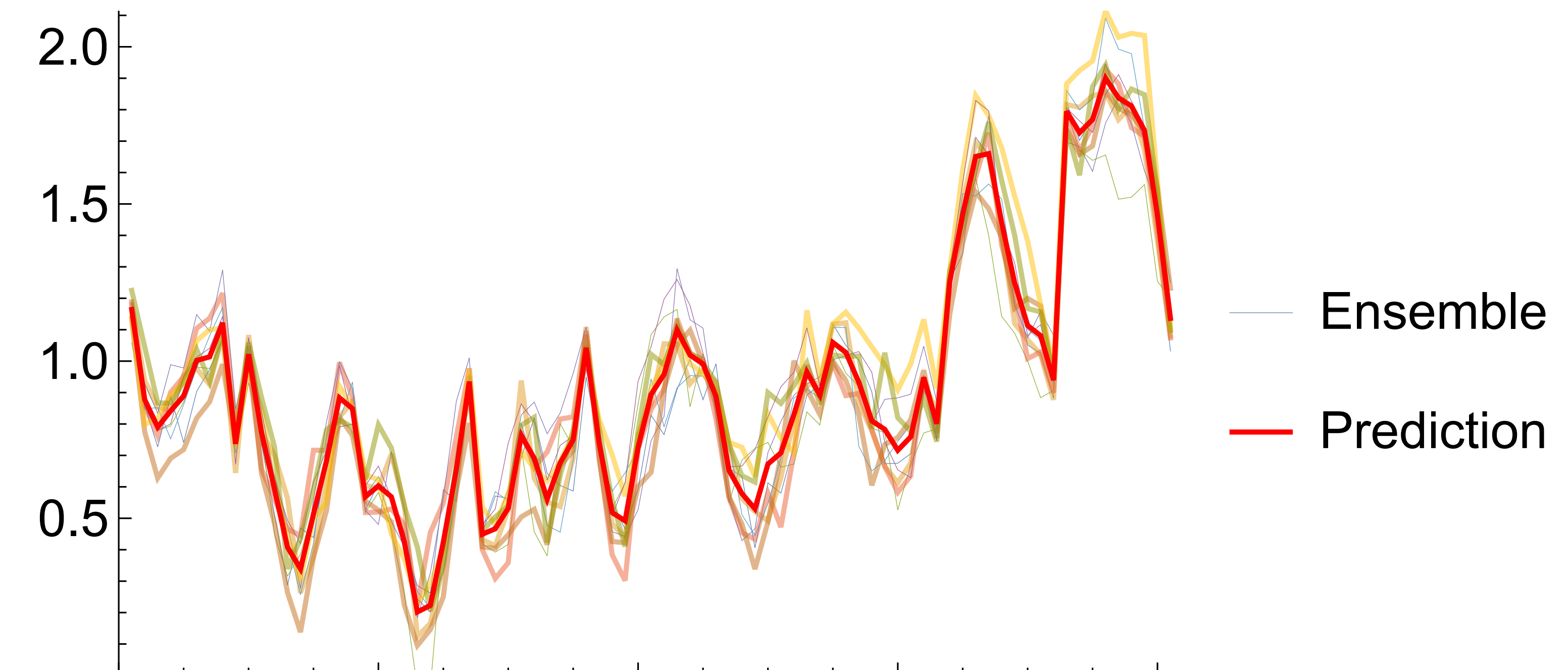
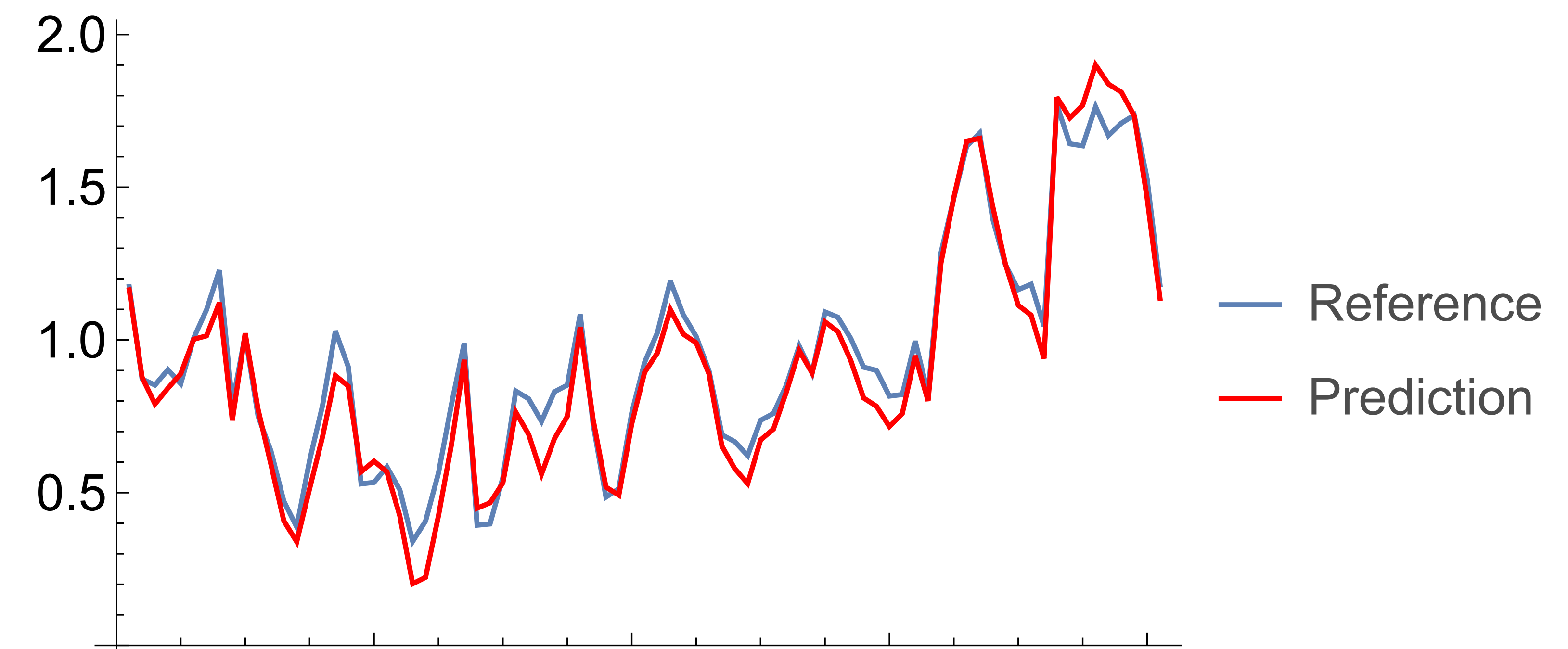
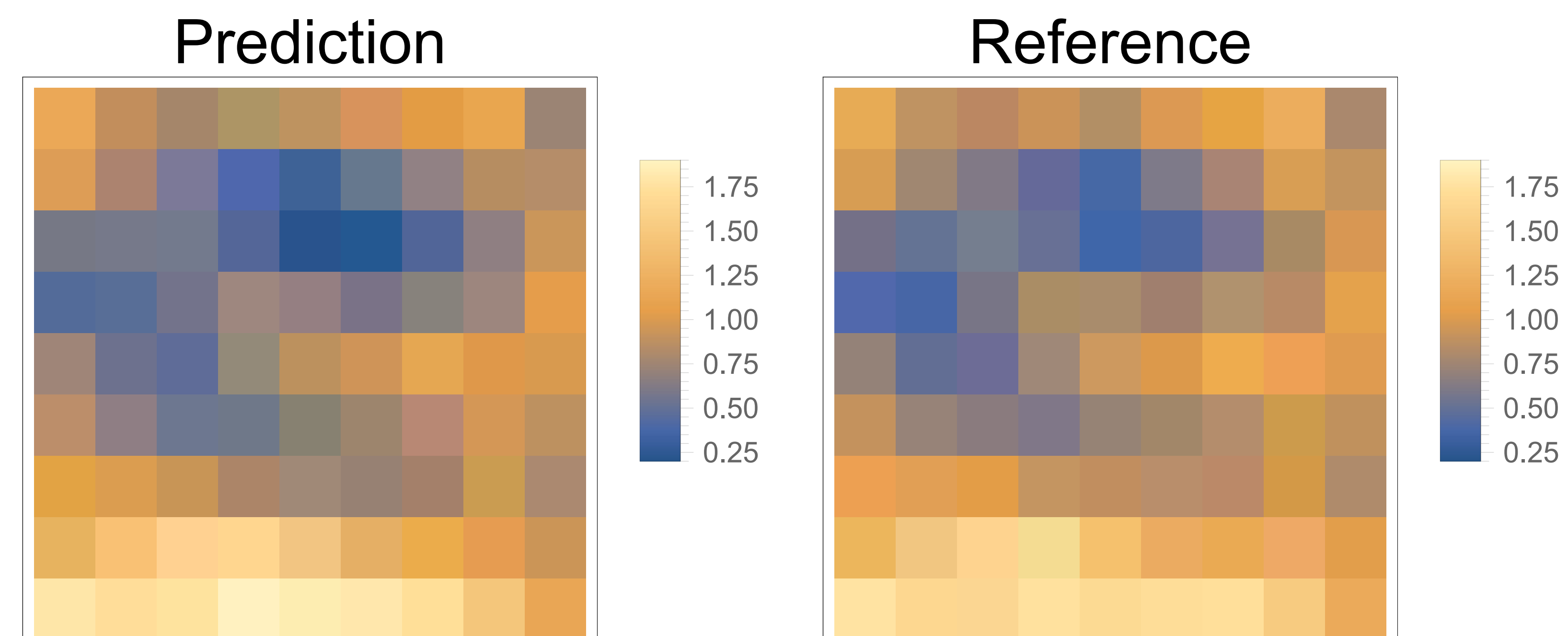


Statistical loss



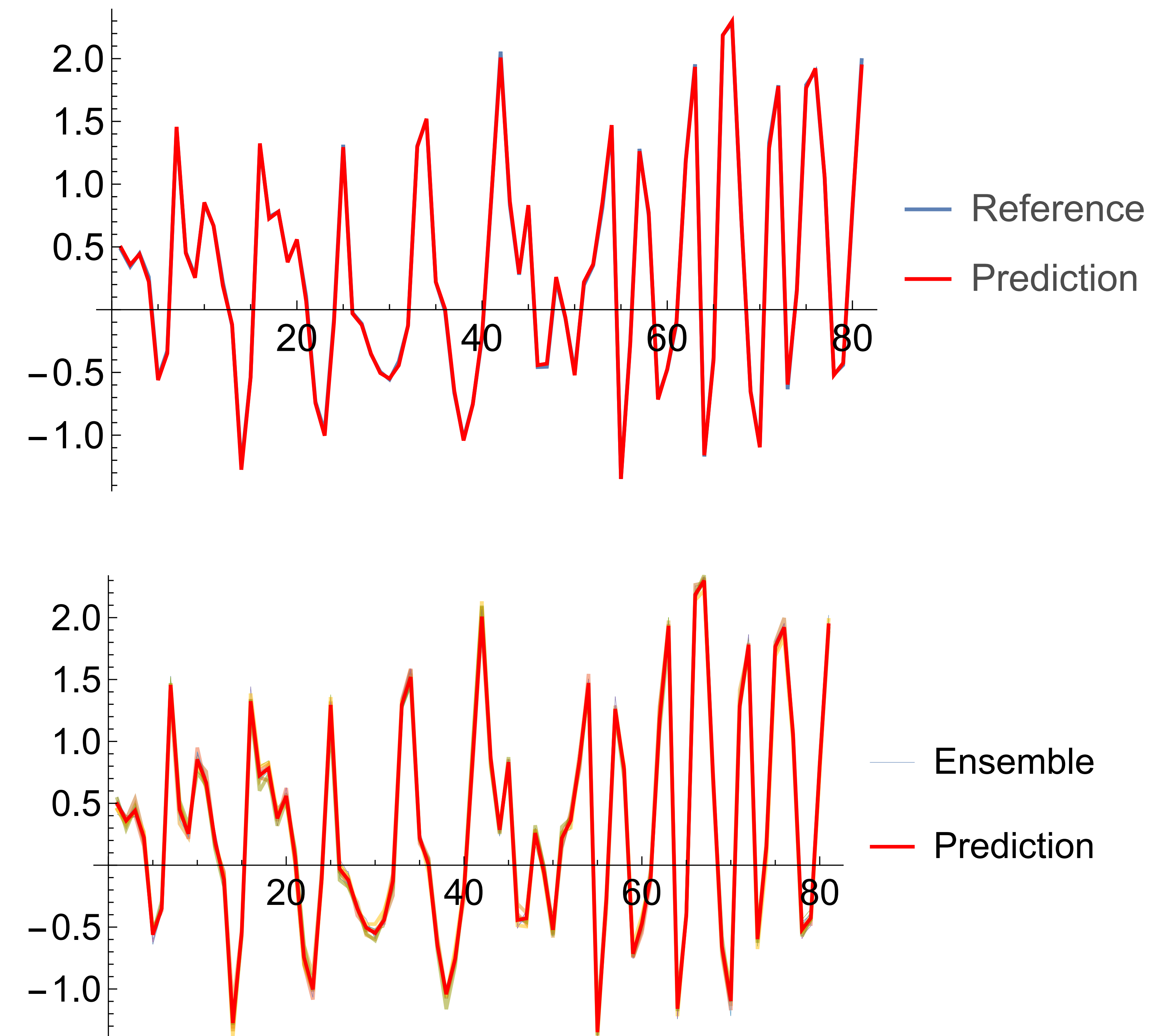
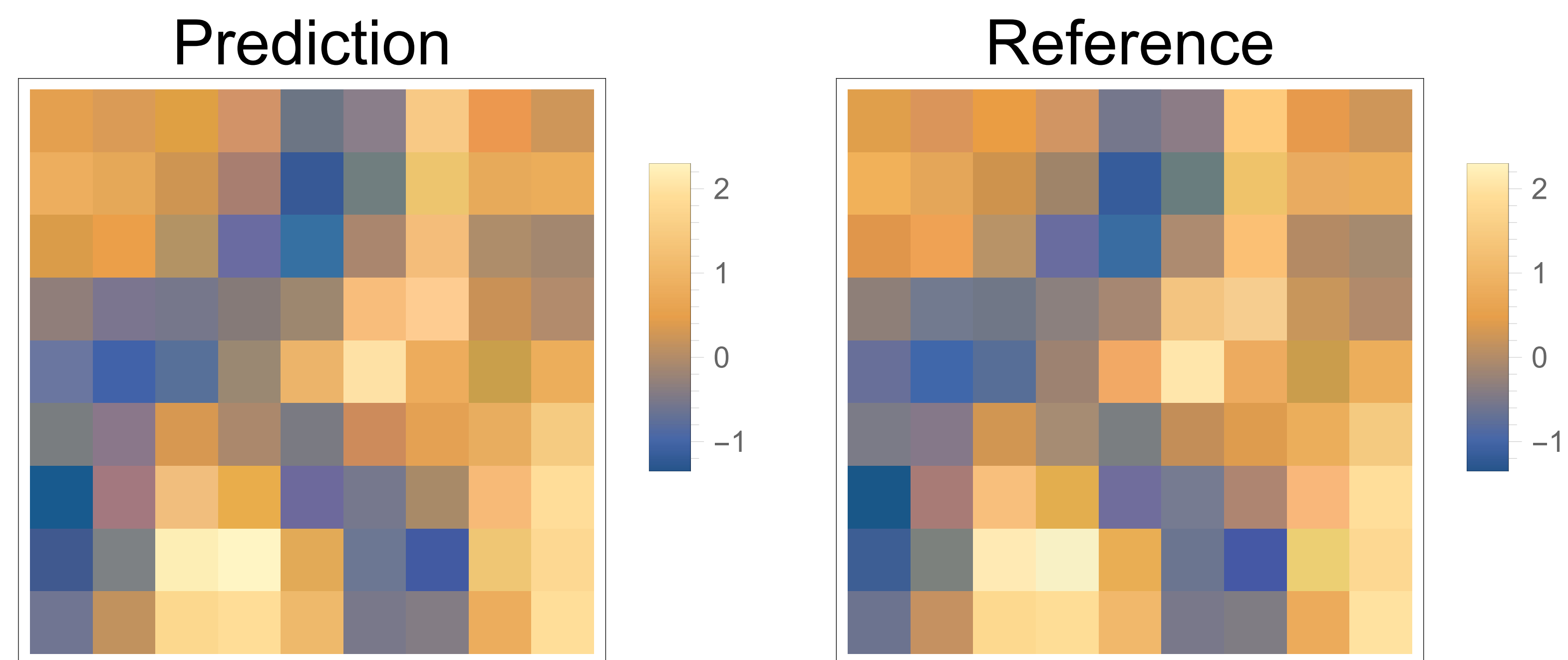
Statistical loss

- Predictions:



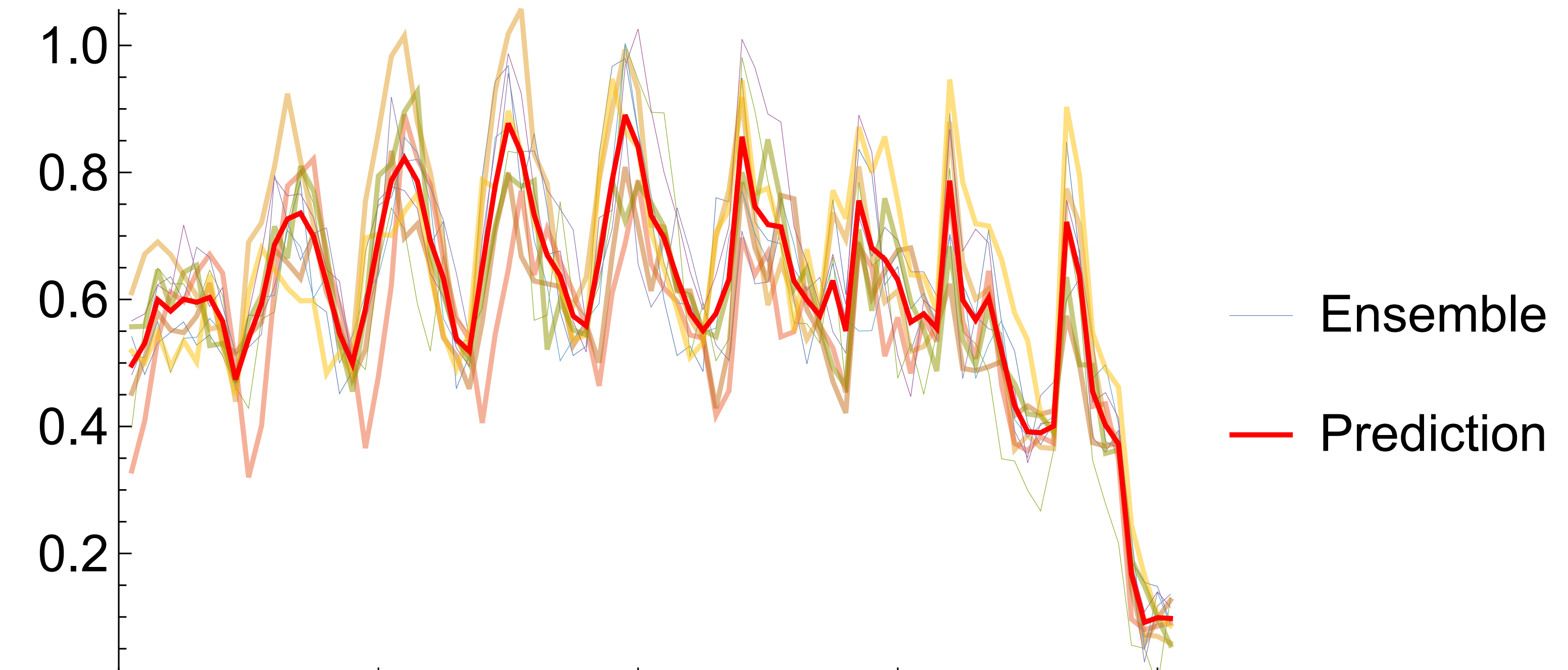
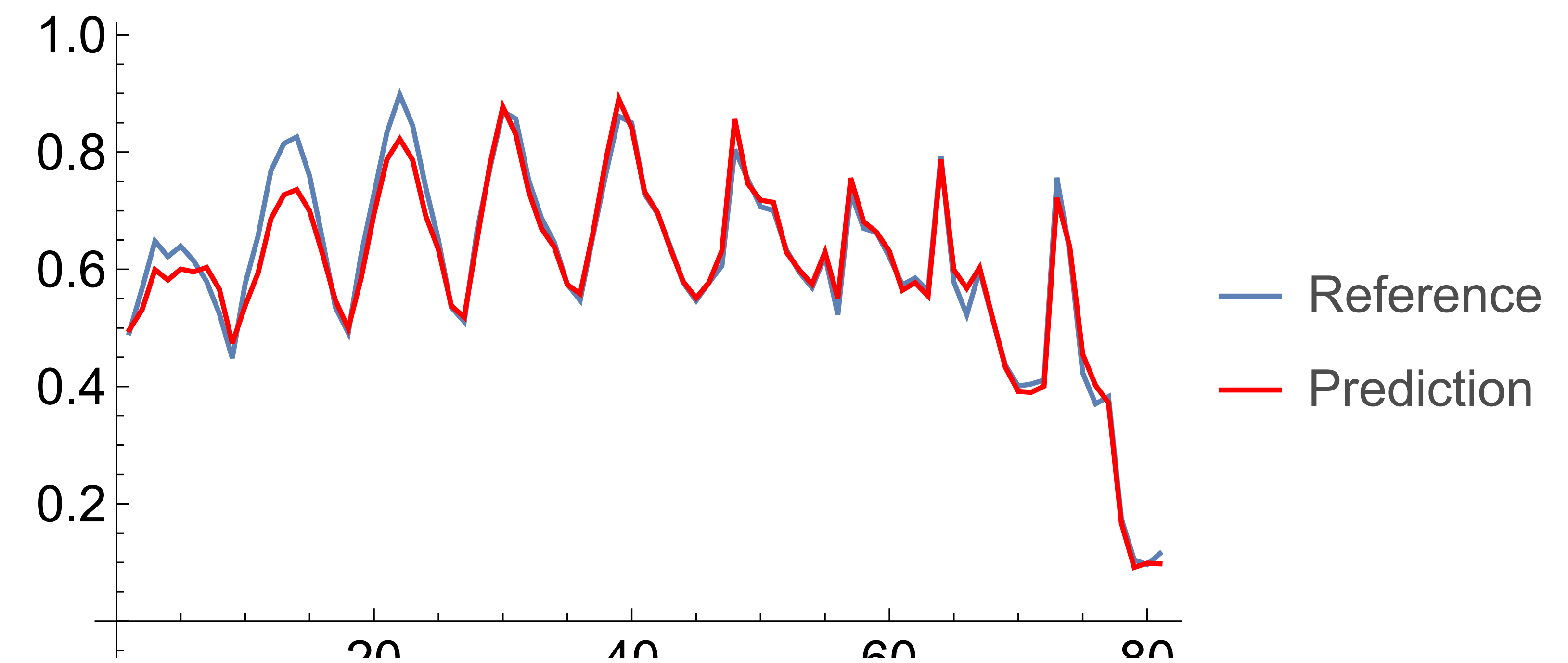
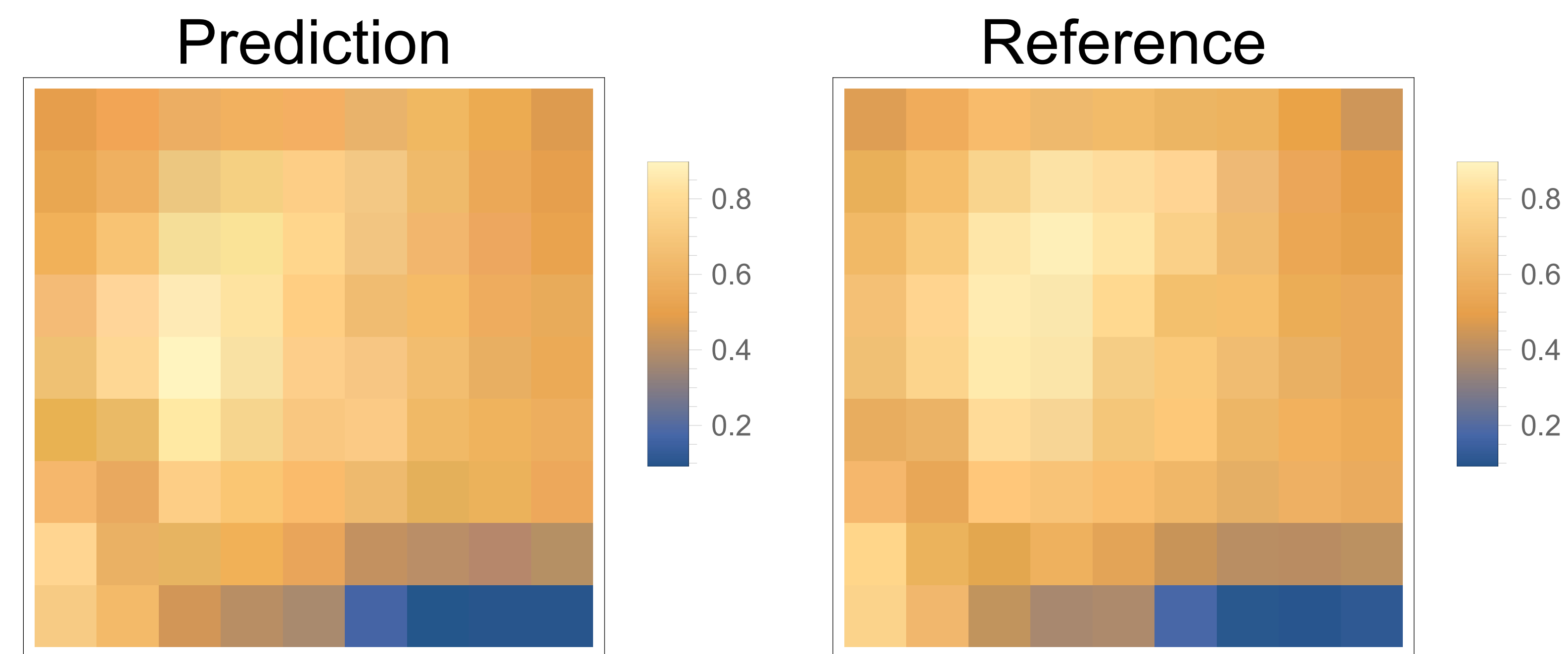
Statistical loss

- Predictions:

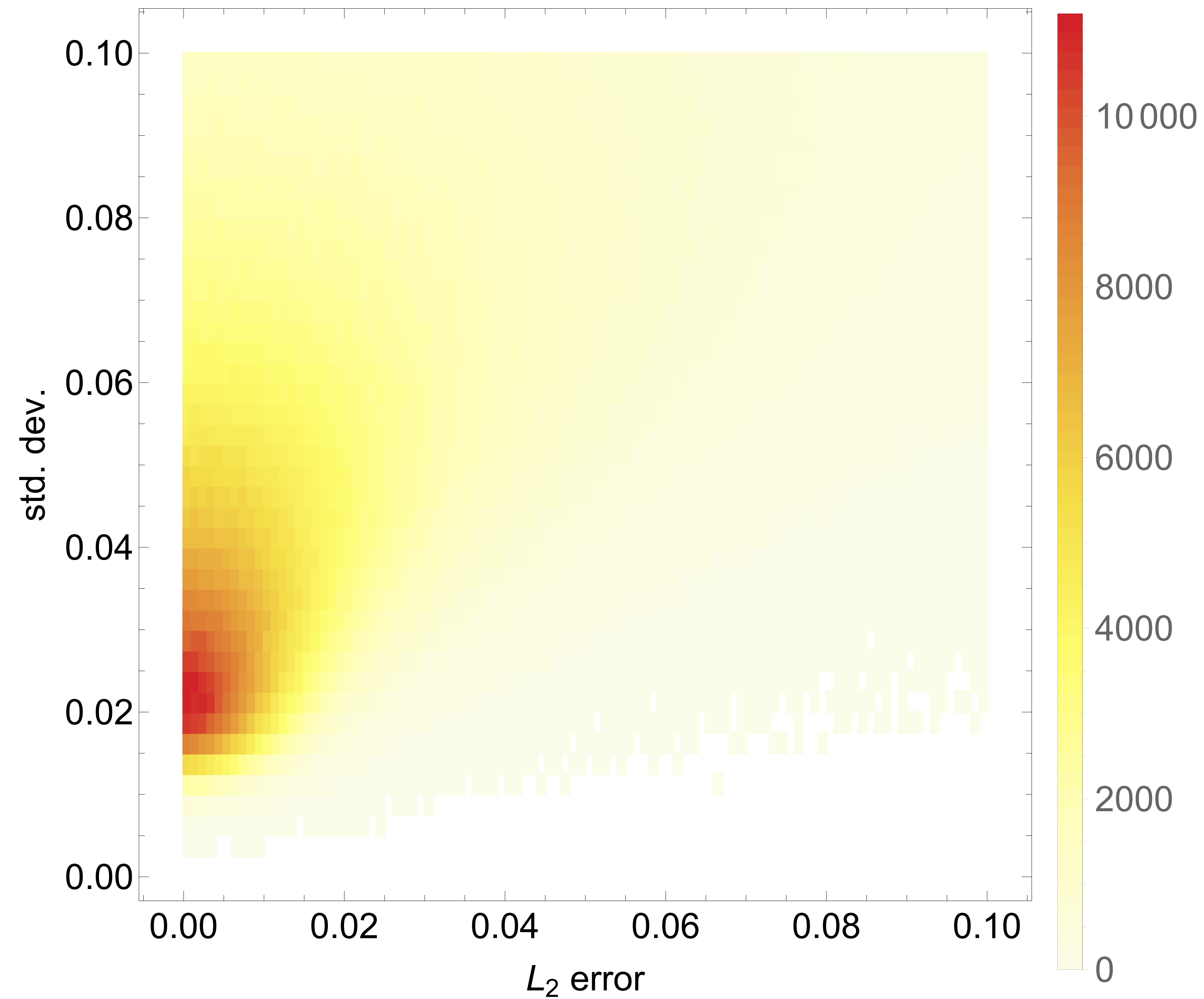


Statistical loss

- Predictions:



Statistical loss



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

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation
 - › Language models: chat programs, translation, auto-correction, ... from training on next sentence prediction task
 - › Natural language used to specify task

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation
 - › Language models: chat programs, translation, auto-correction, ... from training on next sentence prediction task
 - › Natural language used to specify task

What is in-context learning for AtmoRep $p_{\theta}(y|x, \alpha)$?

AtmoRep: in-context learning

- The model $p_{\theta}(y|x, \alpha)$ implies that what we want to “control” is the output state y without re-learning

AtmoRep: in-context learning

- The model $p_{\theta}(y|x, \alpha)$ implies that what we want to “control” is the output state y without re-learning

$$p_{\theta}(y|x, \alpha, \tau)$$

↑
positional encoding
for output

AtmoRep: in-context learning

- The model $p_{\theta}(y|x, \alpha)$ implies that what we want to “control” is the output state y without re-learning

$$p_{\theta}(y|x, \alpha, \tau)$$

- › τ : spatial and temporal location, resolution, quality
- › Few shot: “explain” τ to the network

AtmoRep: in-context learning

- The model $p_{\theta}(y|x, \alpha)$ implies that what we want to “control” is the output state y without re-learning

$$p_{\theta}(y|x, \alpha, \tau)$$

- › τ : spatial and temporal location, resolution, quality
- › Few shot: “explain” τ to the network
- Does this make the network a scientific model?

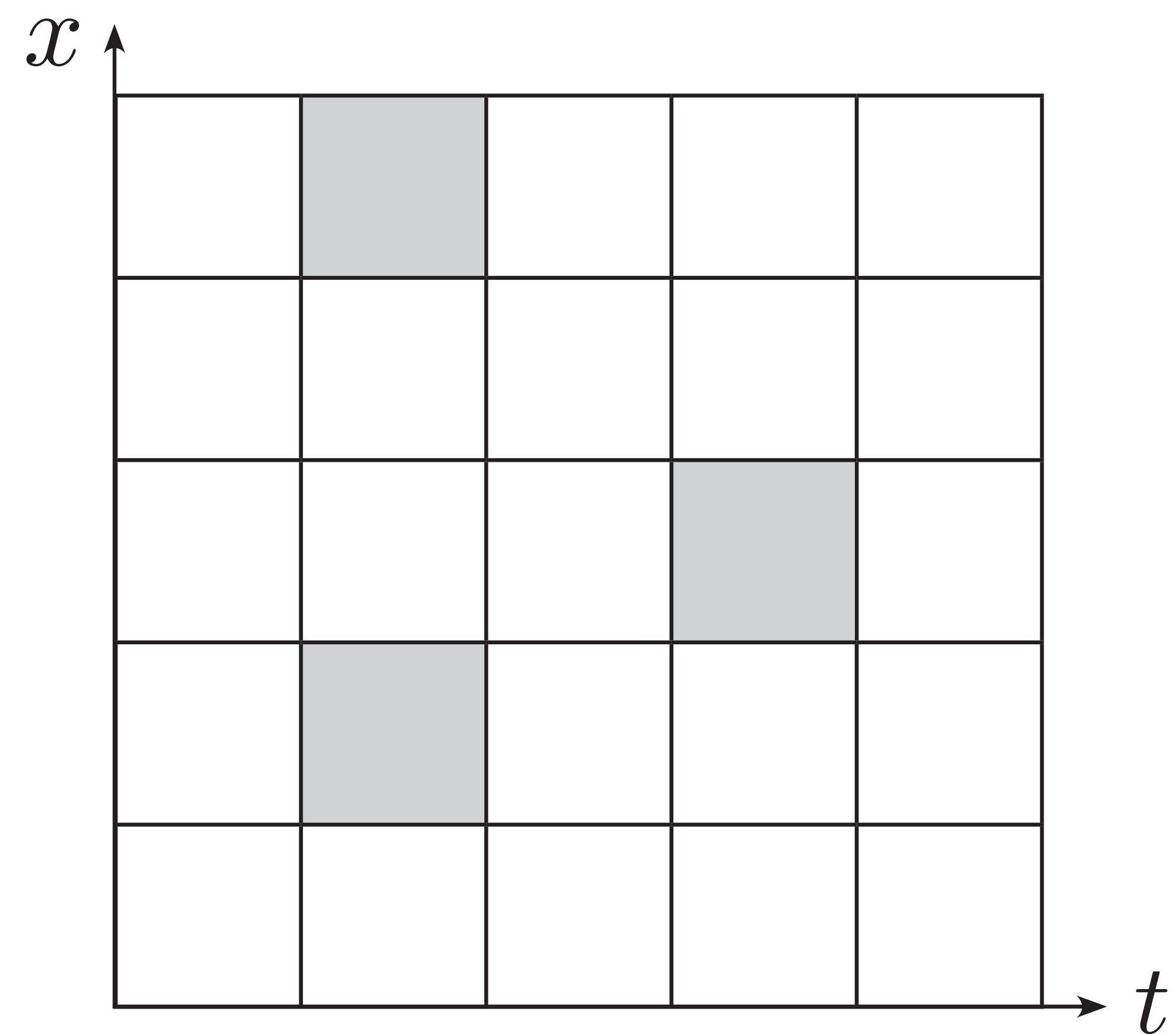
“I confess even to this day that I still don’t understand quantum mechanics, and I’m not even sure I really know how to use it all that well. And a lot of this has to do with the fact that I still don’t understand it.”

John Clauser, 2002

Quoted from <https://www.nytimes.com/2022/10/04/science/nobel-prize-physics-winner.html>

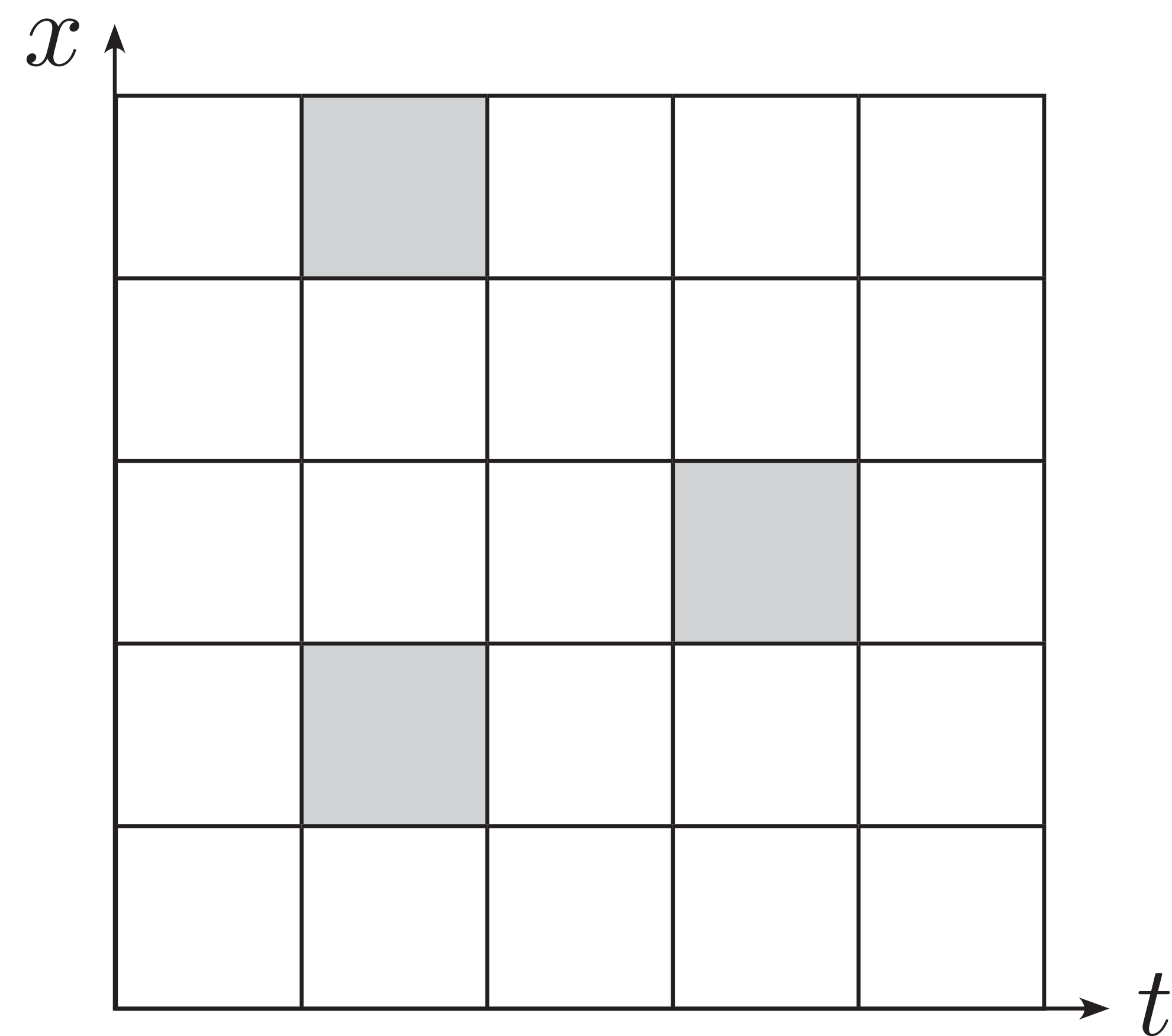
Zero shot forecasting

BERT

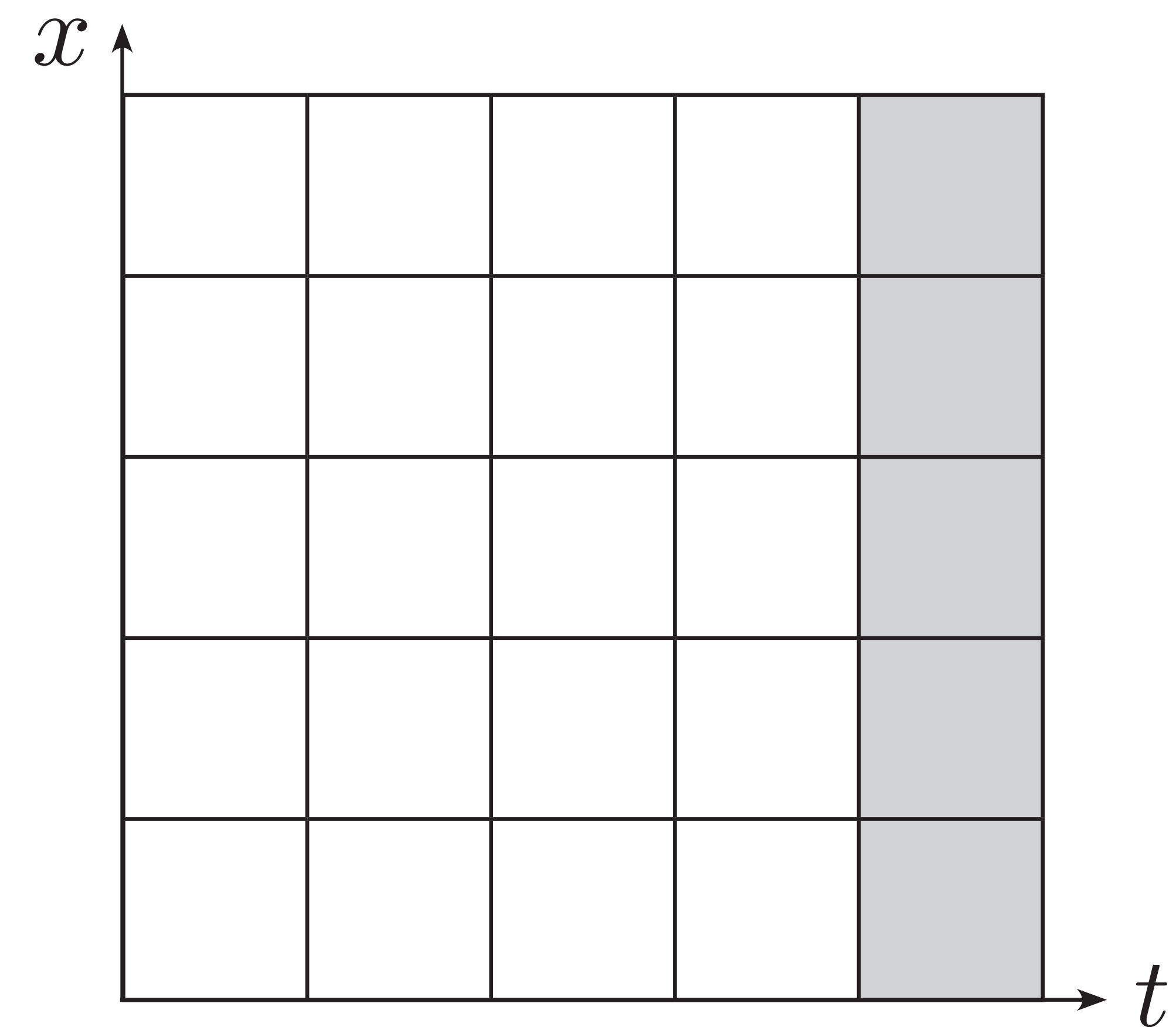


Zero shot forecasting

BERT

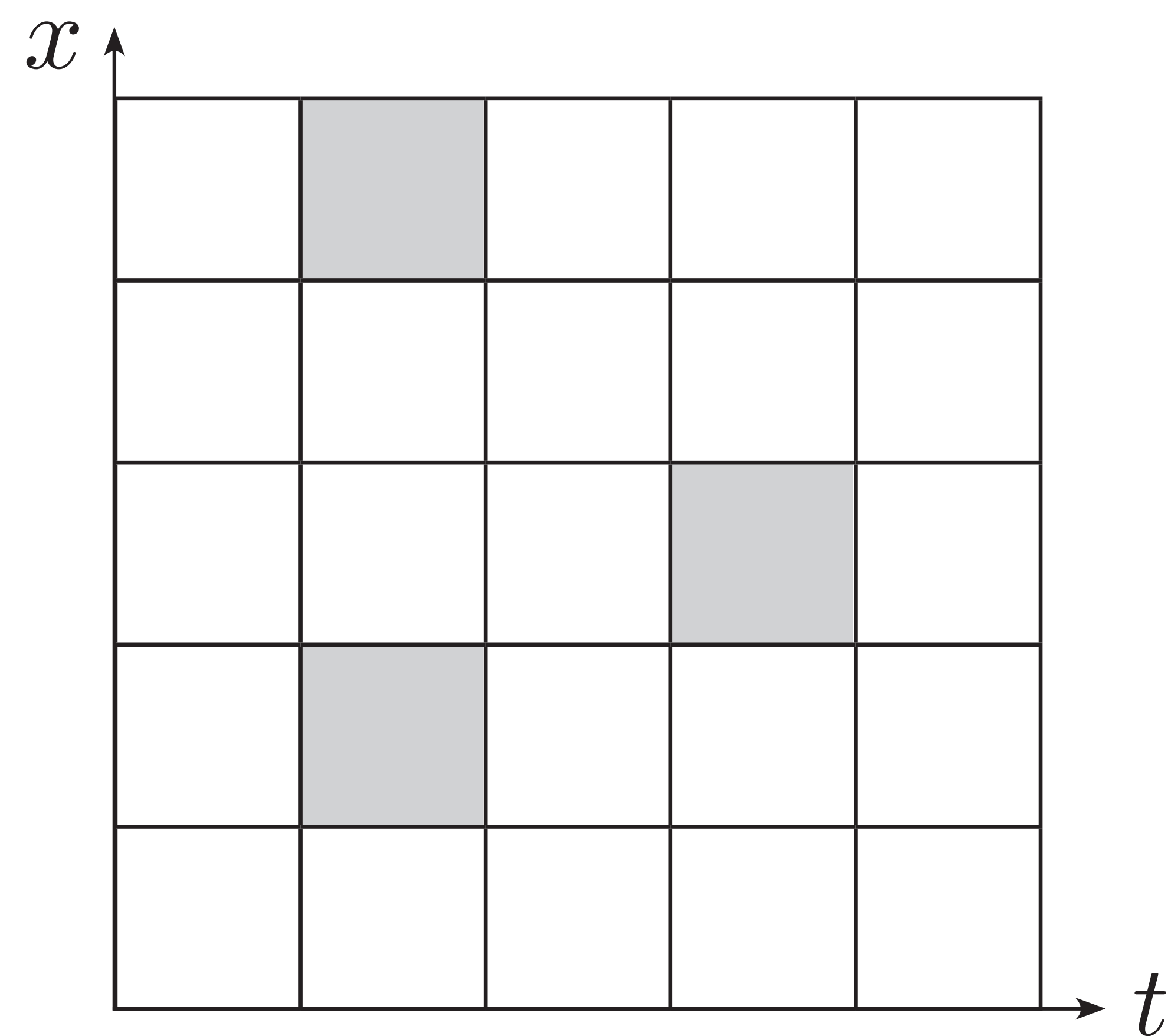


BERT-Forecast

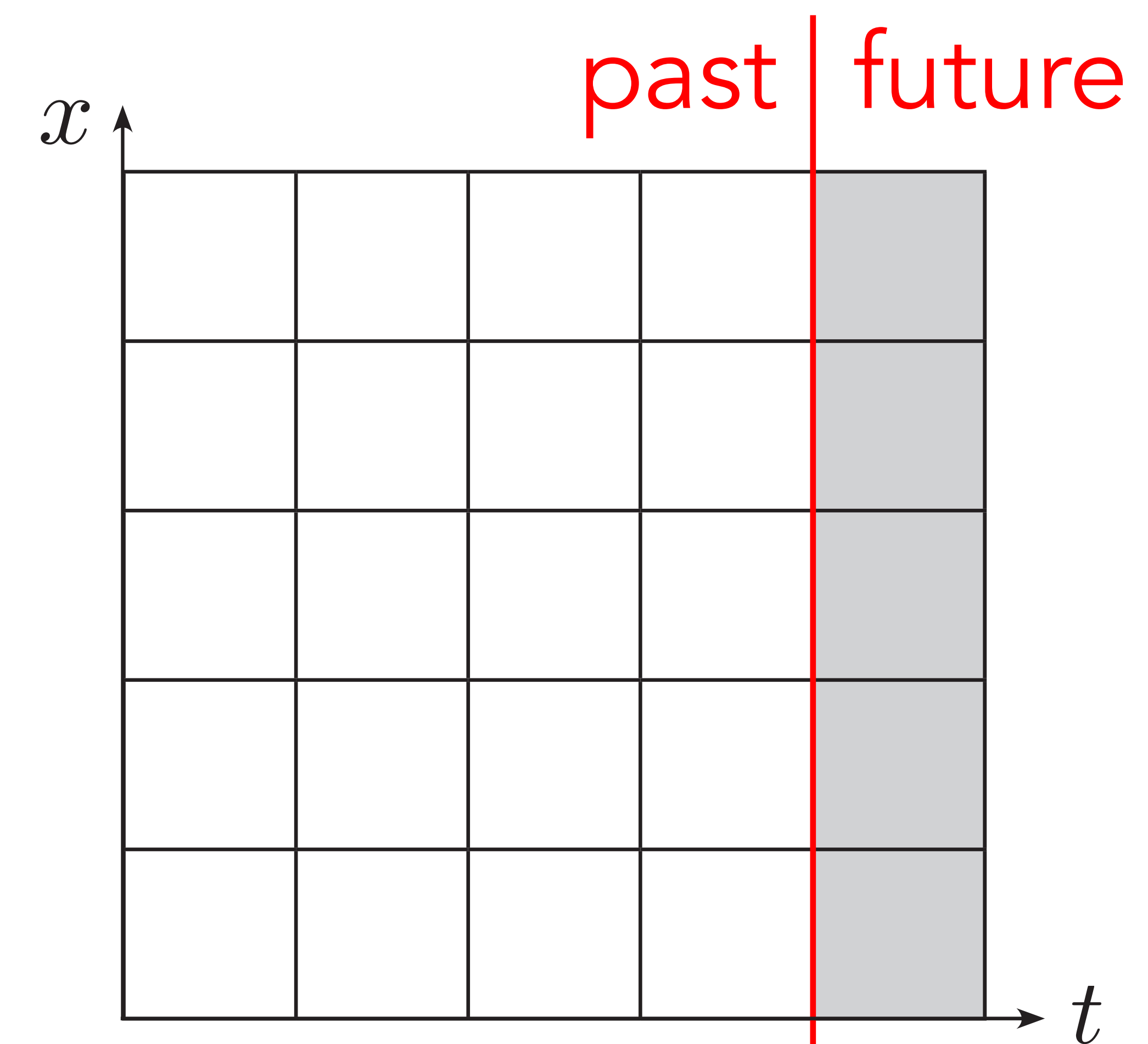


Zero shot forecasting

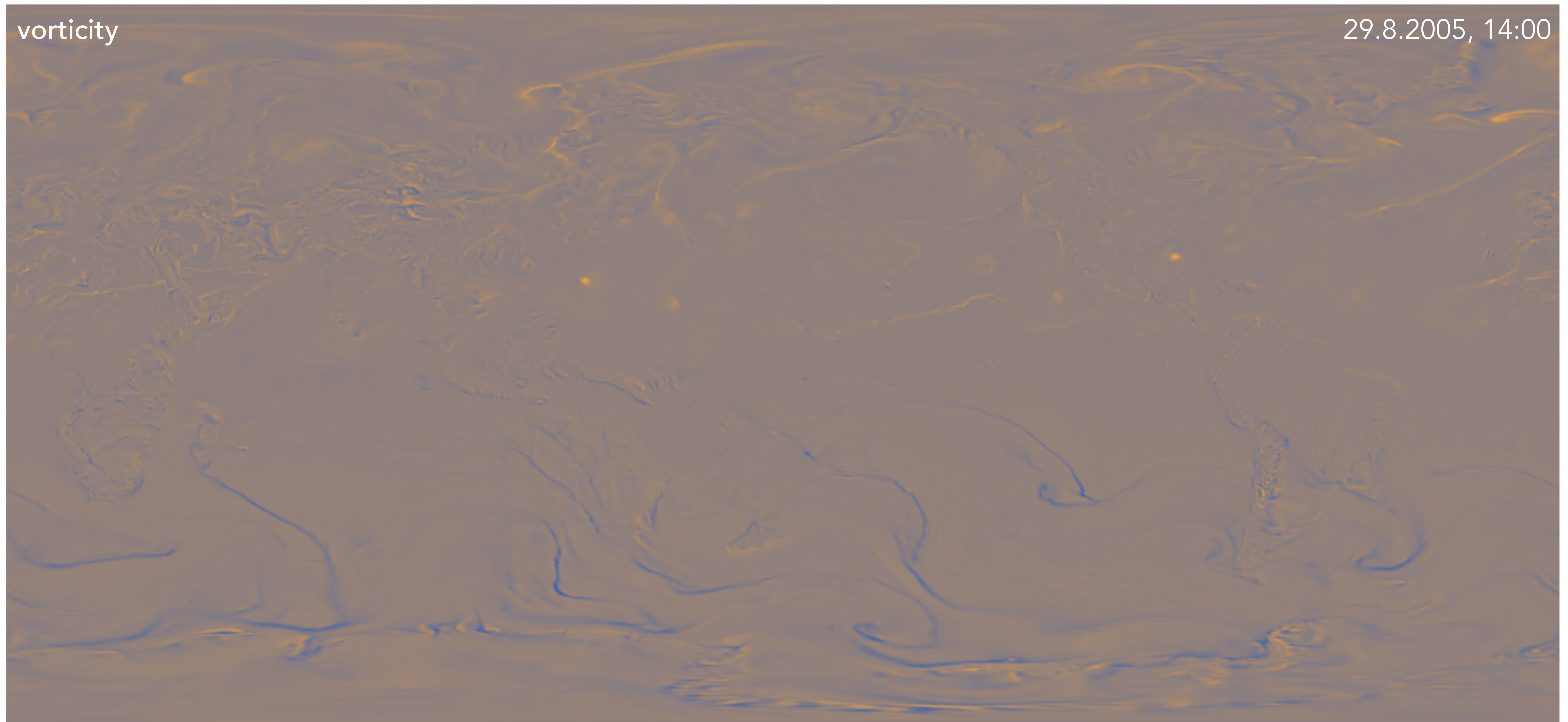
BERT



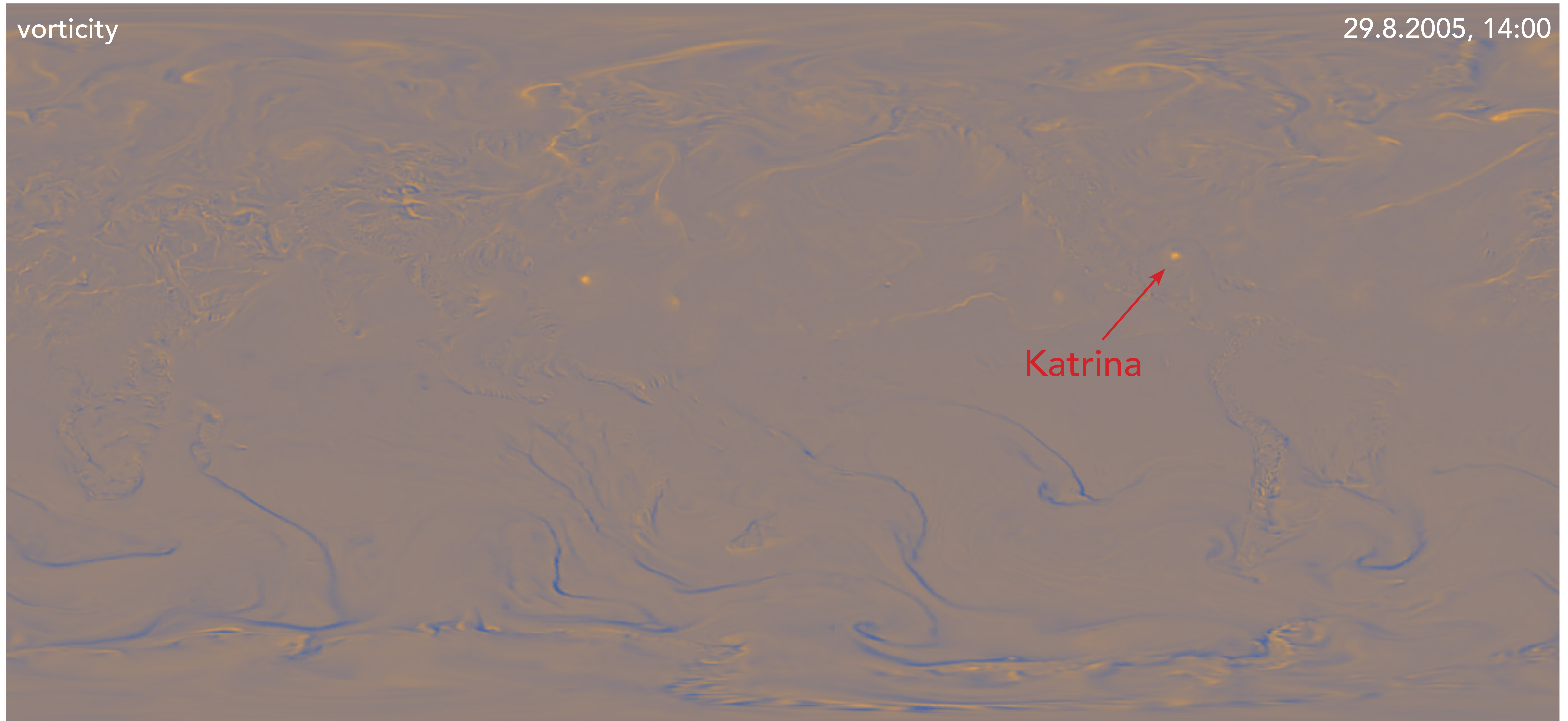
BERT-Forecast



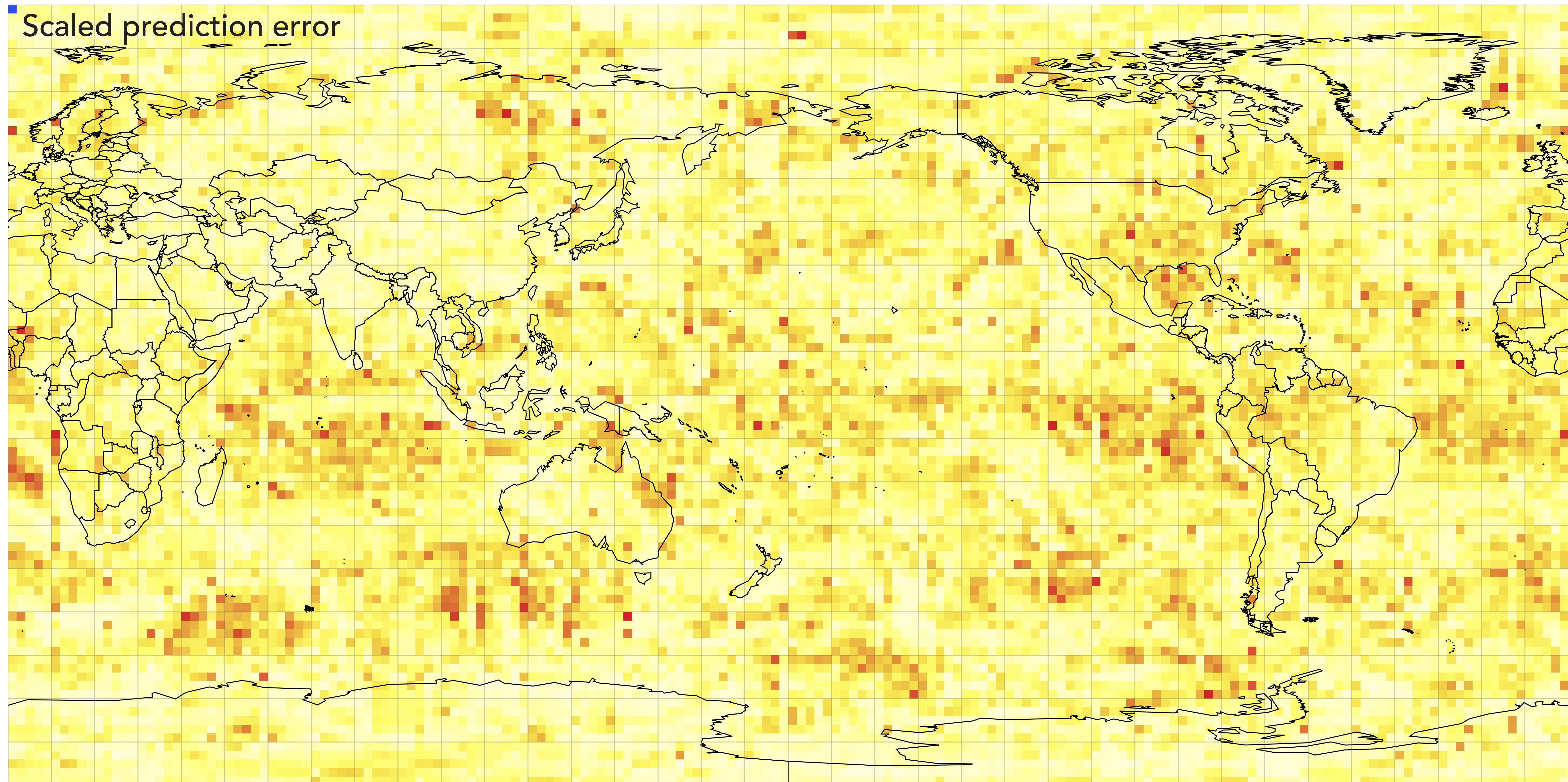
Zero shot forecasting



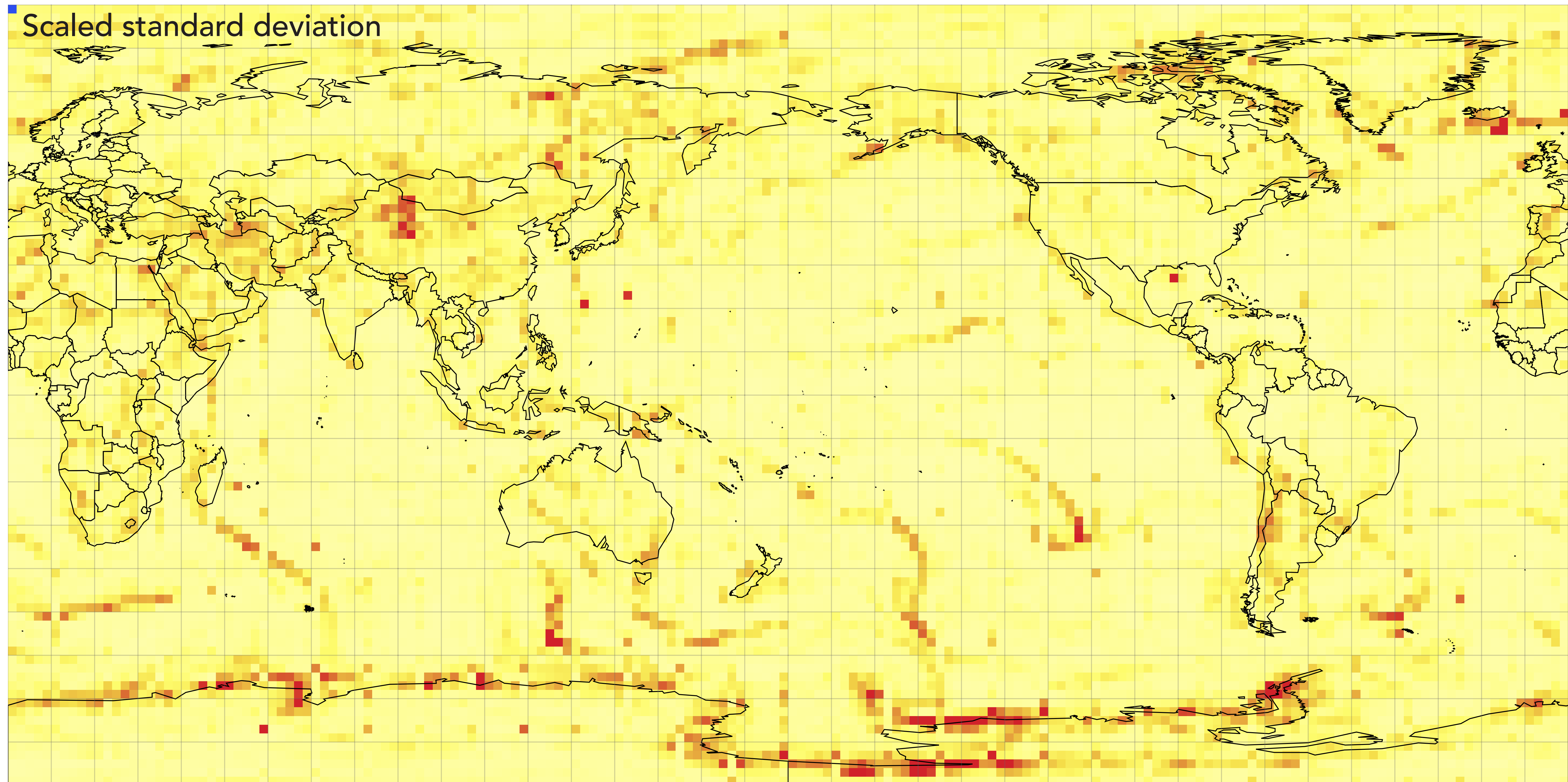
Zero shot forecasting



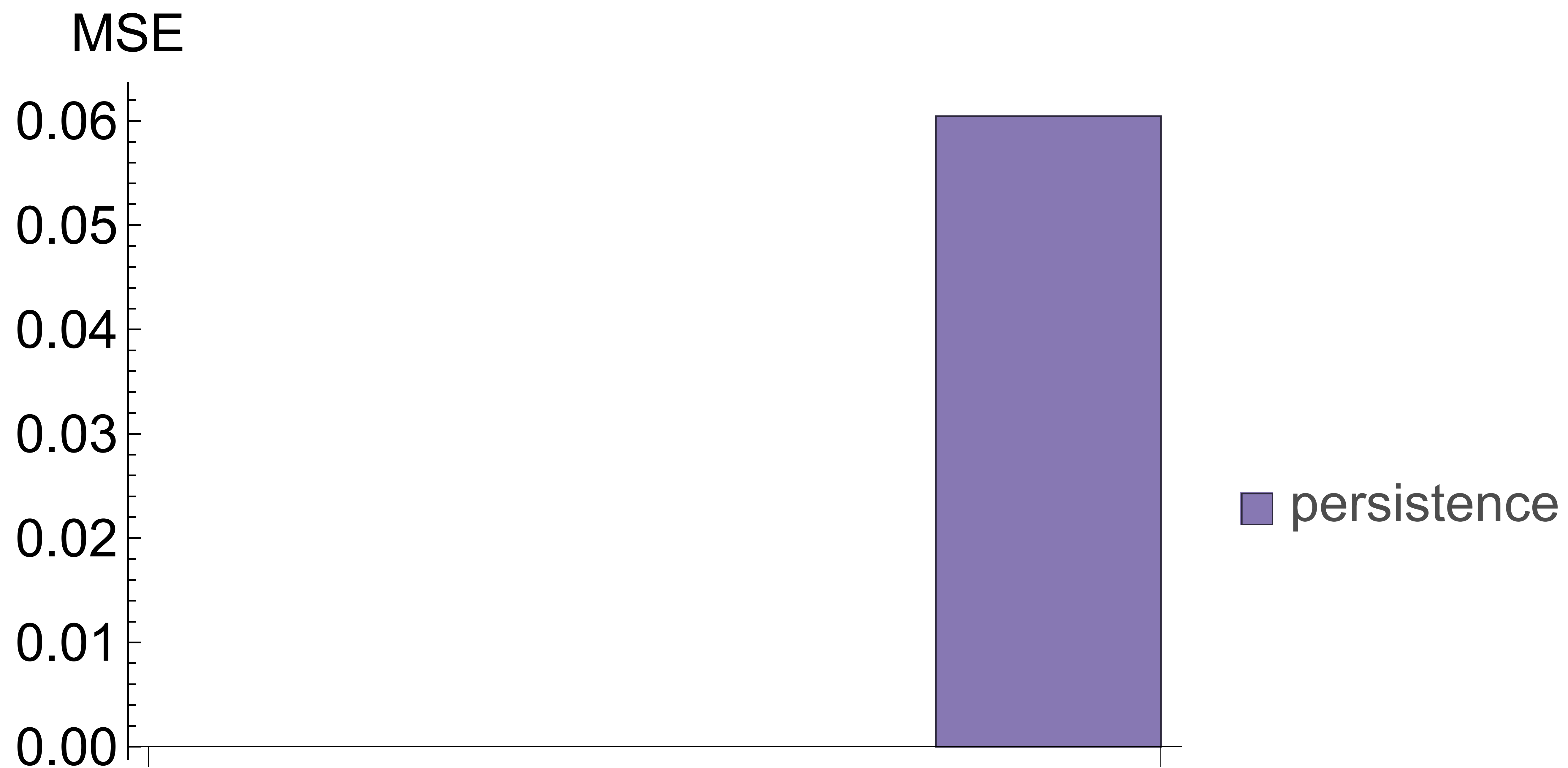
Zero shot forecasting



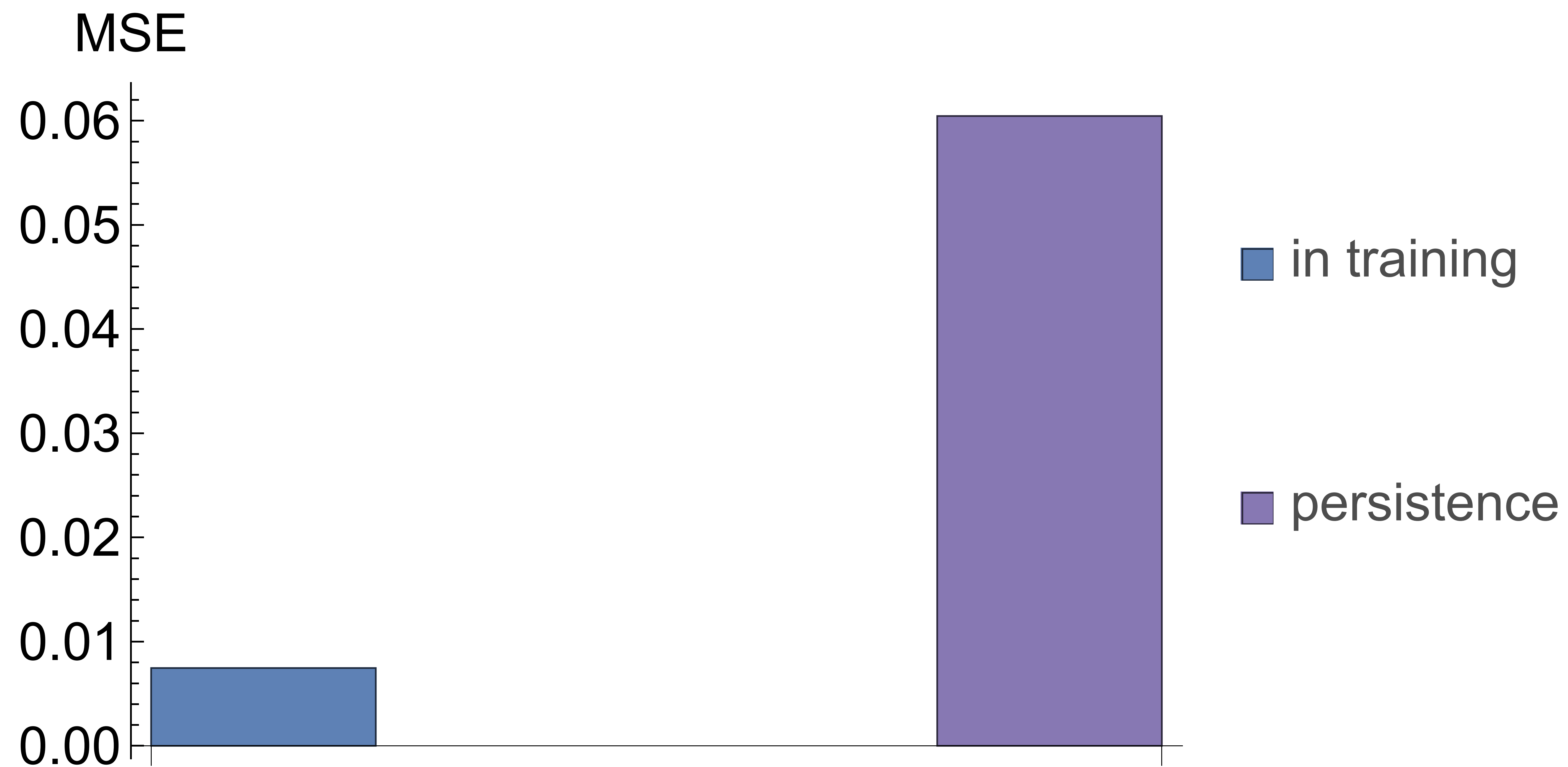
Zero shot forecasting



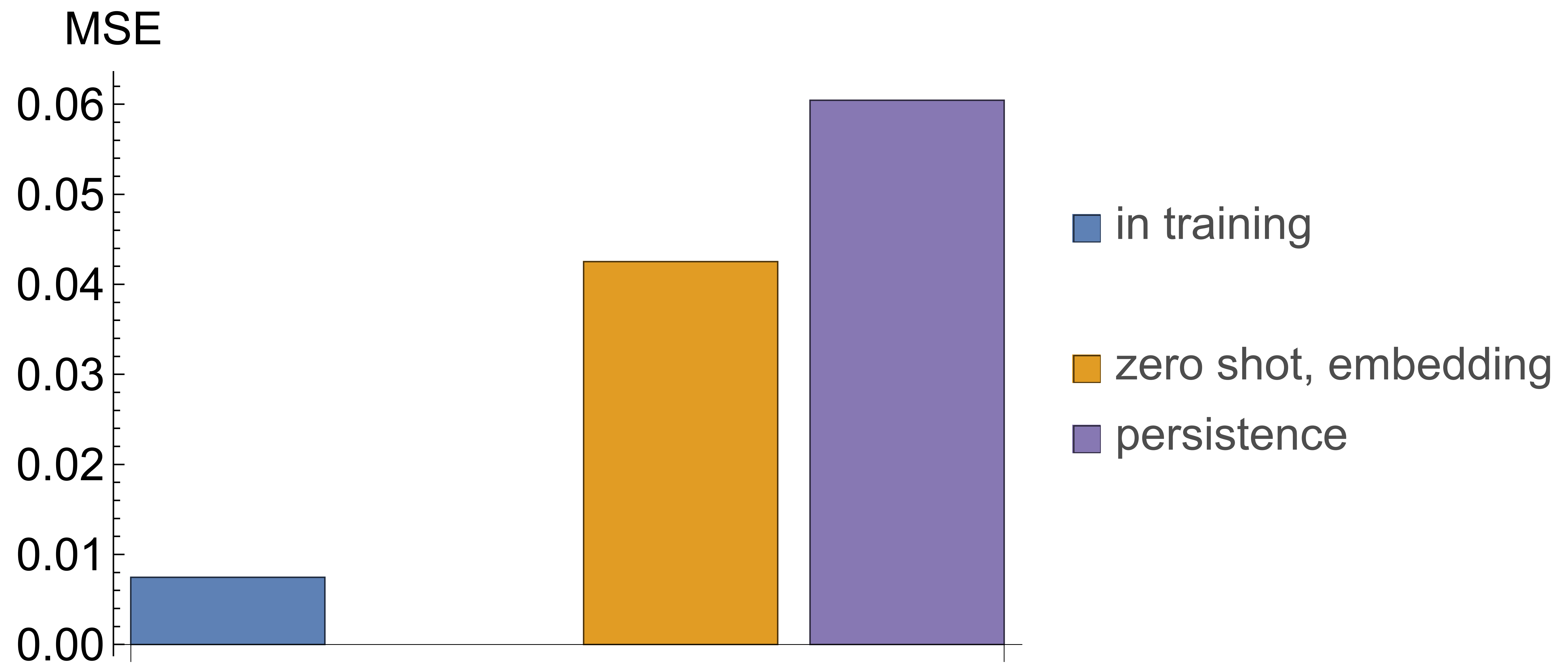
Zero shot forecasting



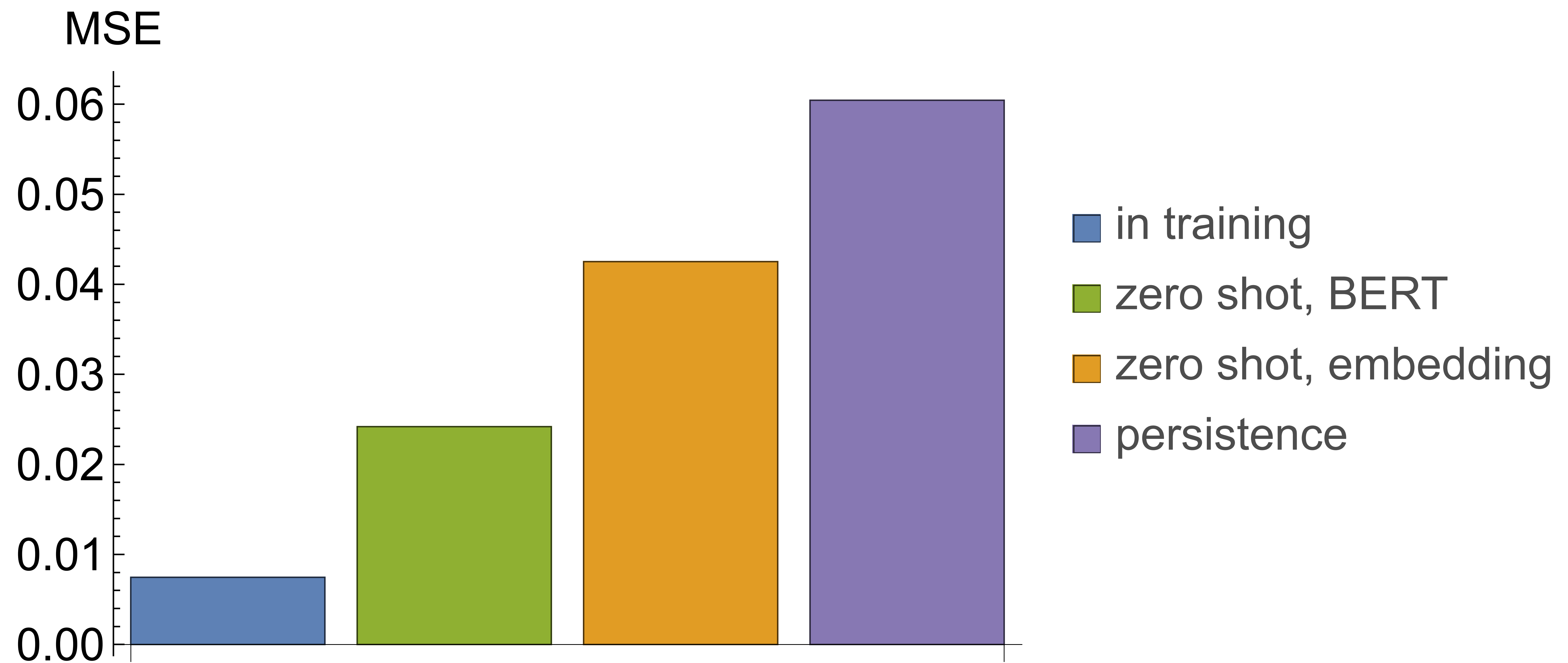
Zero shot forecasting



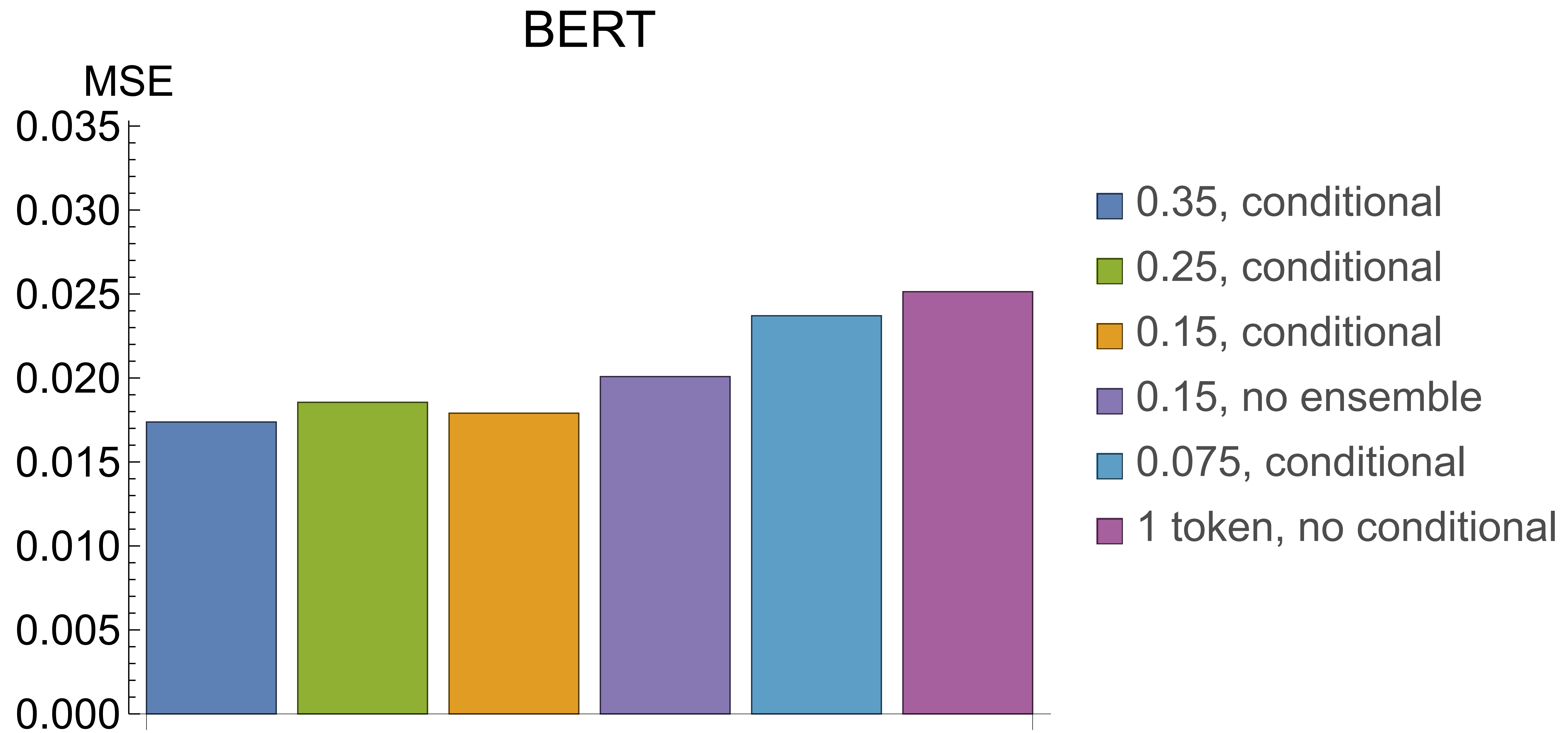
Zero shot forecasting



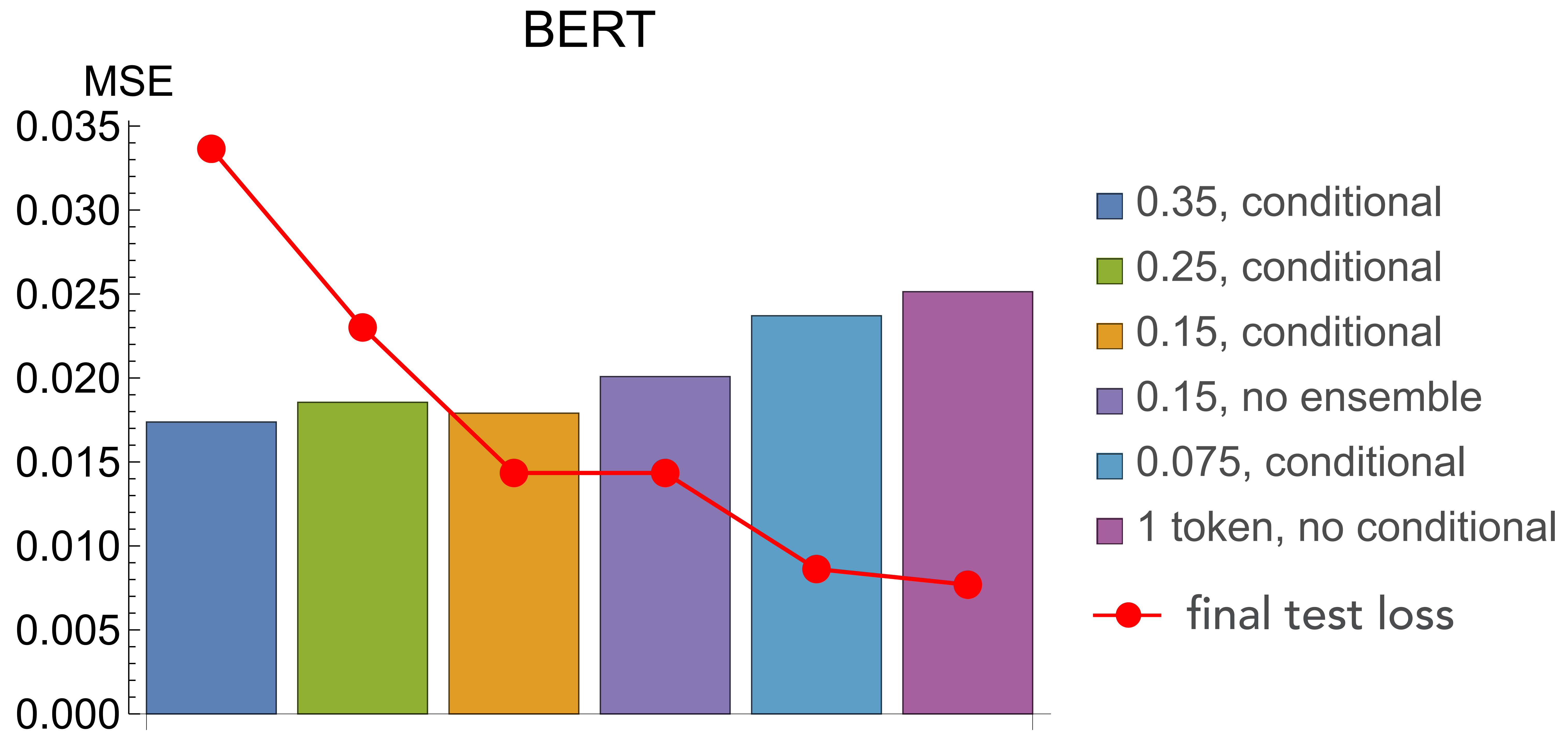
Zero shot forecasting



Zero shot forecasting



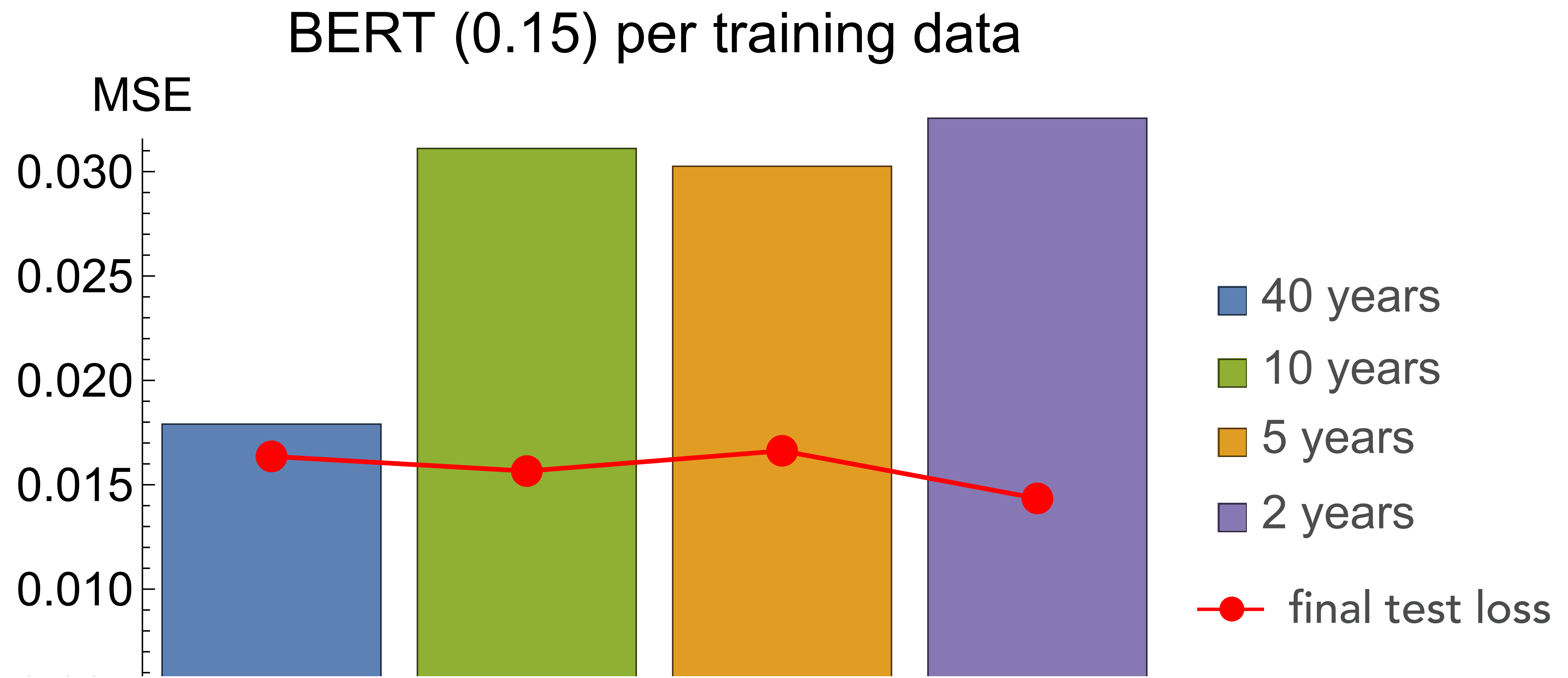
Zero shot forecasting



Zero shot forecasting

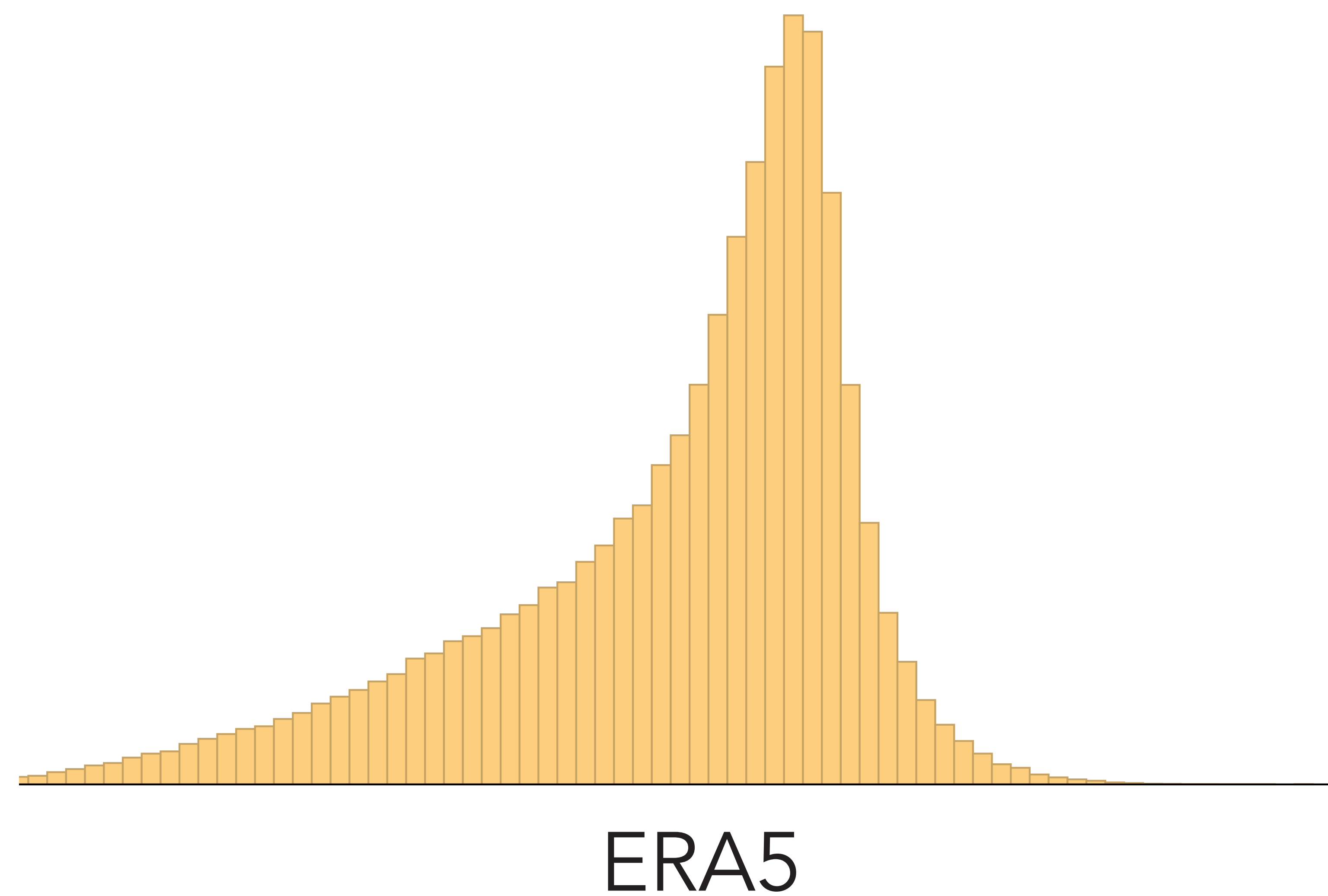


Zero shot forecasting



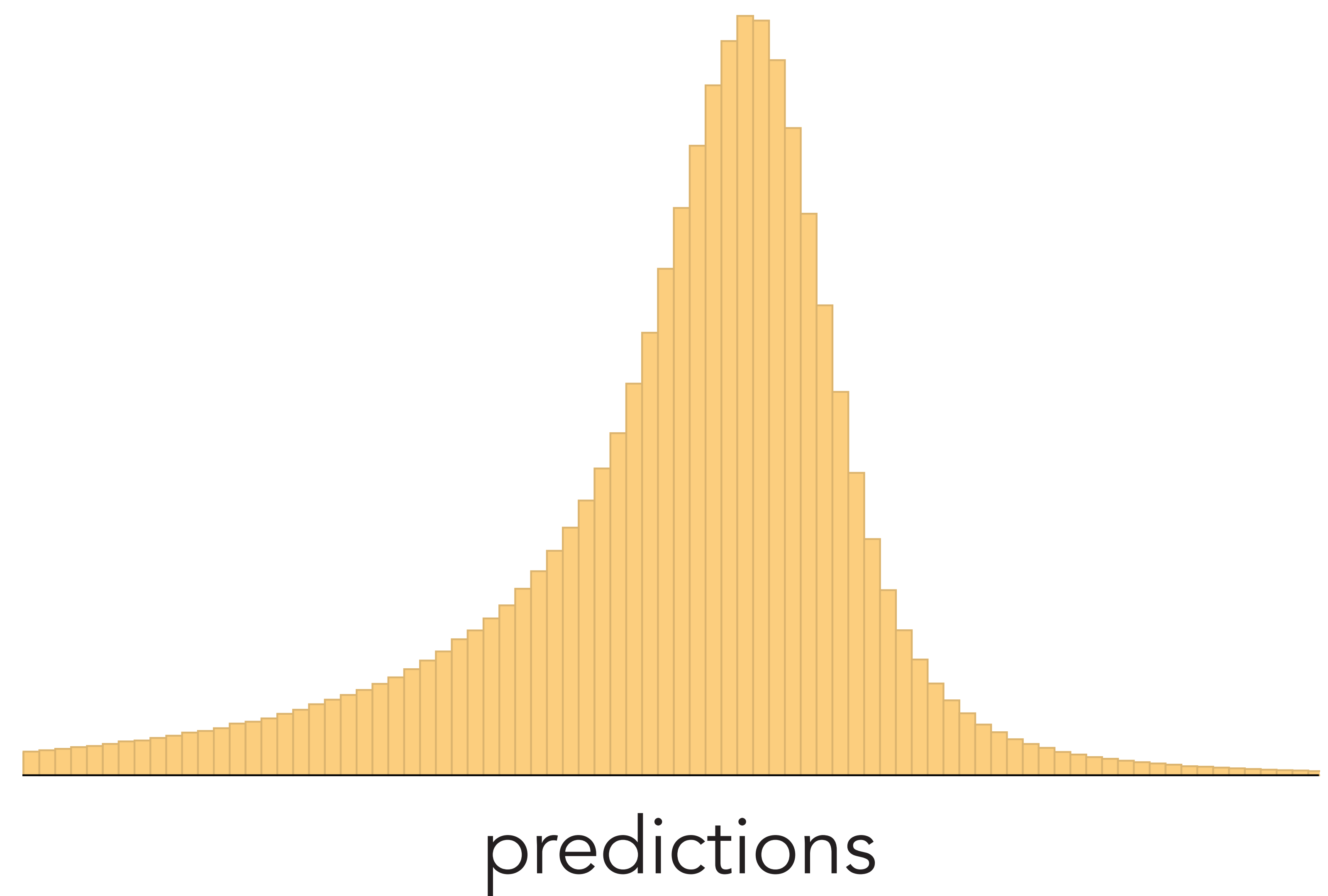
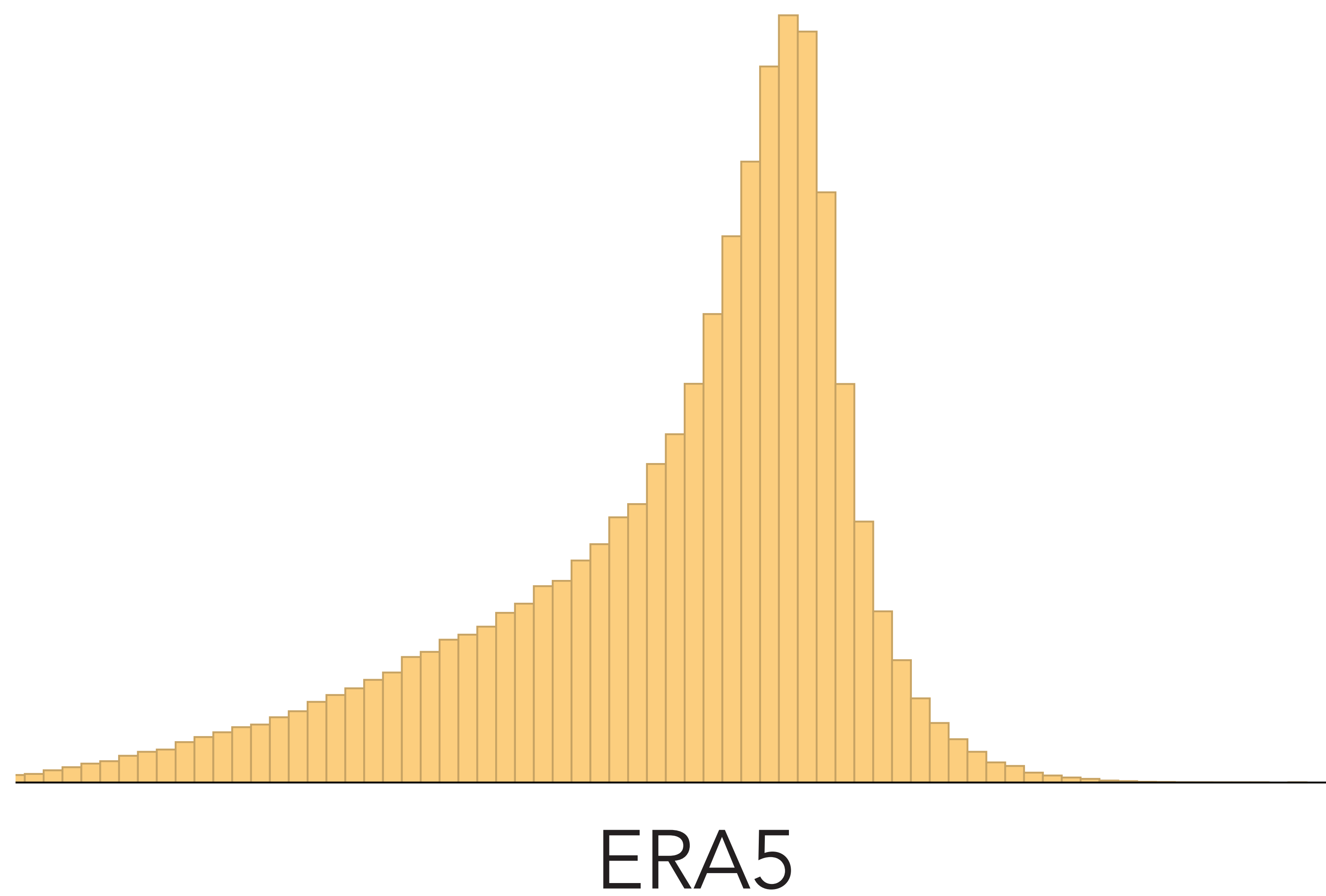
Zero shot forecasting

Cape town: histogram of vorticity



Zero shot forecasting

Cape town: histogram of vorticity

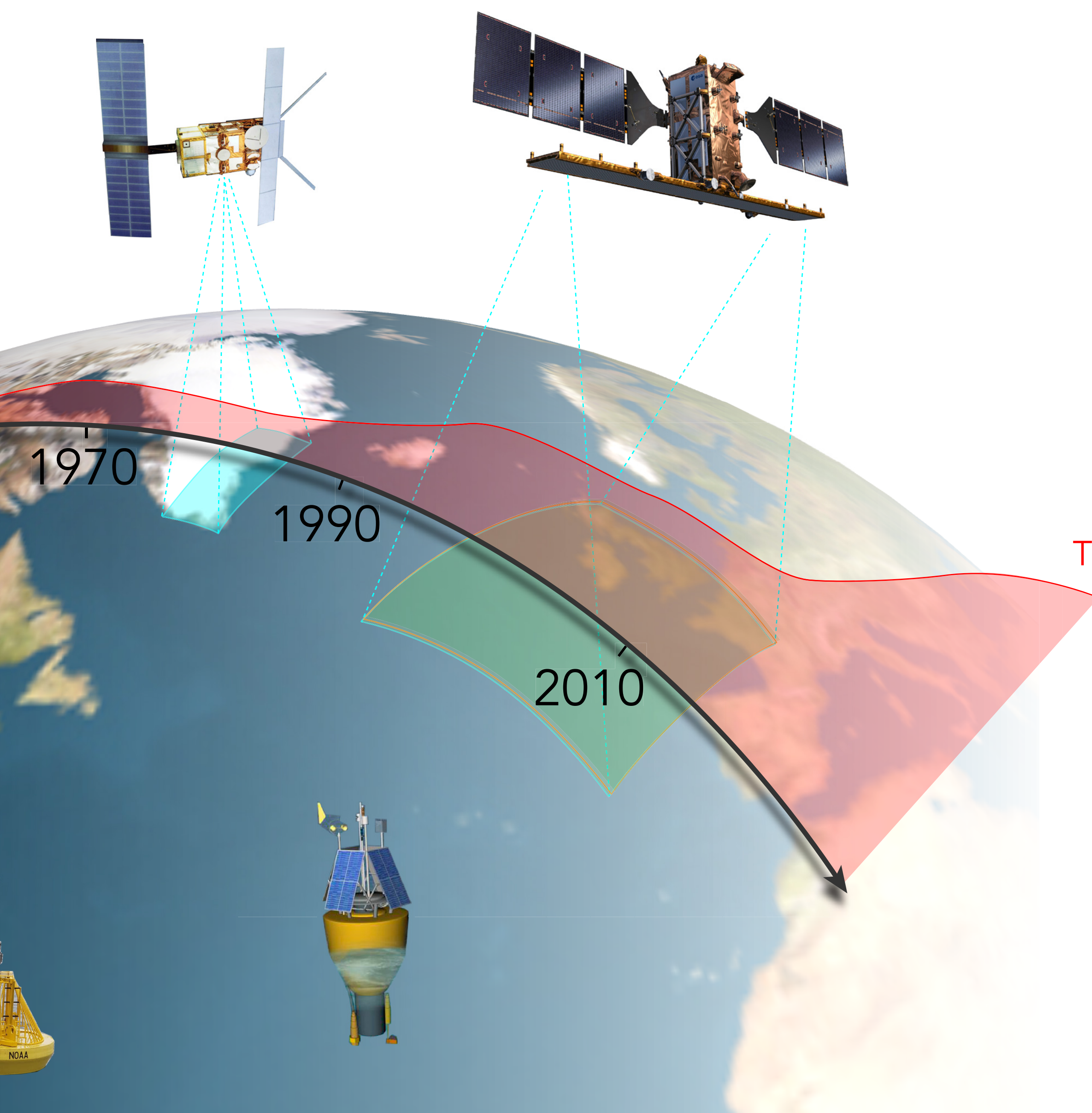


AtmoRep: longer term objectives

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

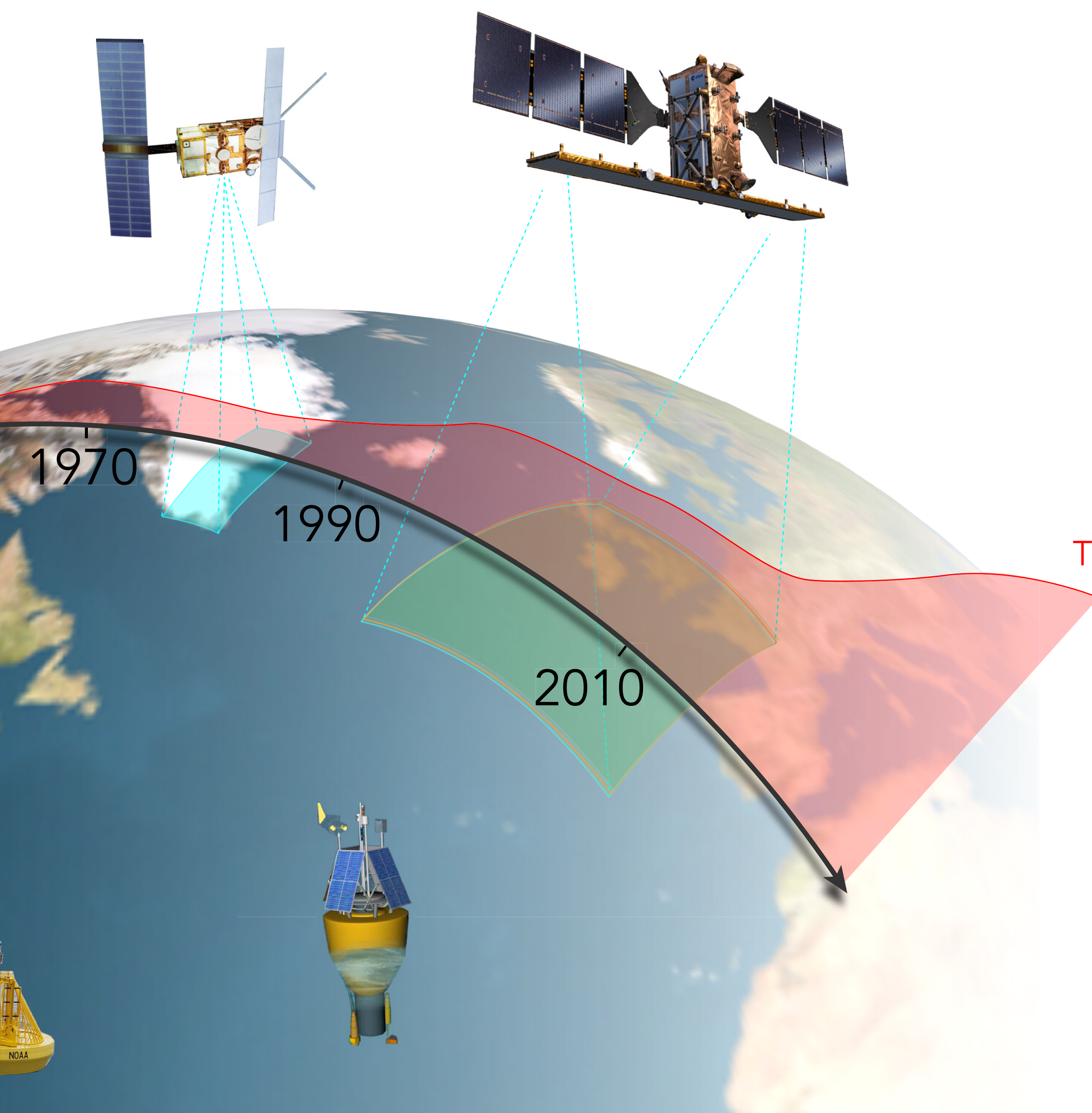
Current / next steps

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



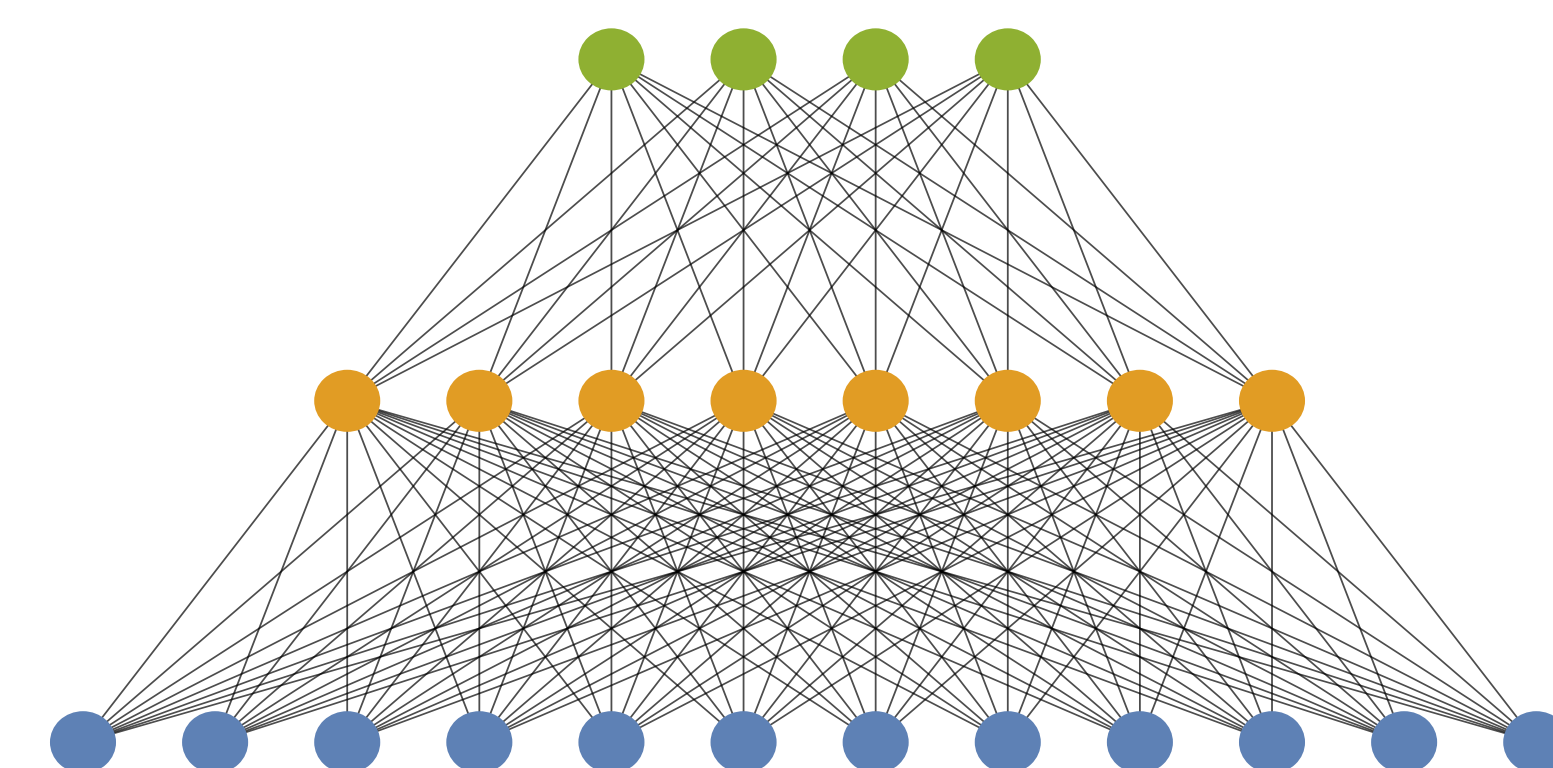
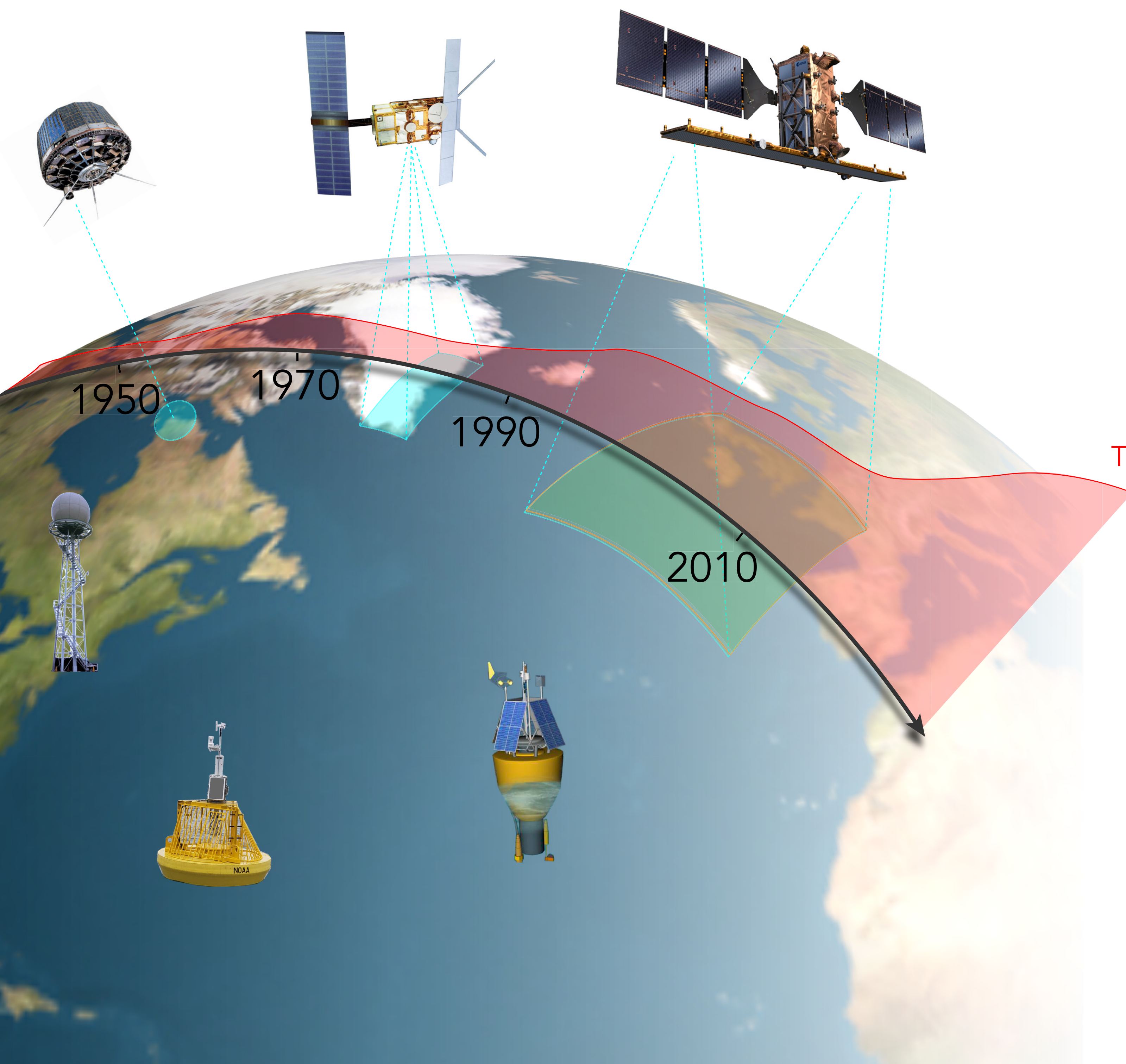
Current / next steps

- Complete representation learning model
- Downstream applications
 - › Weather forecasting
 - › Downscaling
 - › Model correction
 - › ...



AtmoRep

Large scale representation learning of atmospheric dynamics

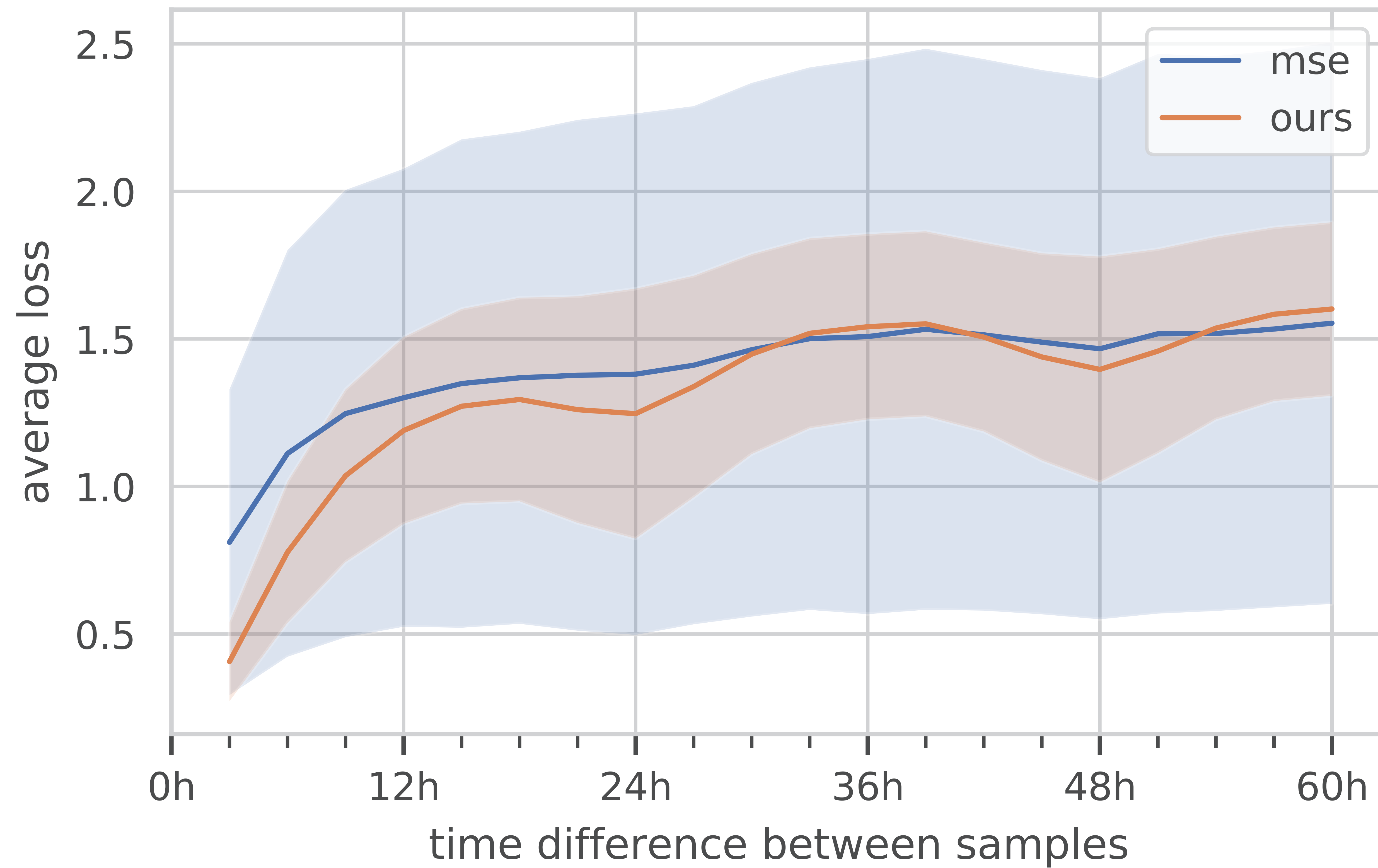


large scale machine learning

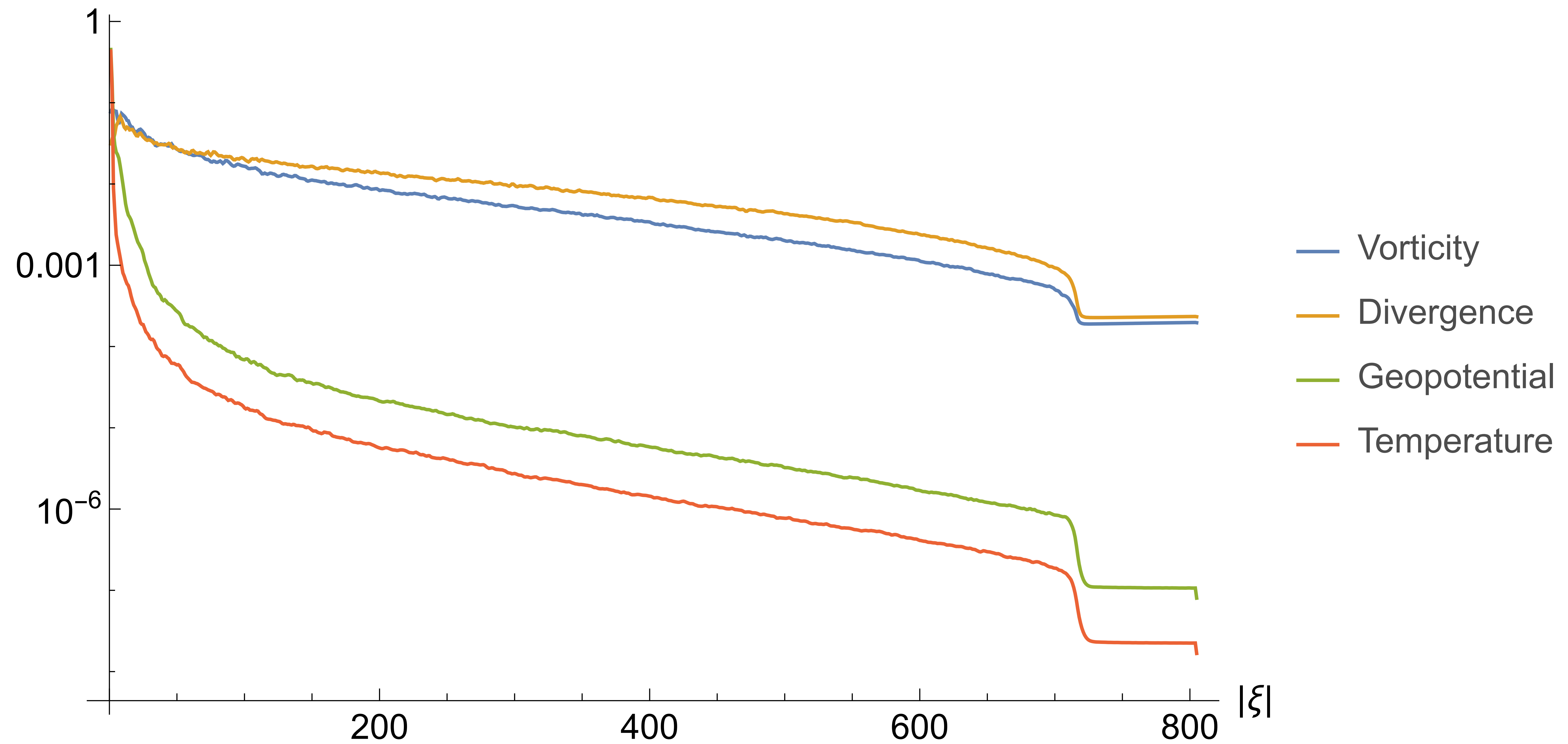
address climate change

scientific insight

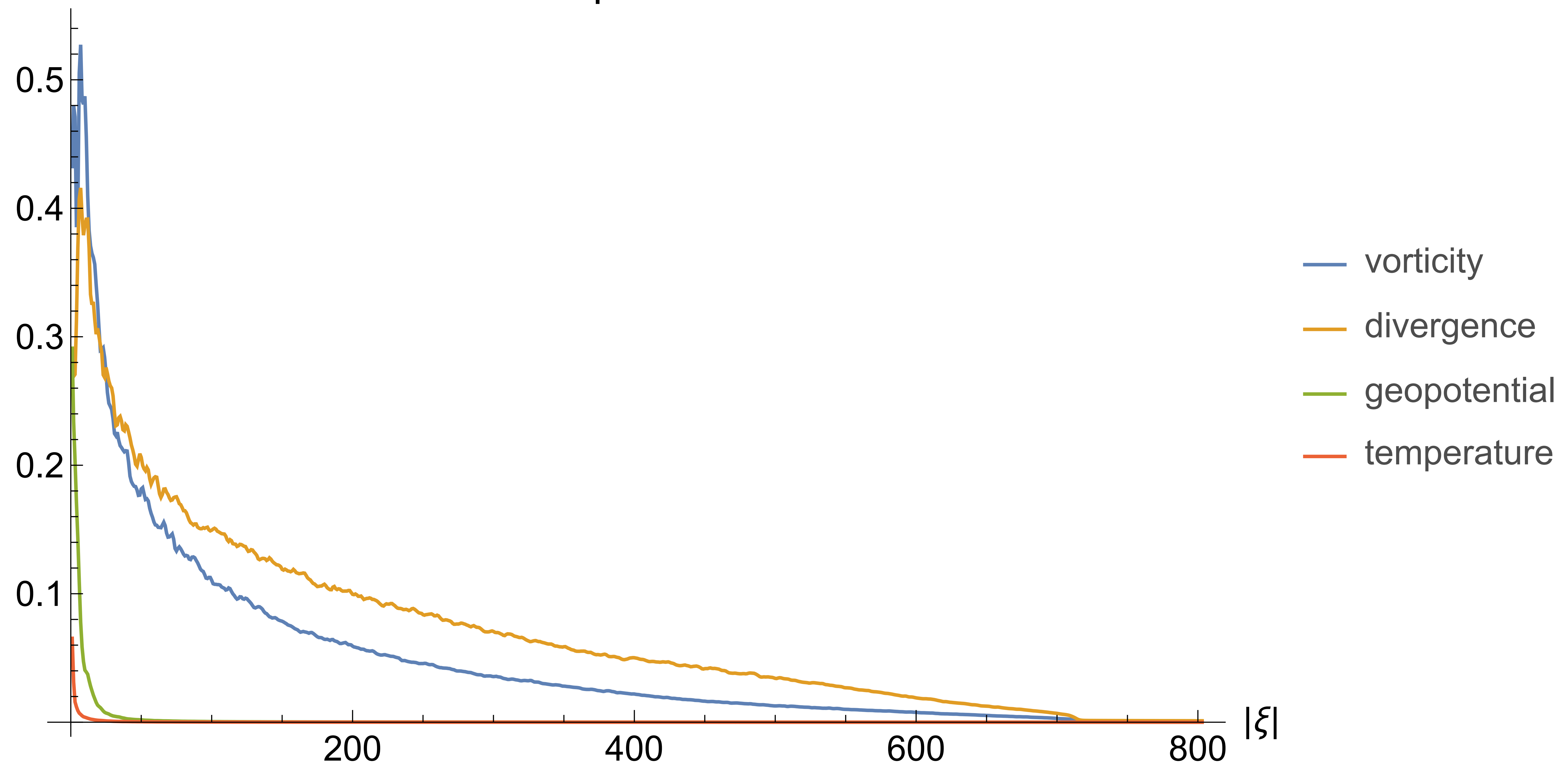
AtmoDist: evaluation



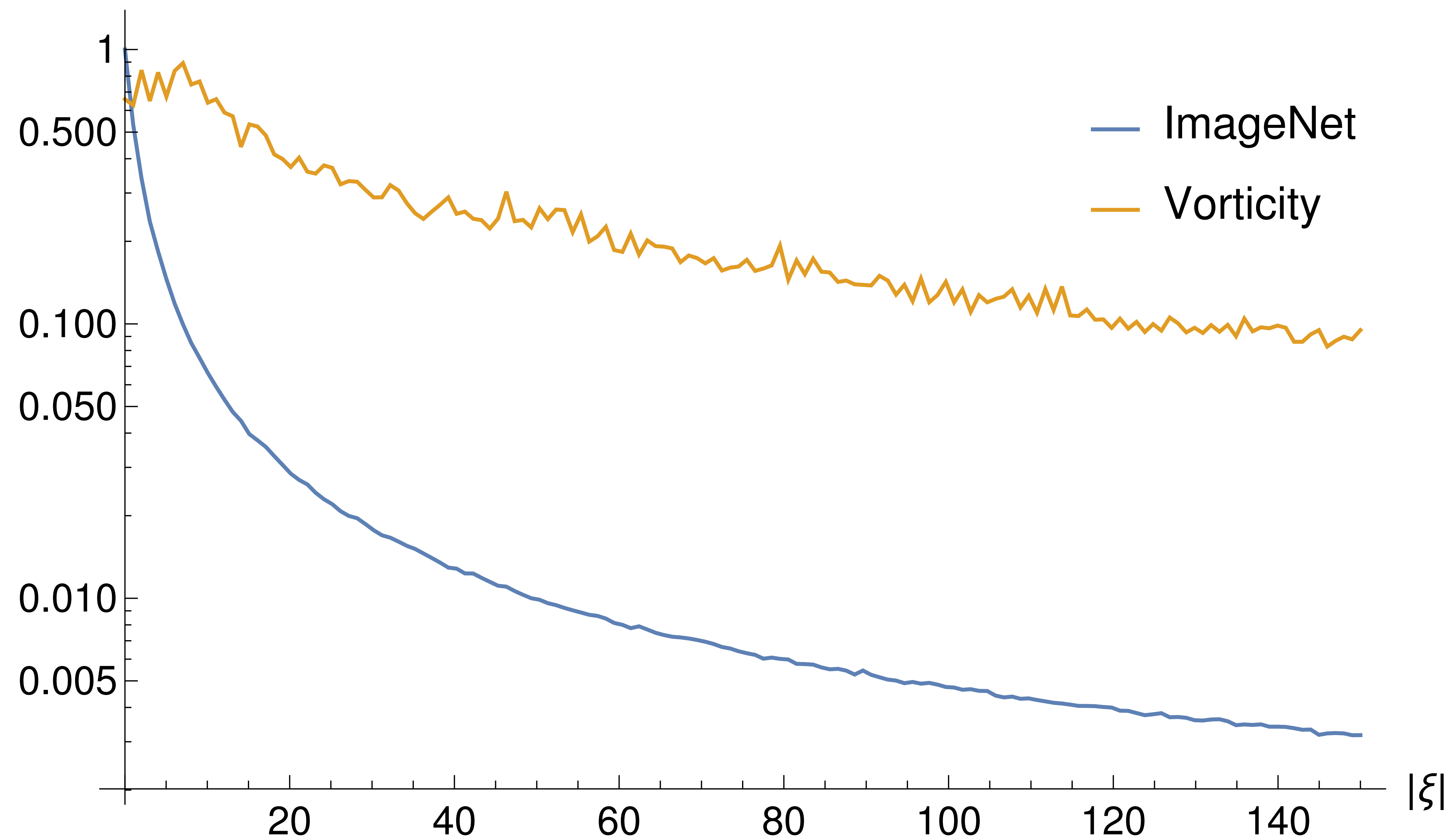
AtmoRep data



AtmoRep data

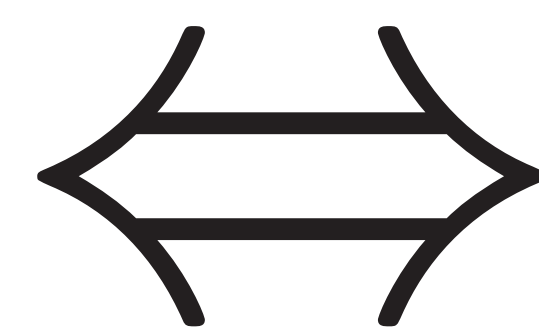


ERA5 versus ImageNet

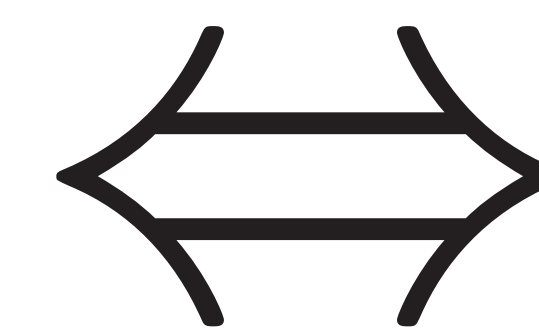


ERA5 versus ImageNet

stream function
velocity potential

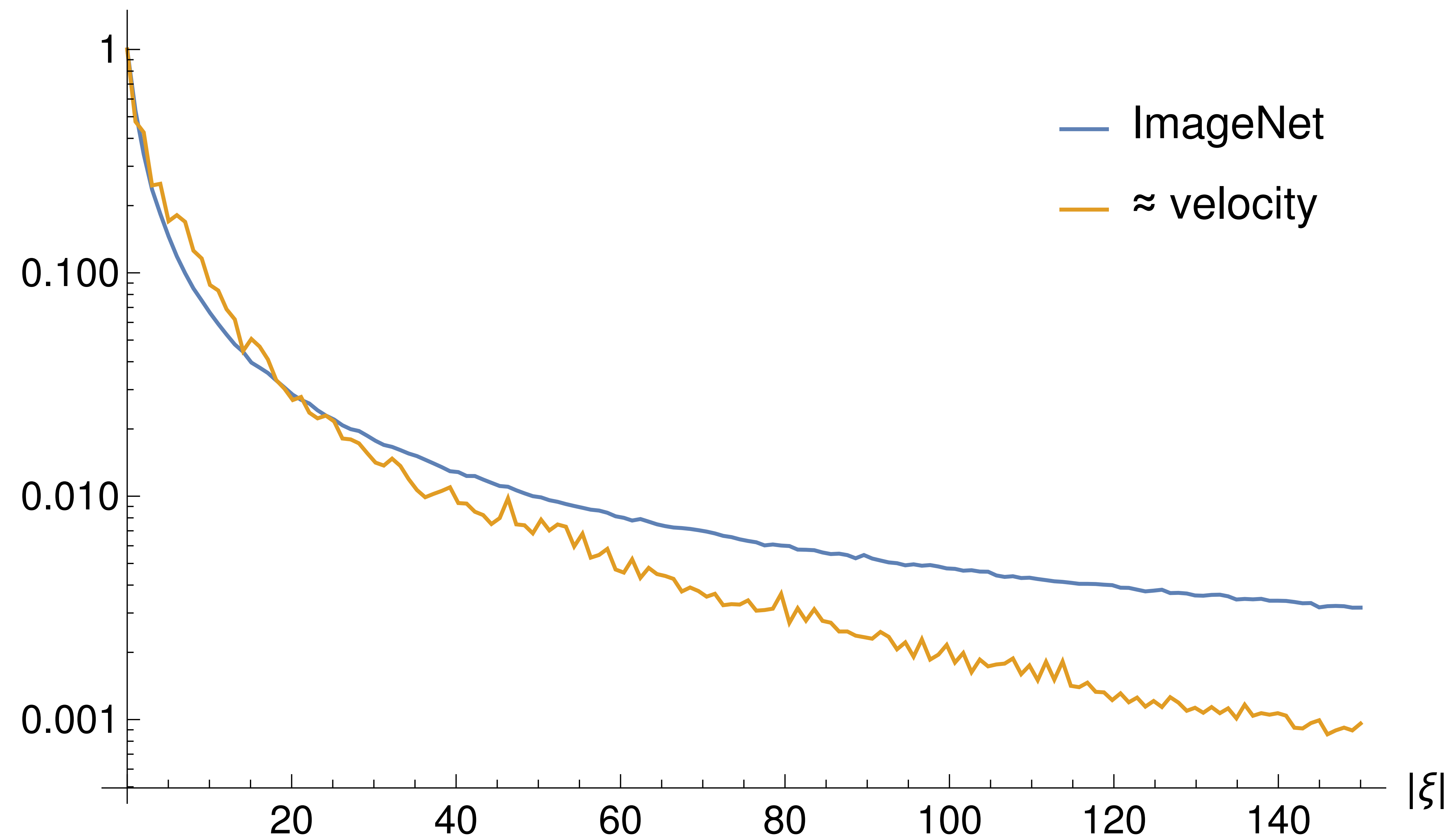


velocity
vector field

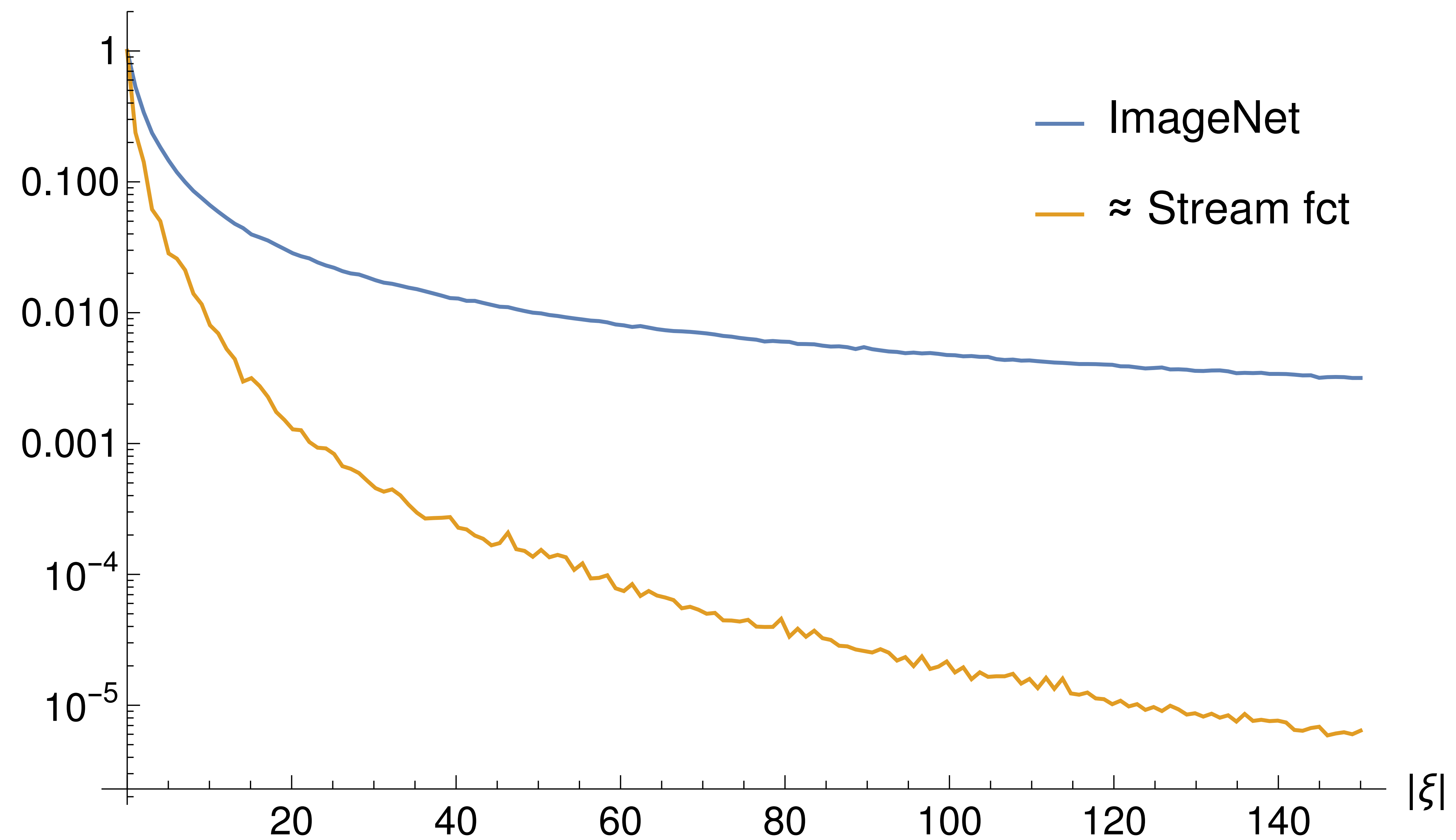


vorticity
divergence

ERA5 versus ImageNet

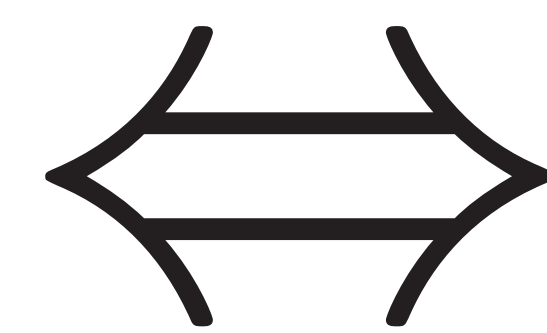


ERA5 versus ImageNet

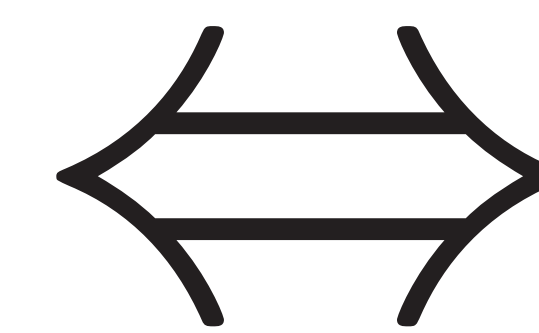


ERA5 versus ImageNet

stream function
velocity potential



velocity
vector field

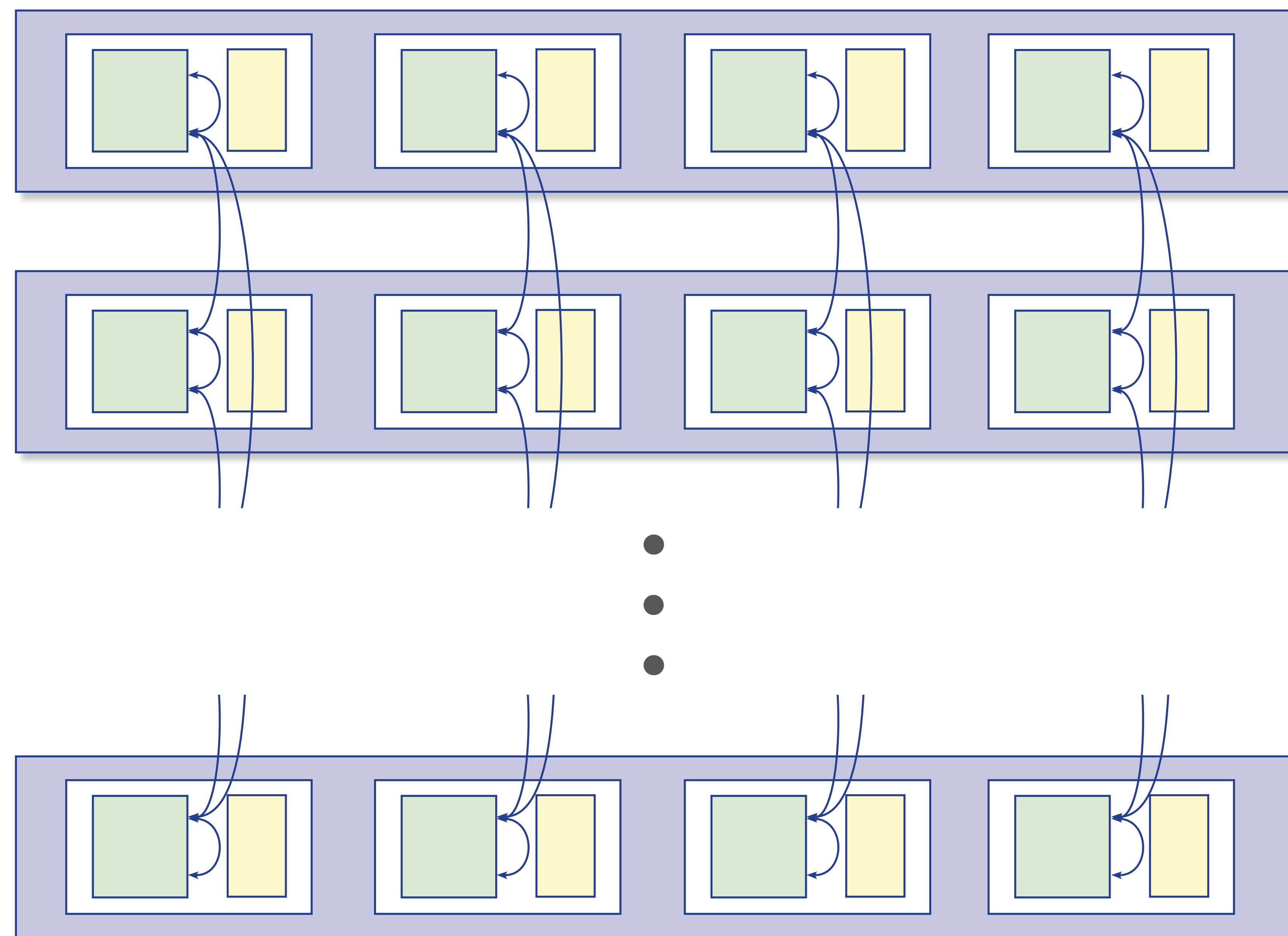


vorticity
divergence

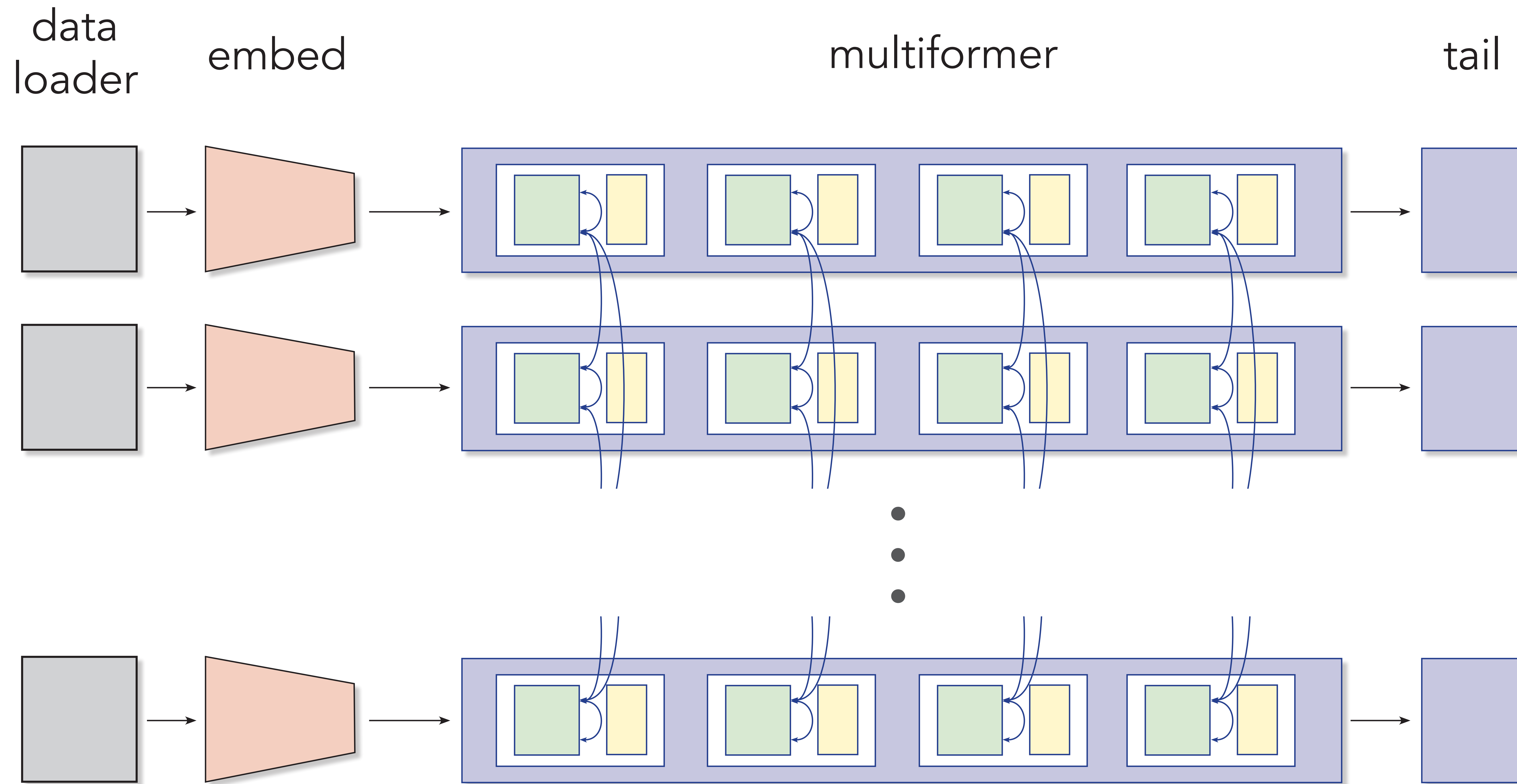
Work with velocity with
norm that emphasizes
small scale detail

Embedding of tokens

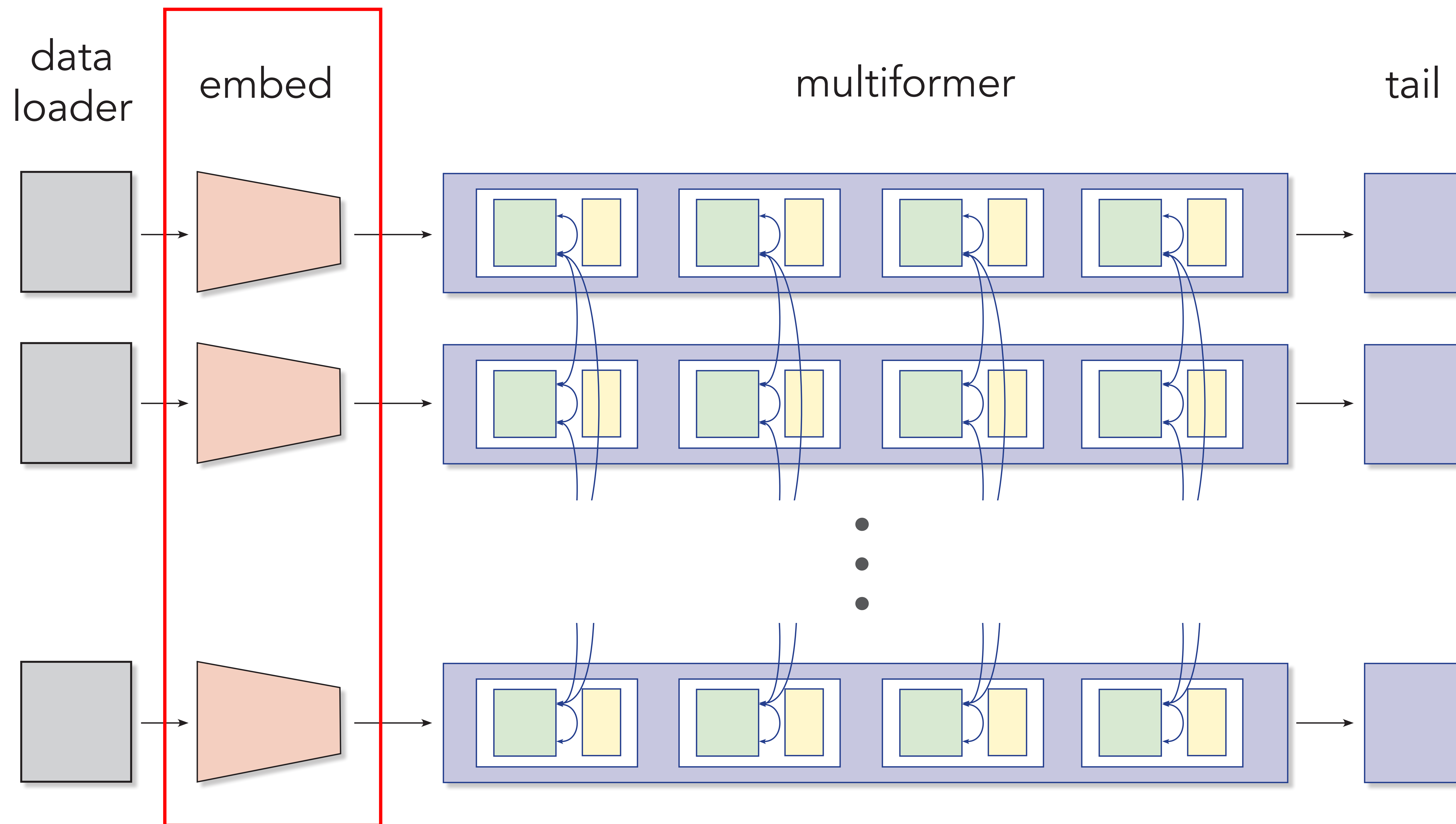
multiformer



Embedding of tokens



Embedding of tokens



Embedding network

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

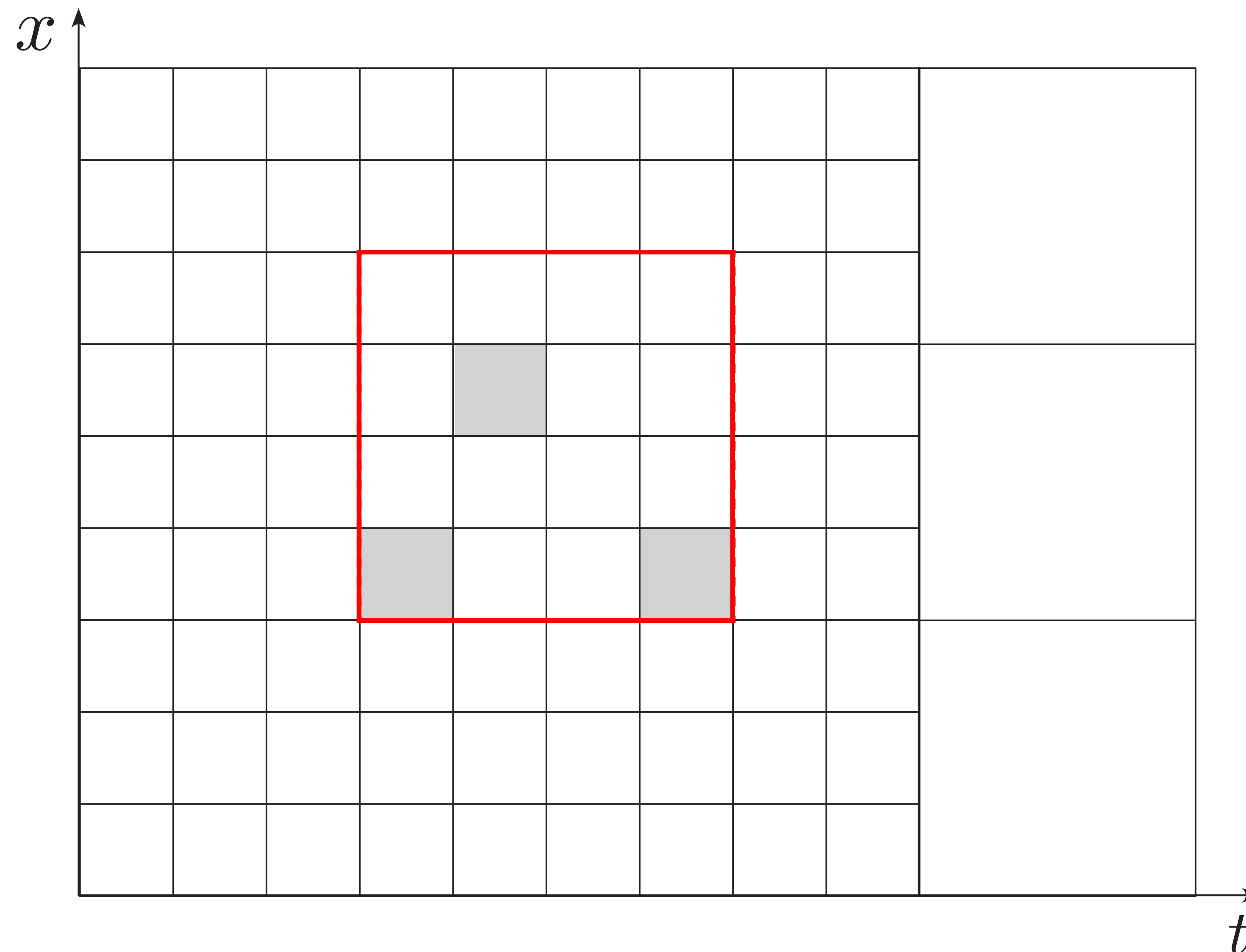
Embedding network

- 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



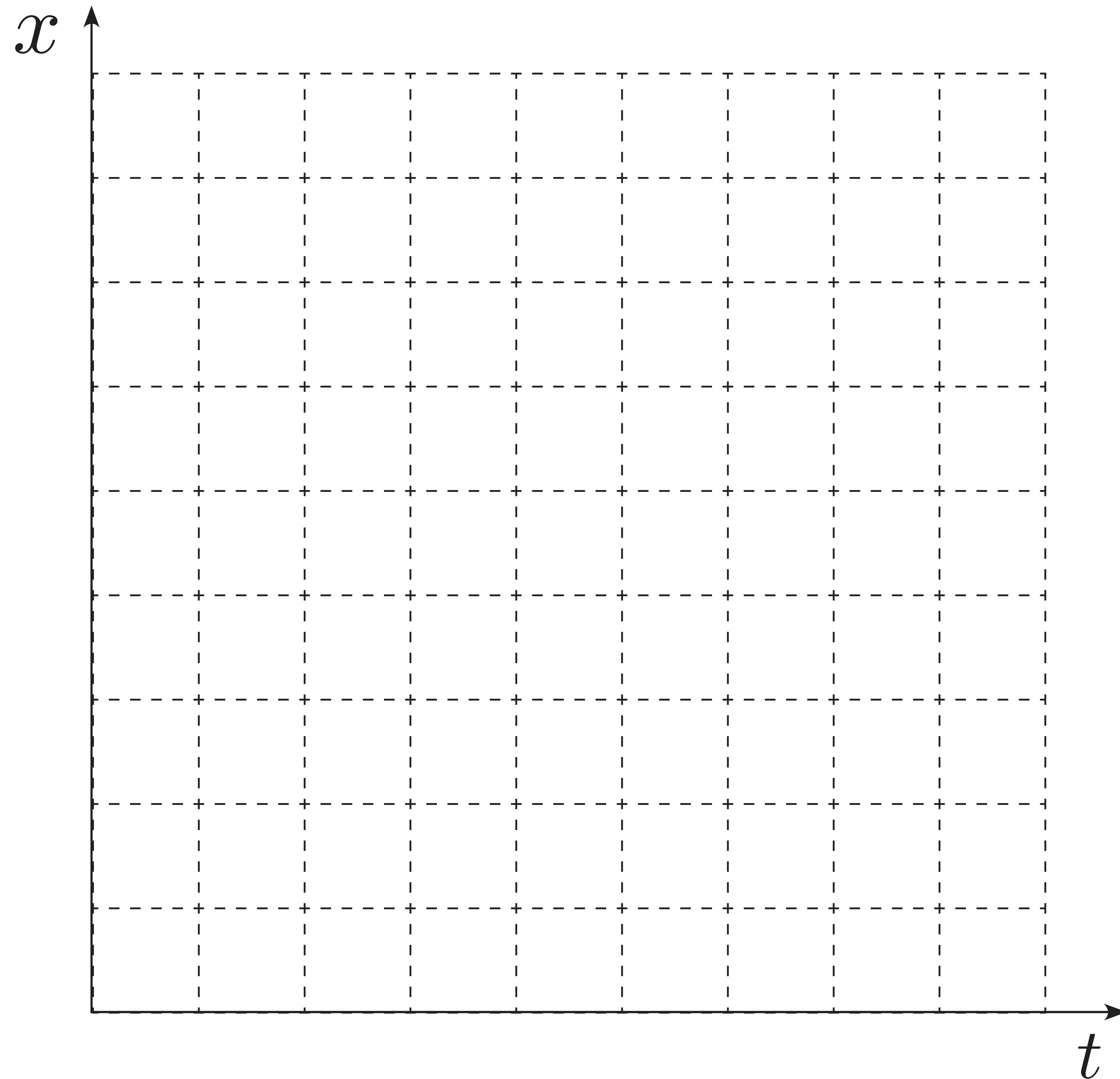
Embedding of tokens

- 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

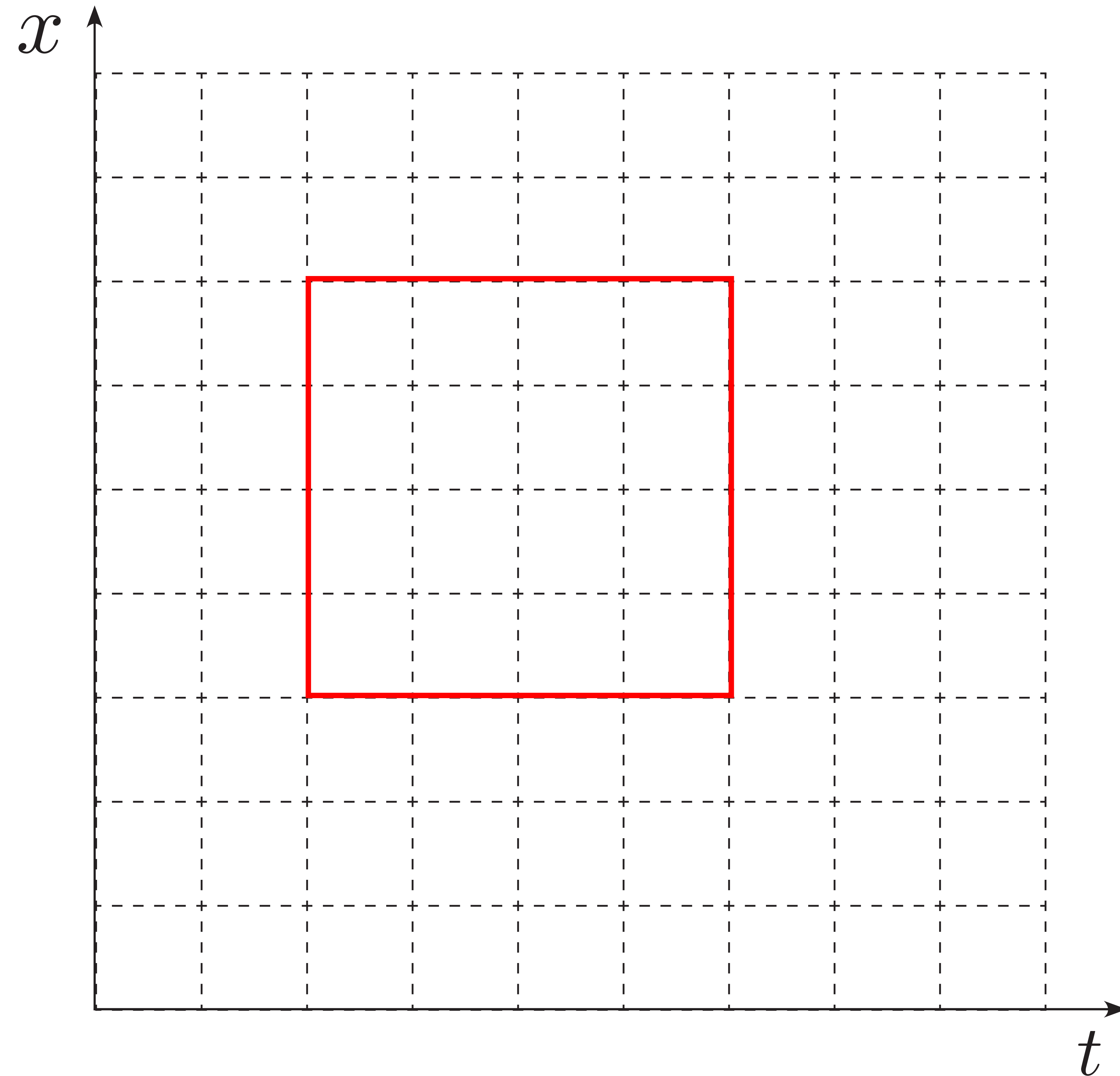
Embedding of tokens

- 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
- ⇒ Use transformer as embedding network

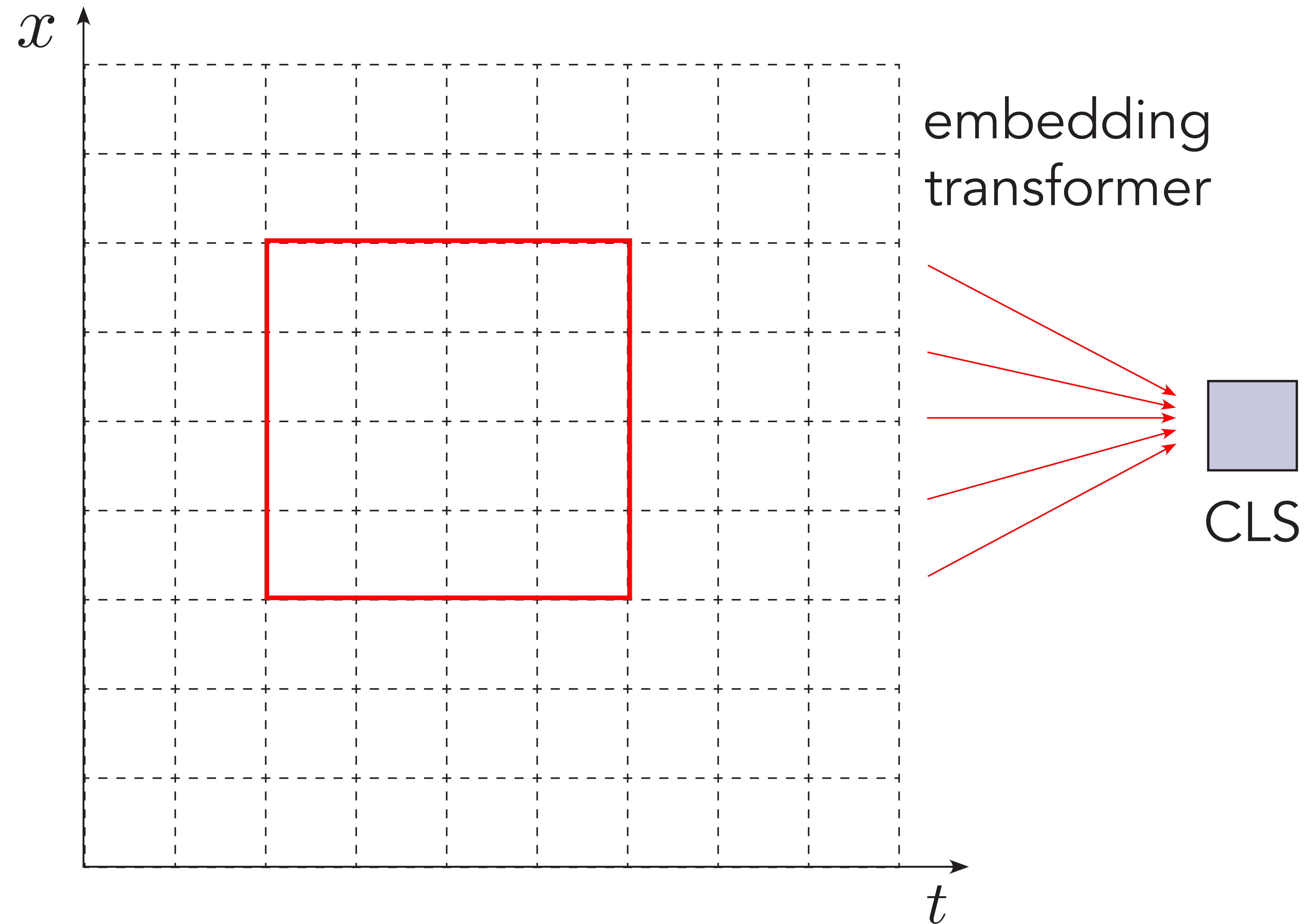
Embedding of tokens



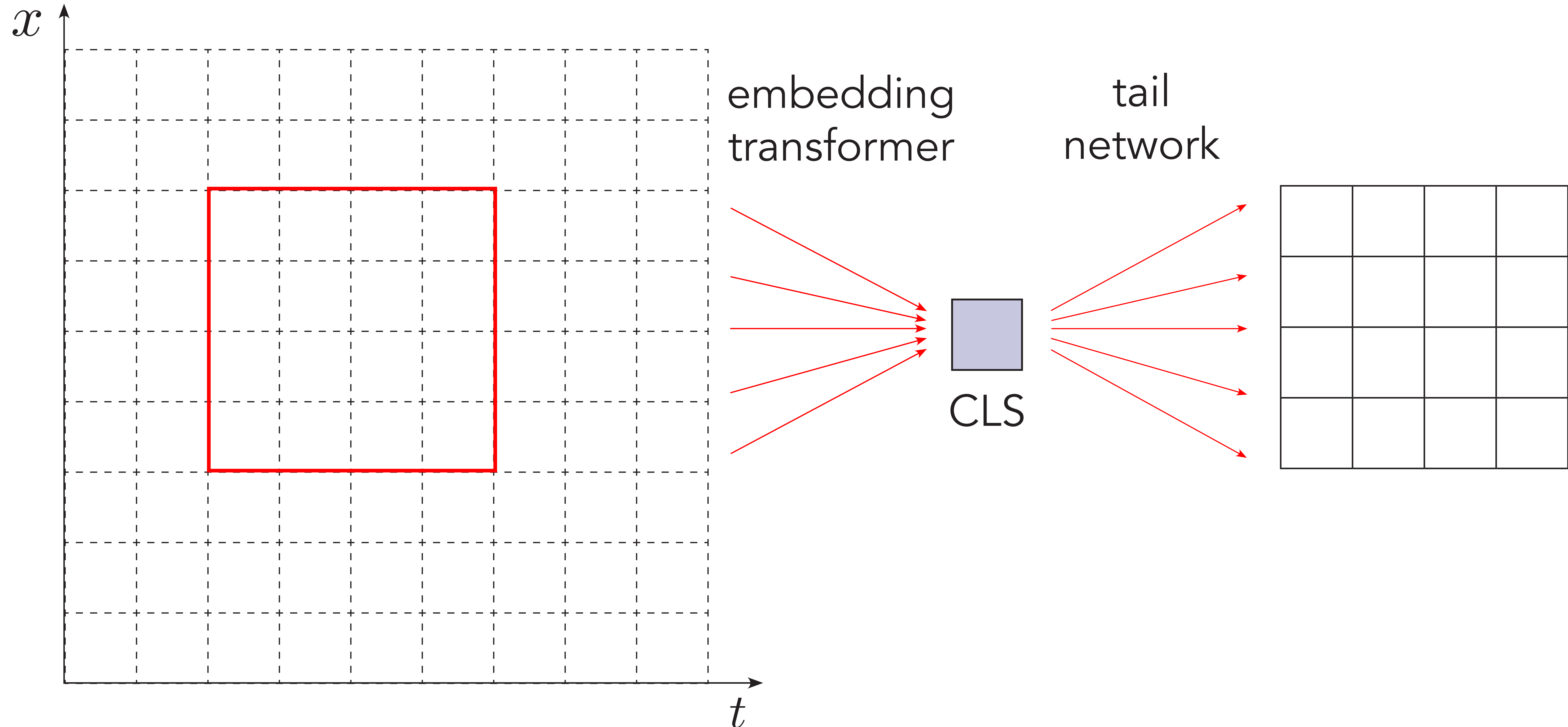
Embedding of tokens



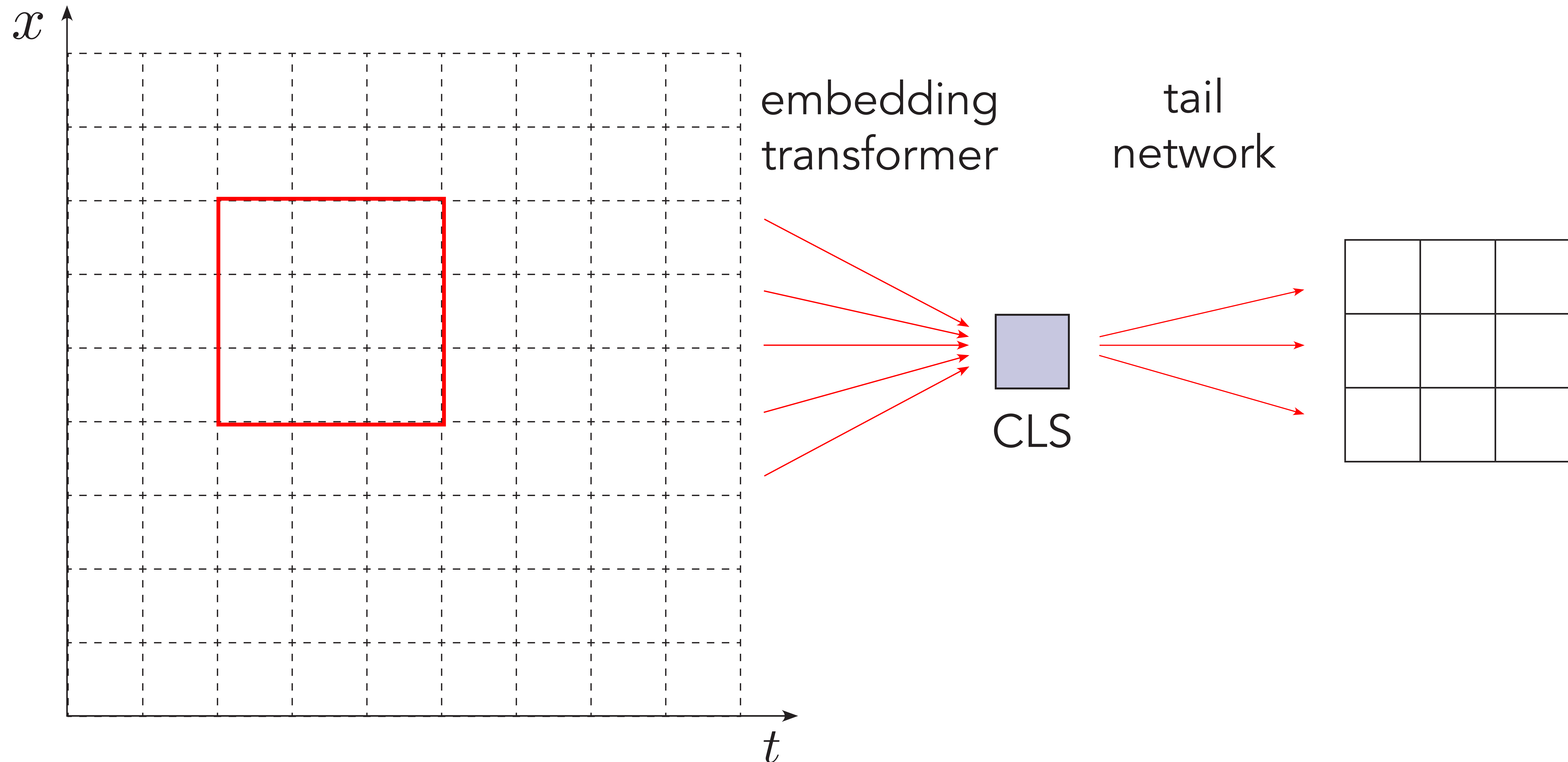
Embedding of tokens



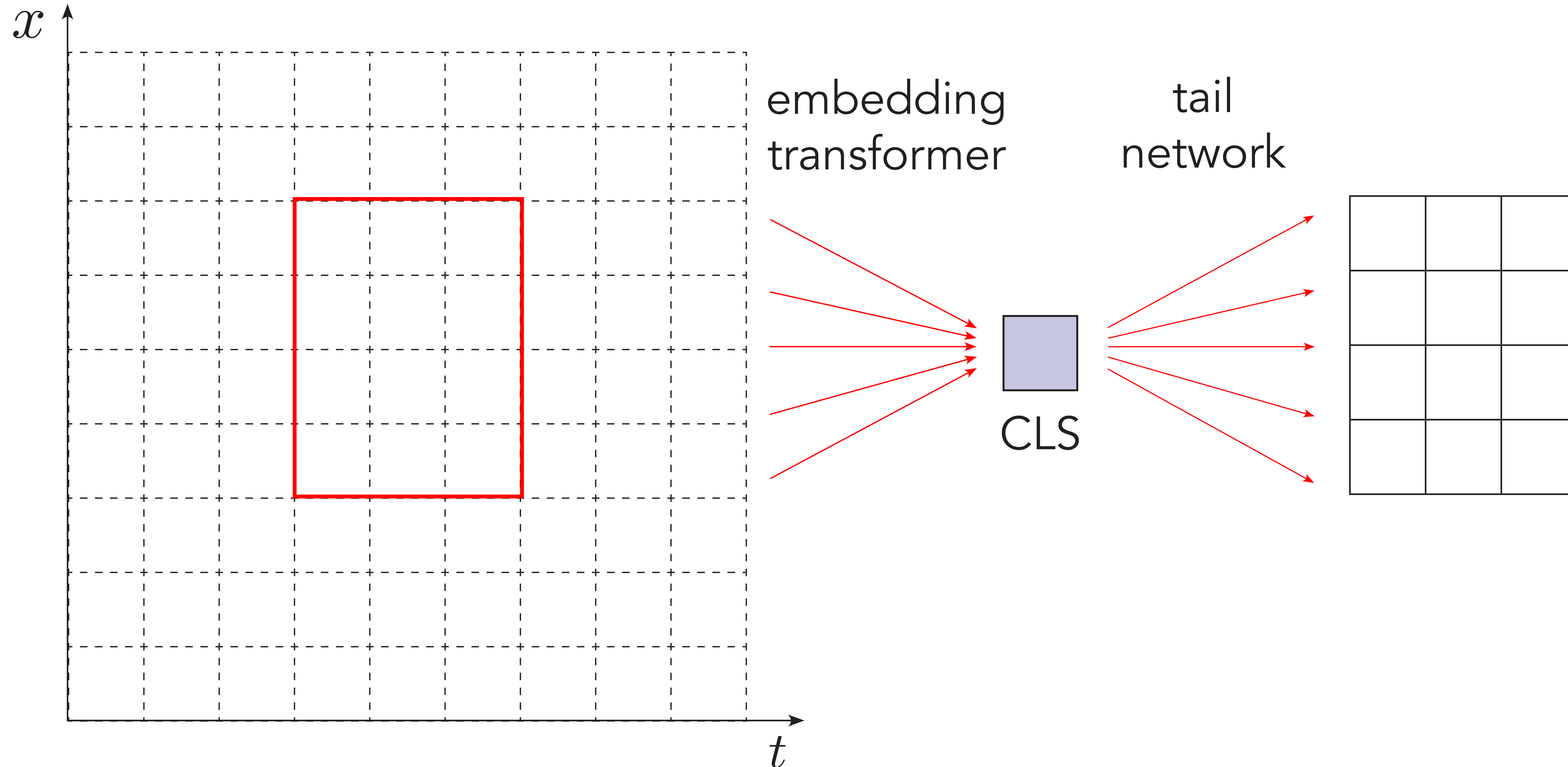
Embedding of tokens



Embedding of tokens



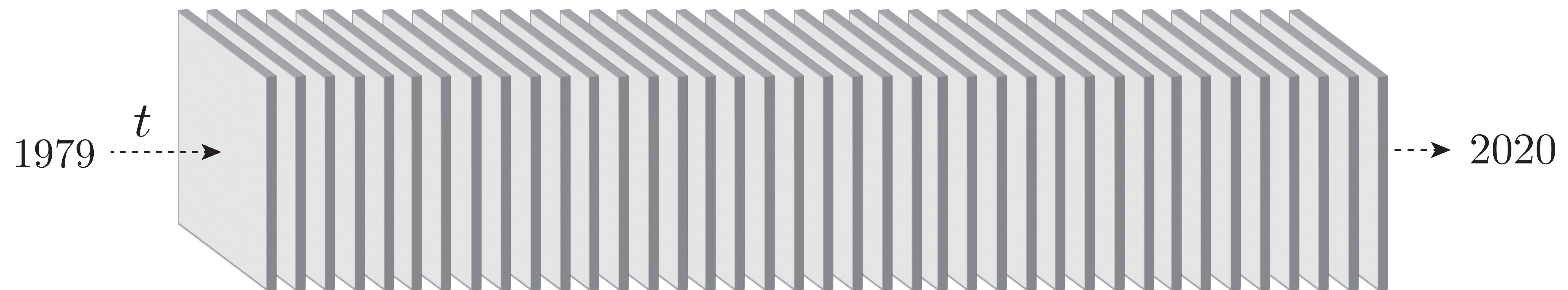
Embedding of tokens



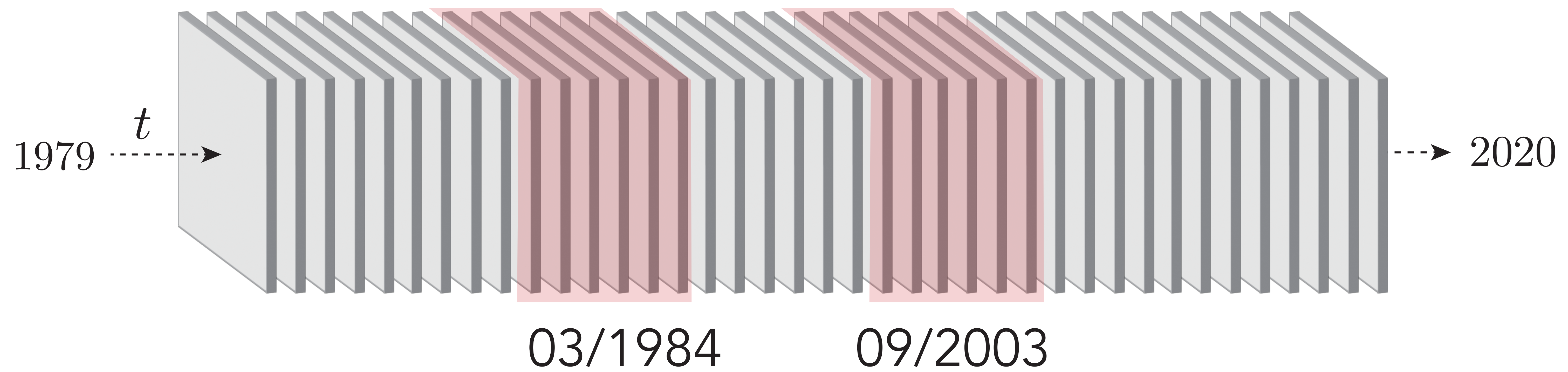
Training

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

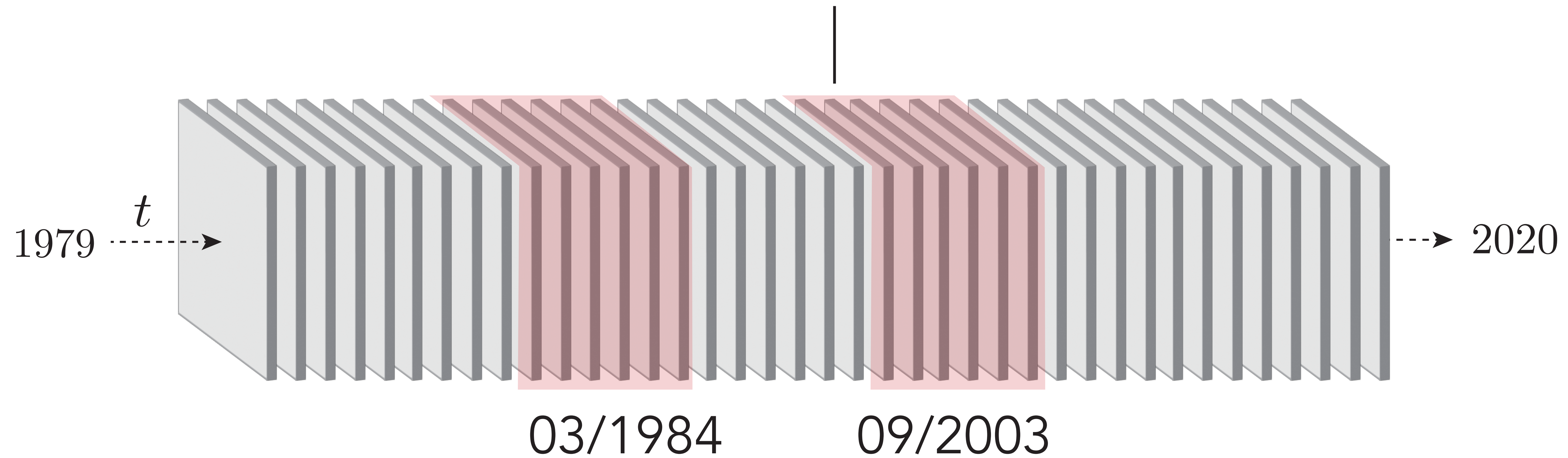
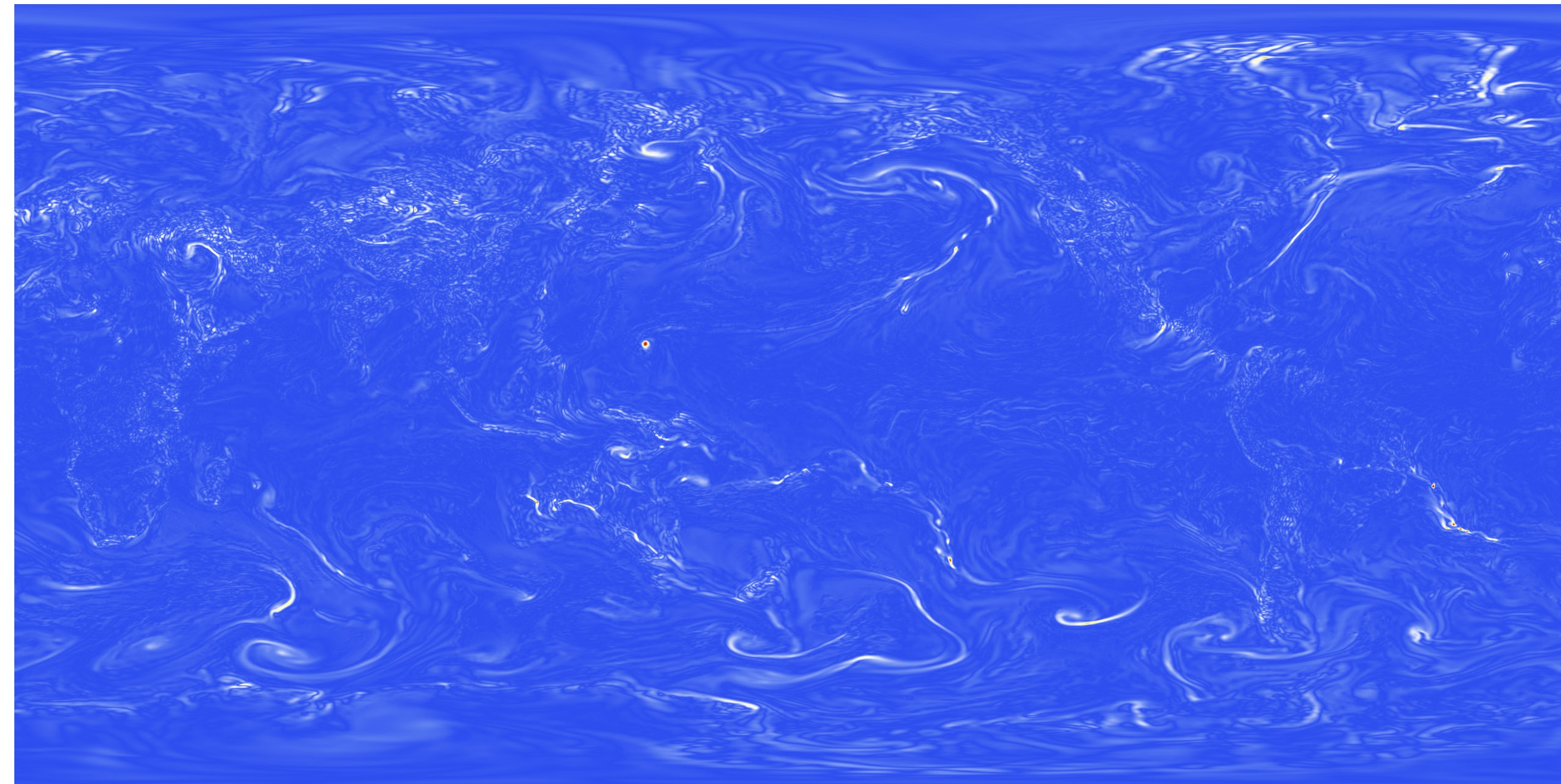
Training



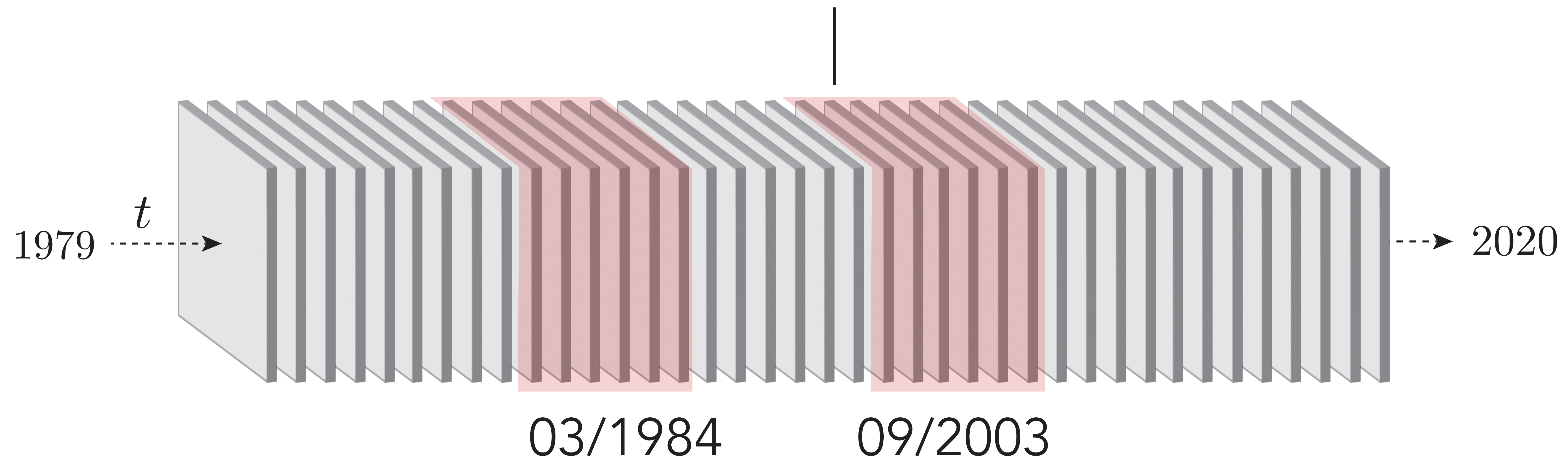
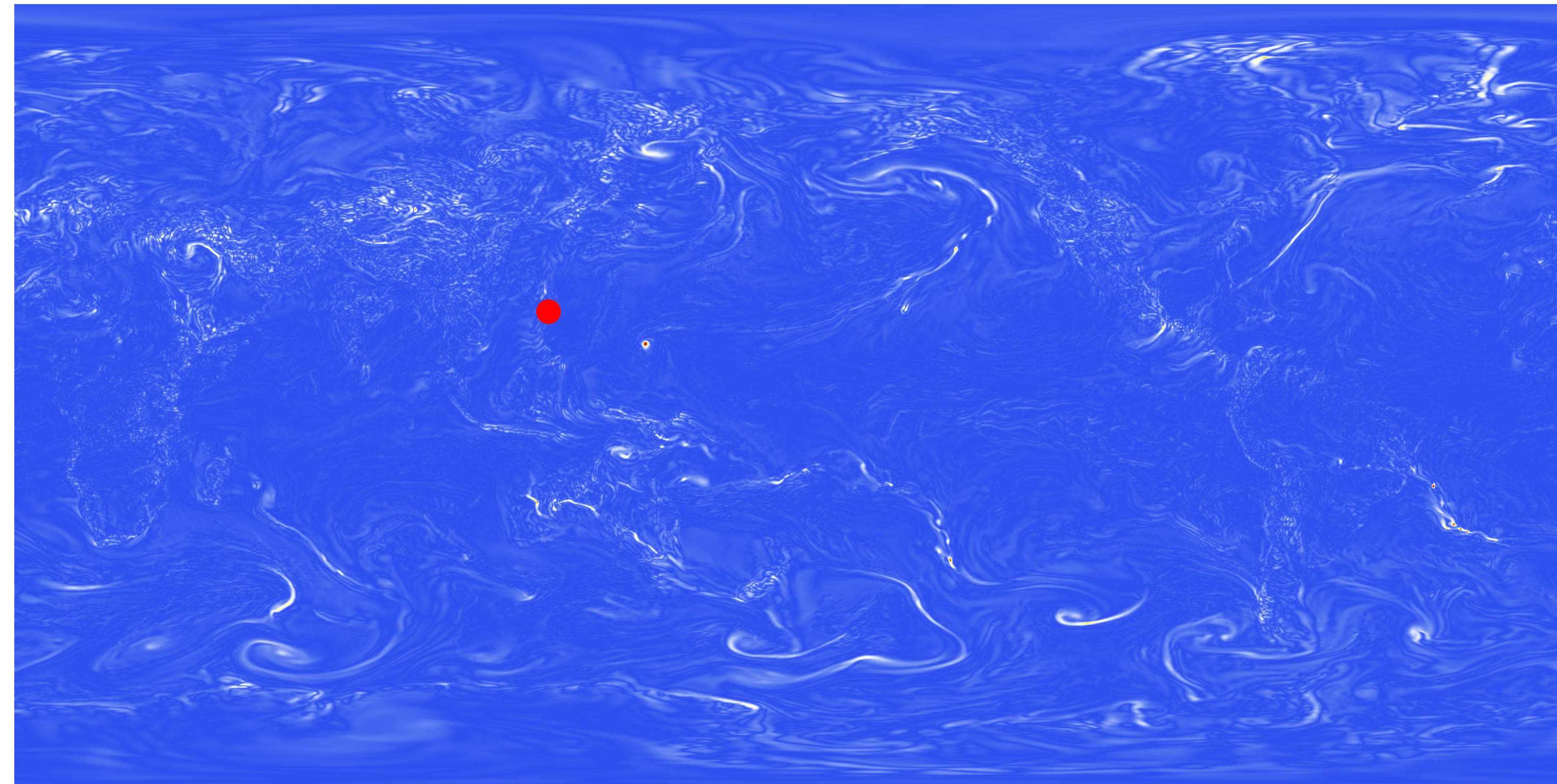
Training



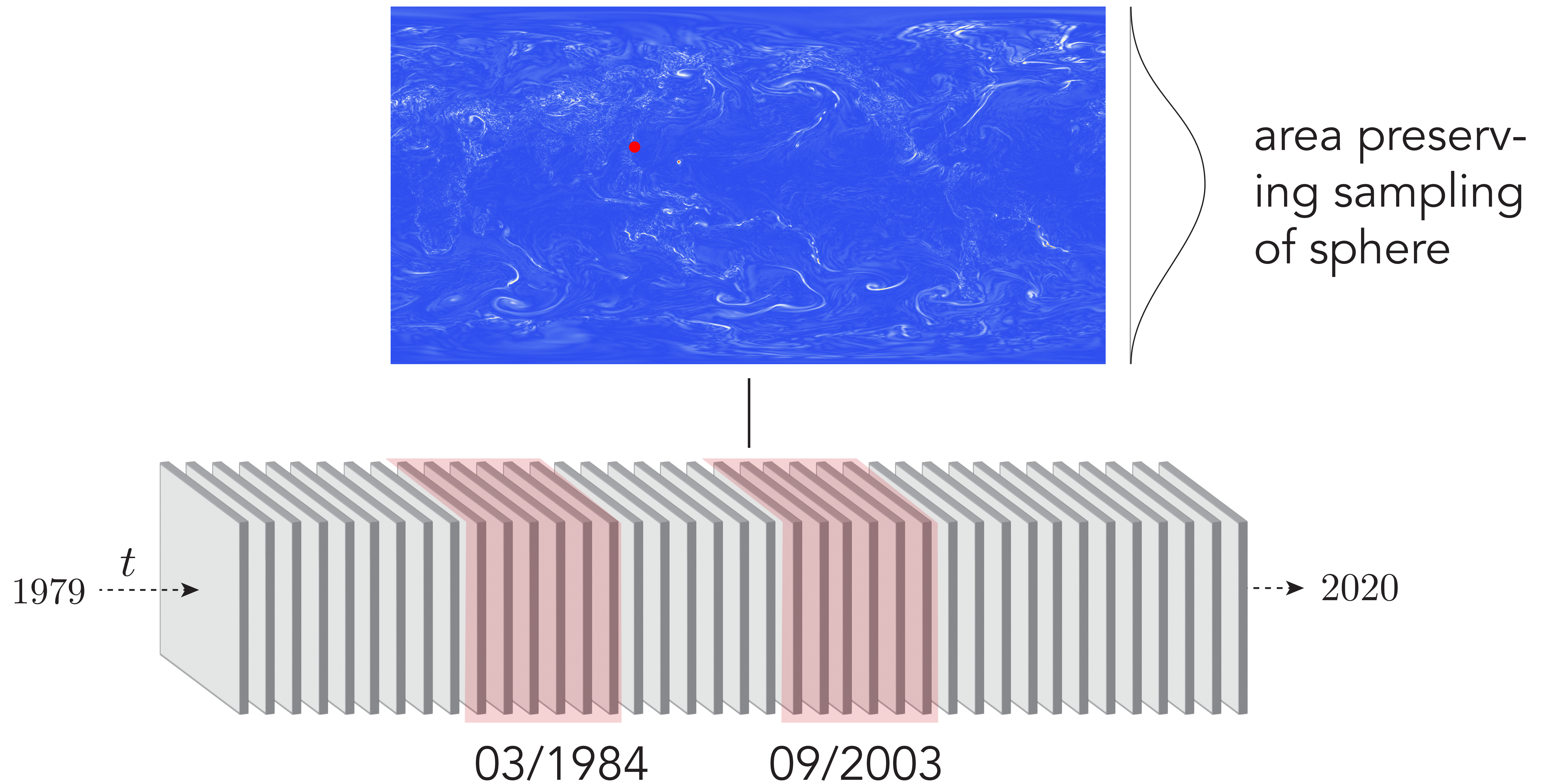
Training



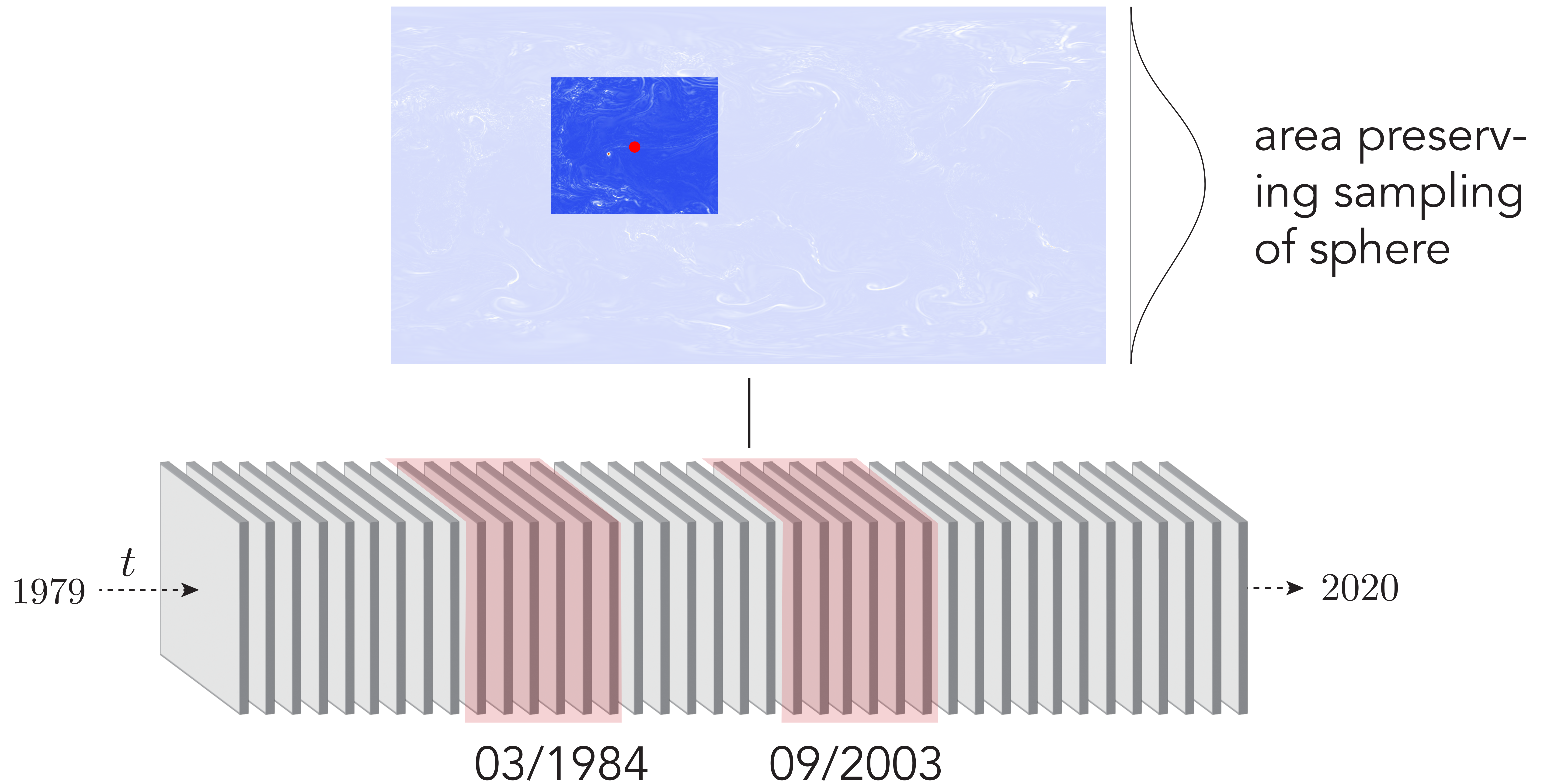
Training



Training



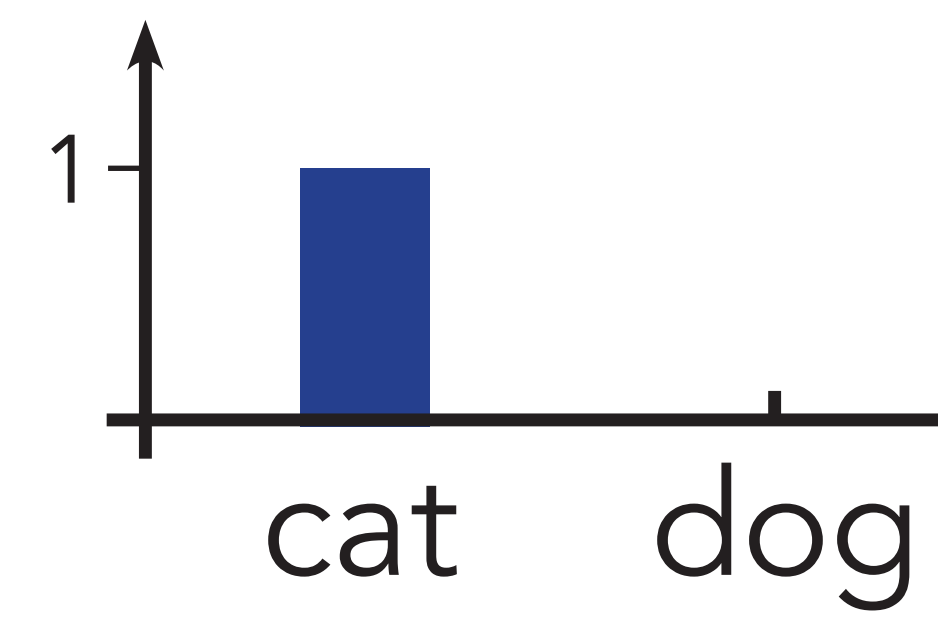
Training



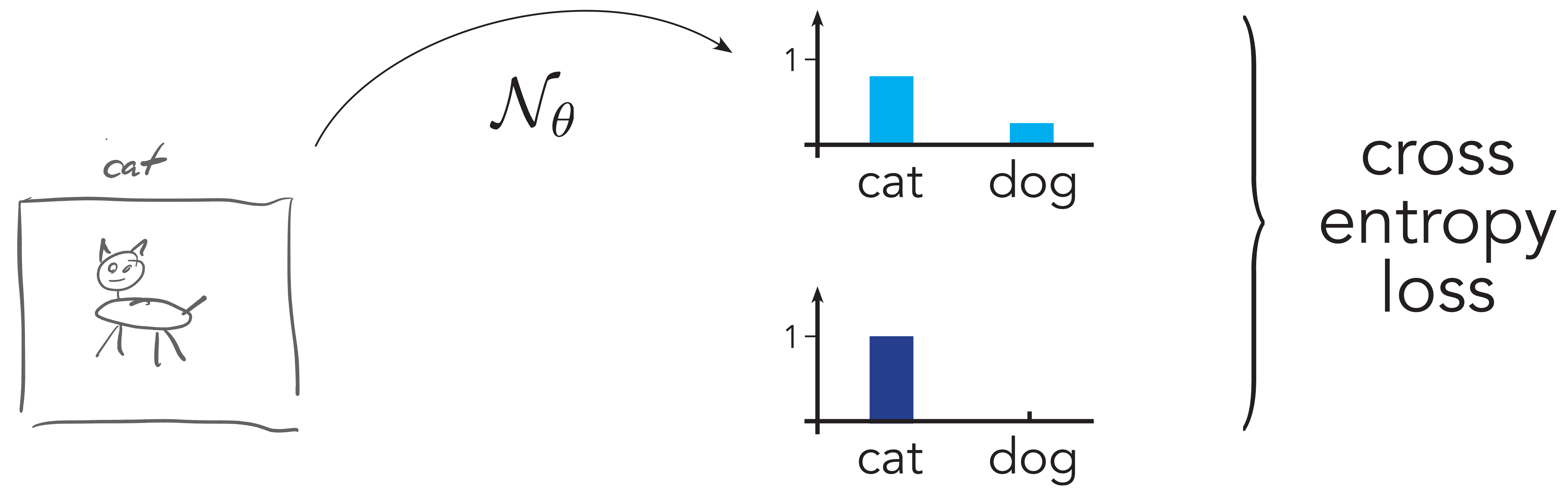
Statistical loss



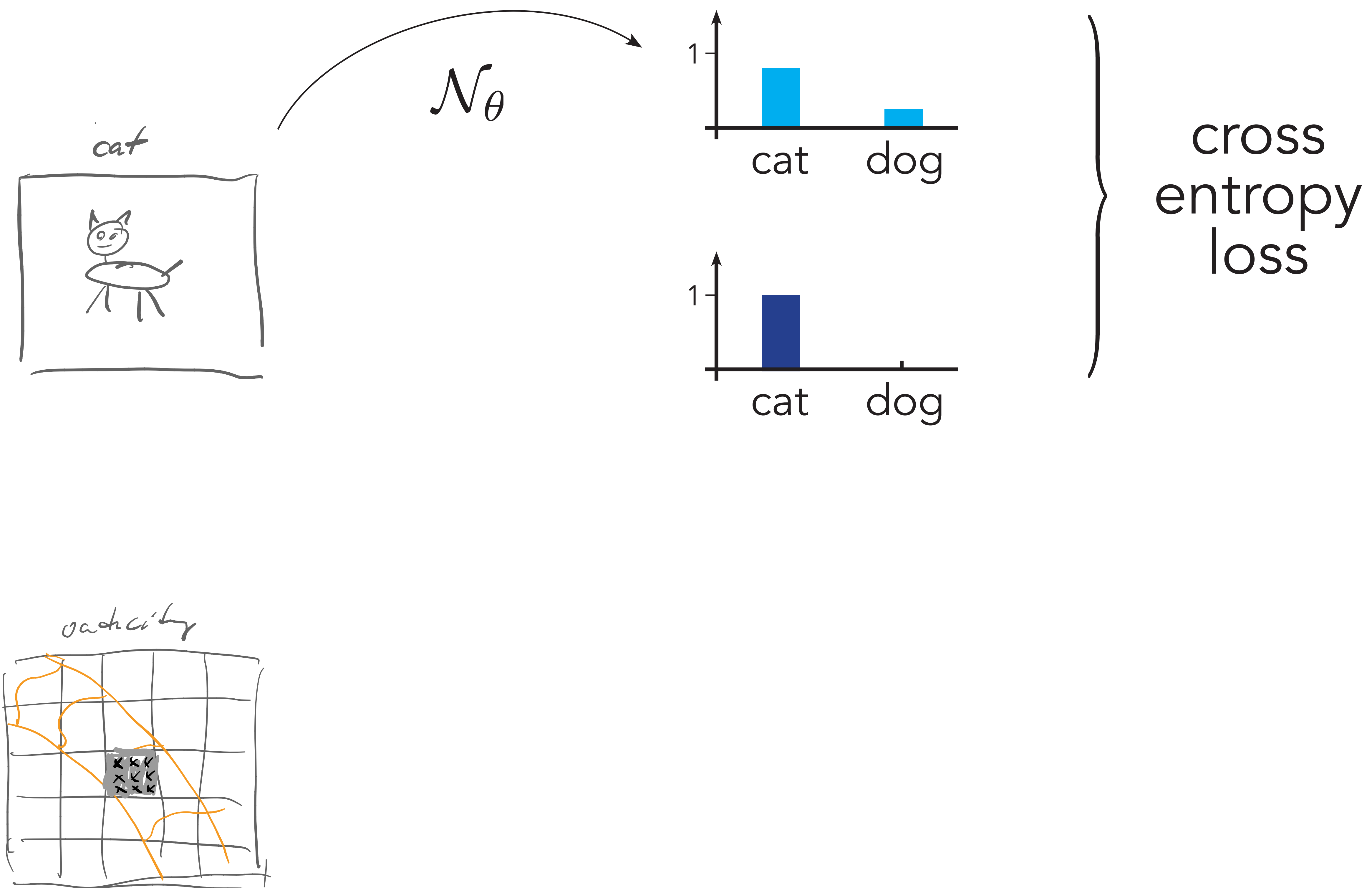
Statistical loss



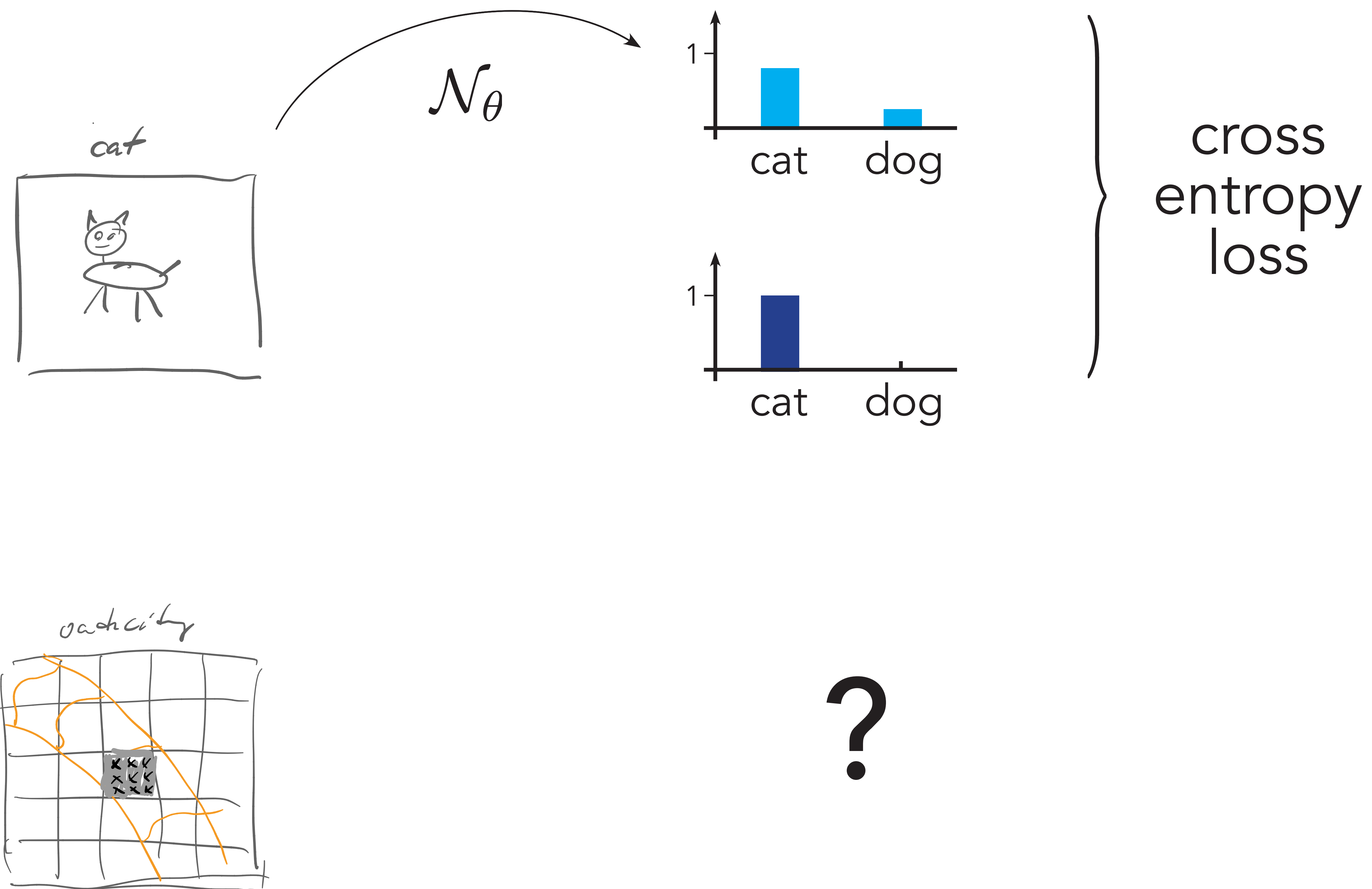
Statistical loss



Statistical loss



Statistical loss



Statistical loss

- Statistical loss:

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

Statistical loss

- Statistical loss:

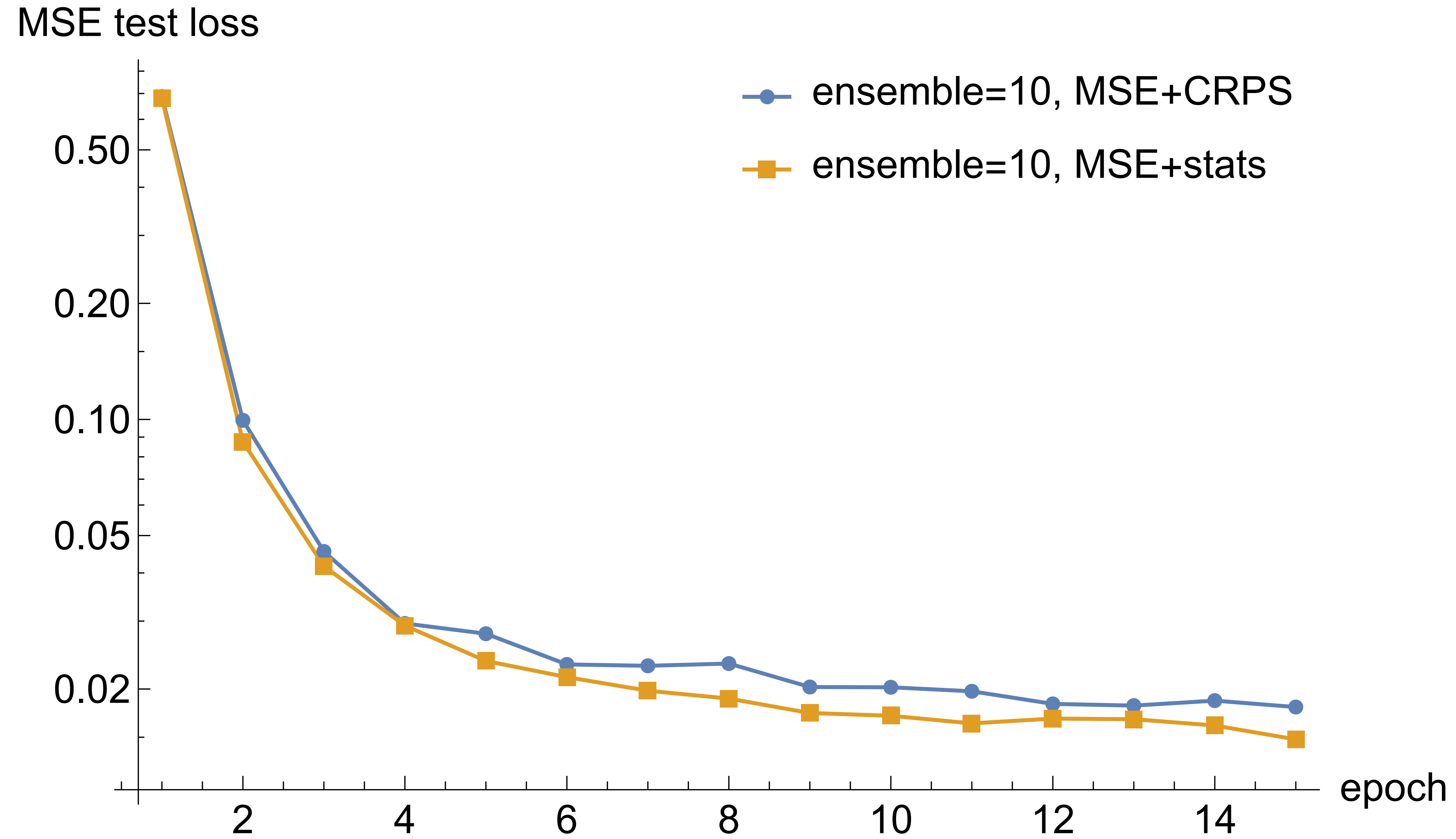
$$\mathcal{L}_{\text{stats}} = \left| 1 - \int_{\mathbb{R}} \delta_y(x) G_{\tilde{\mu}, \tilde{\sigma}}(x) dx \right|^2$$

- CRPS:¹

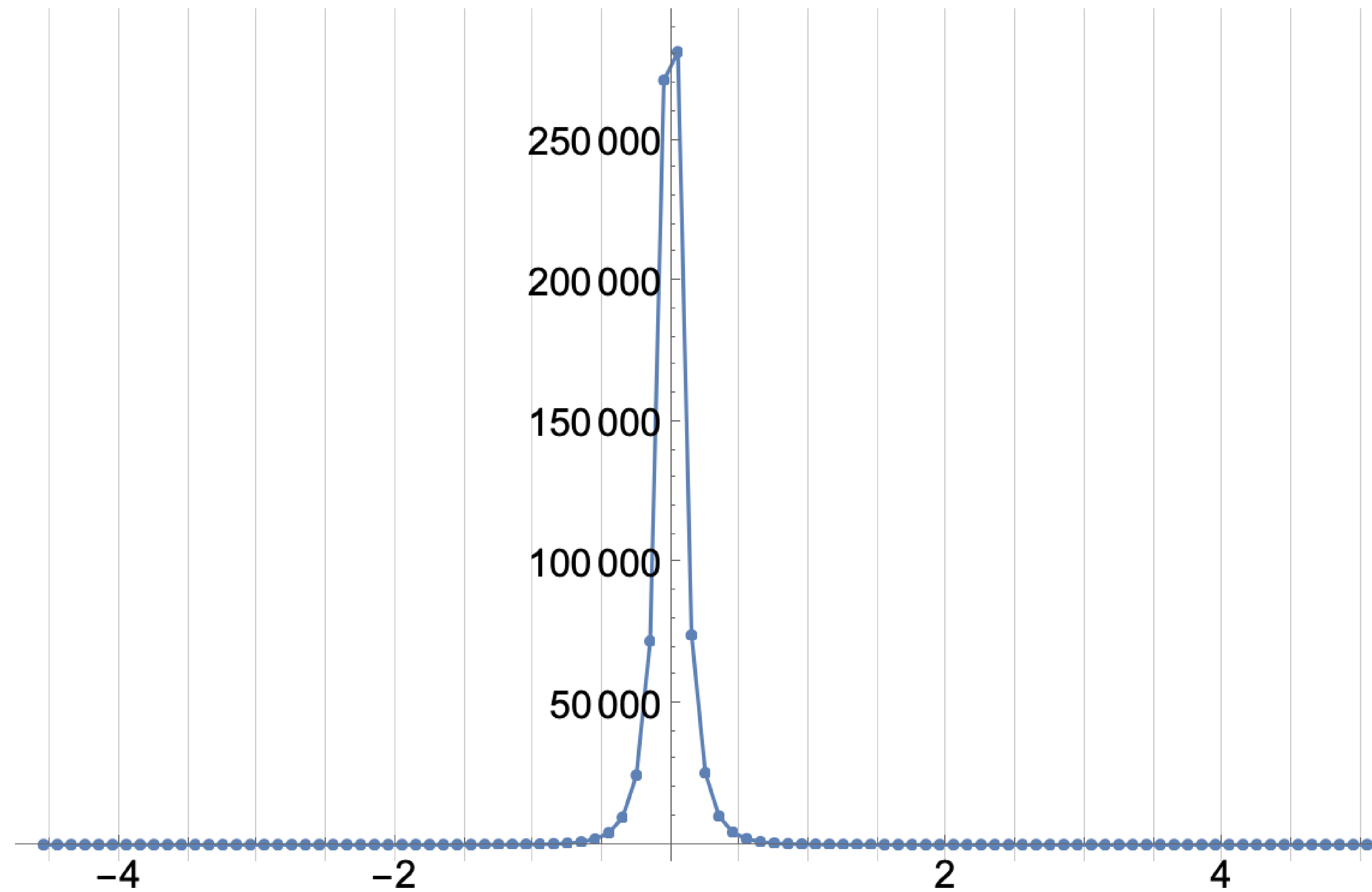
$$\mathcal{L}_{\text{CRPS}} = \int_{\mathbb{R}} \left| H_y(x) \text{erf}_{\tilde{\mu}, \tilde{\sigma}}(x) \right|^2 dx$$

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

Statistical loss



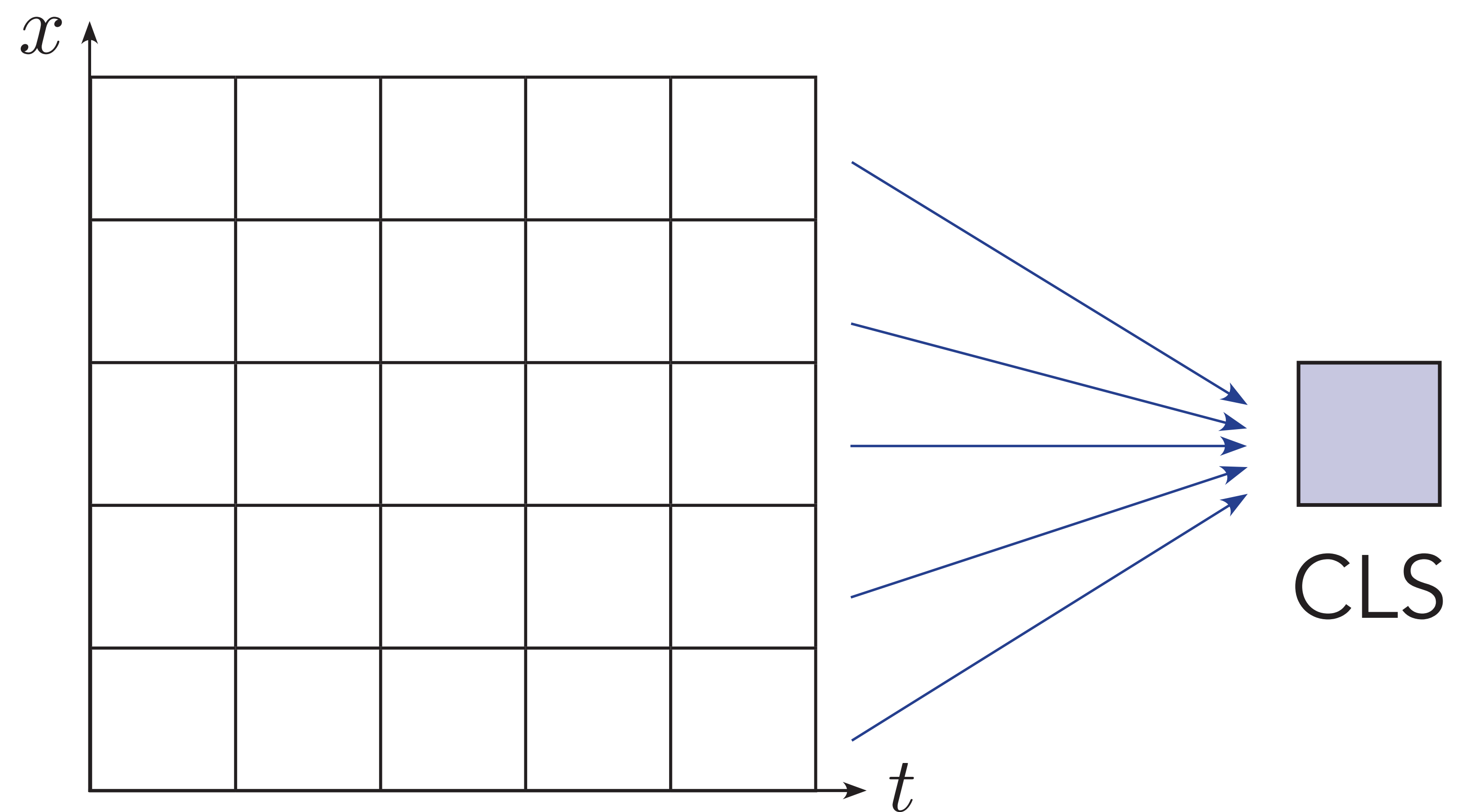
Statistical loss



Histogram
of ensemble
errors

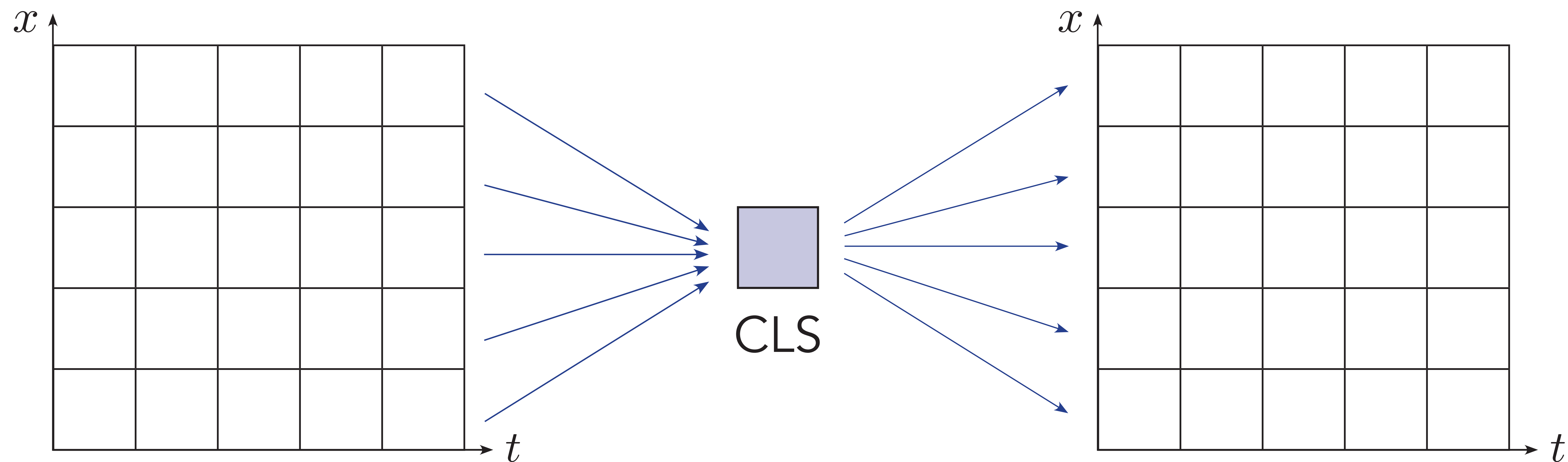
Zero shot evaluation

Embedding



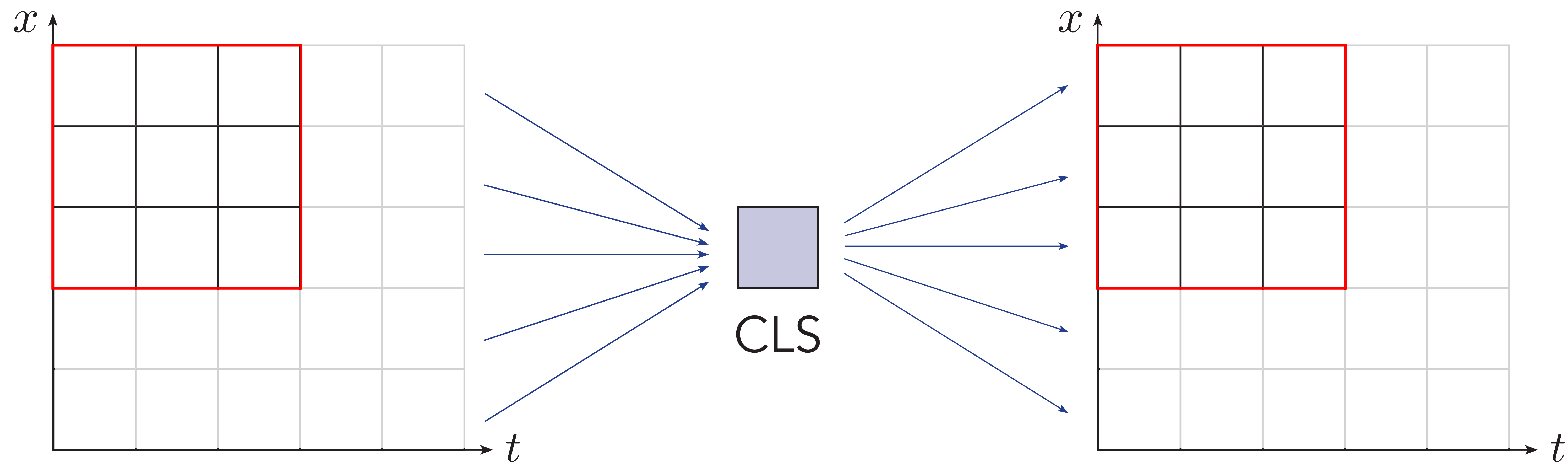
Zero shot evaluation

Embedding



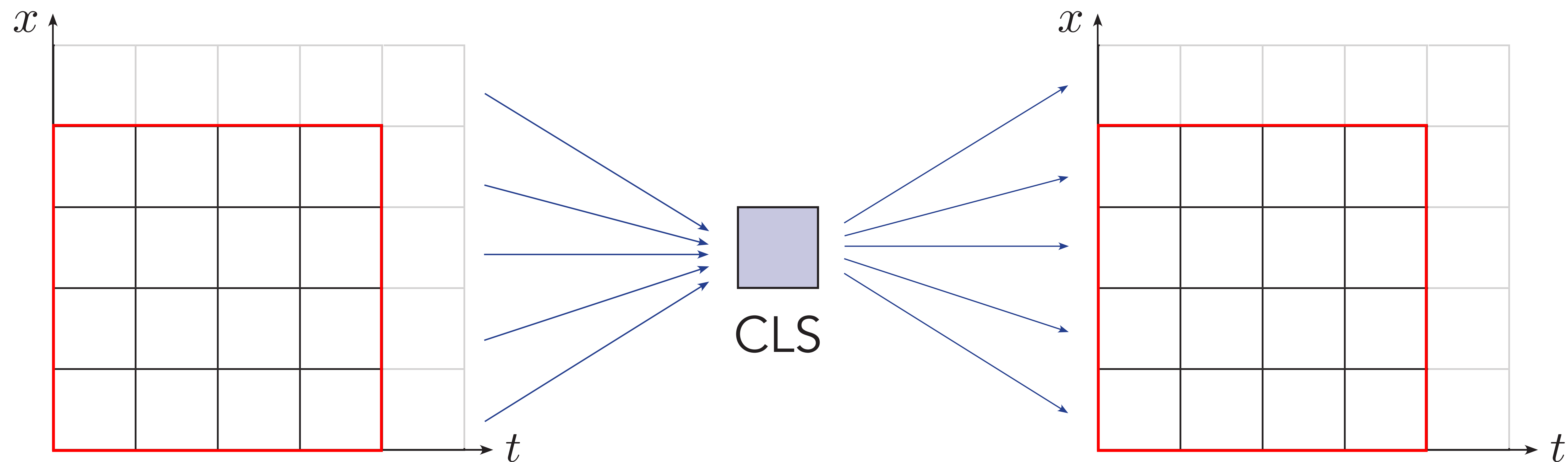
Zero shot evaluation

Embedding



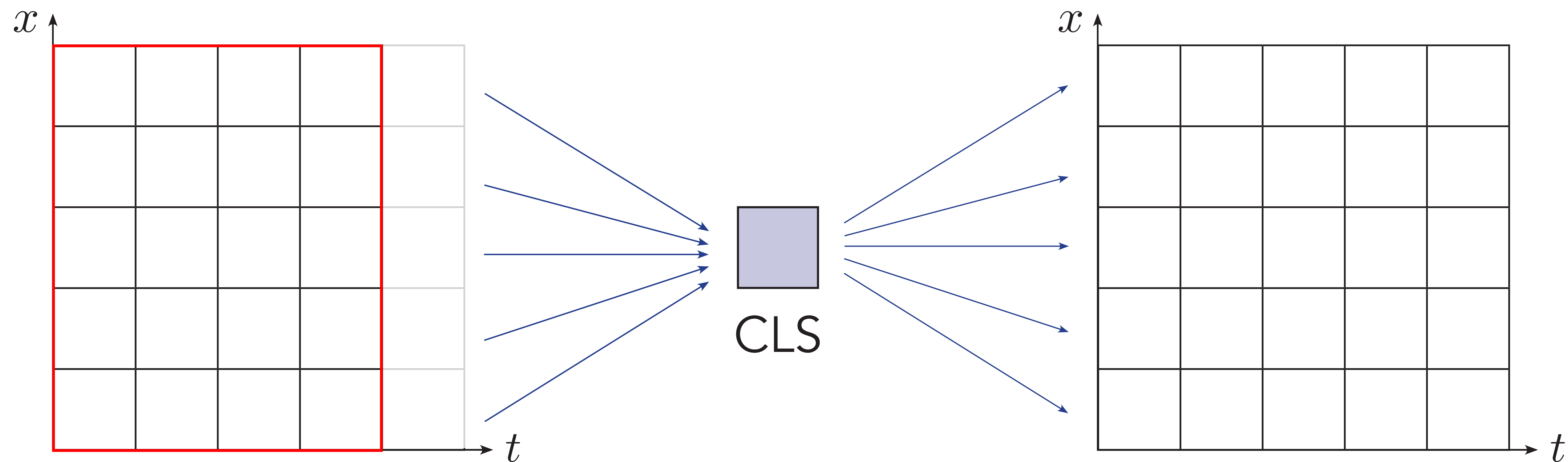
Zero shot evaluation

Embedding



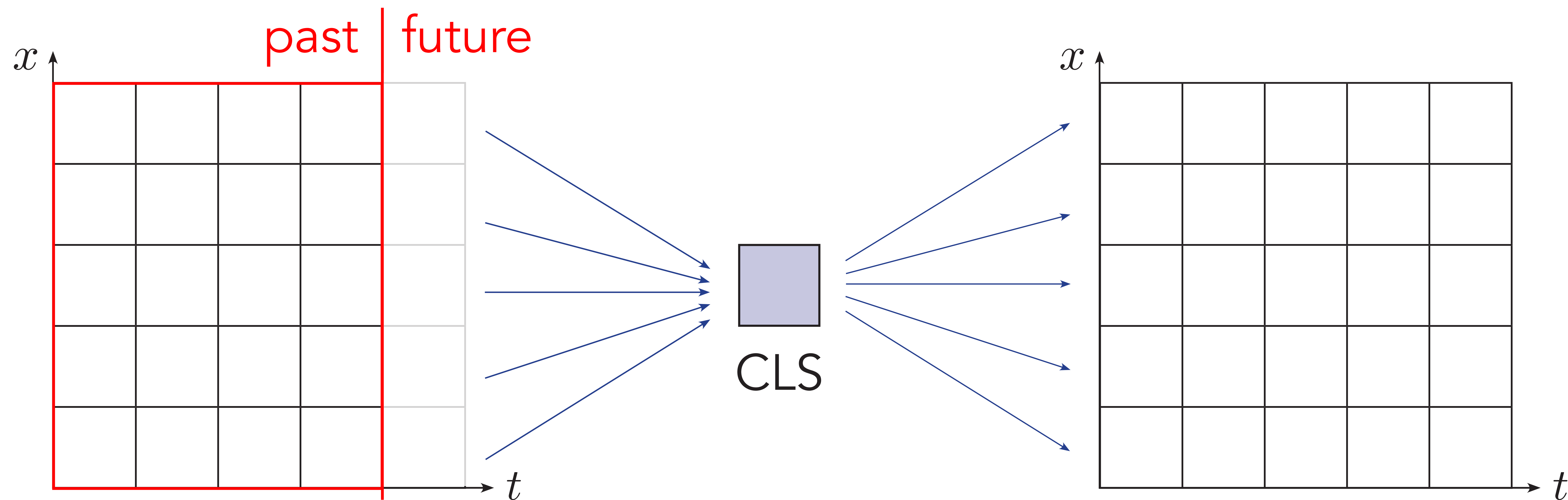
Zero shot evaluation

Embedding-Forecast



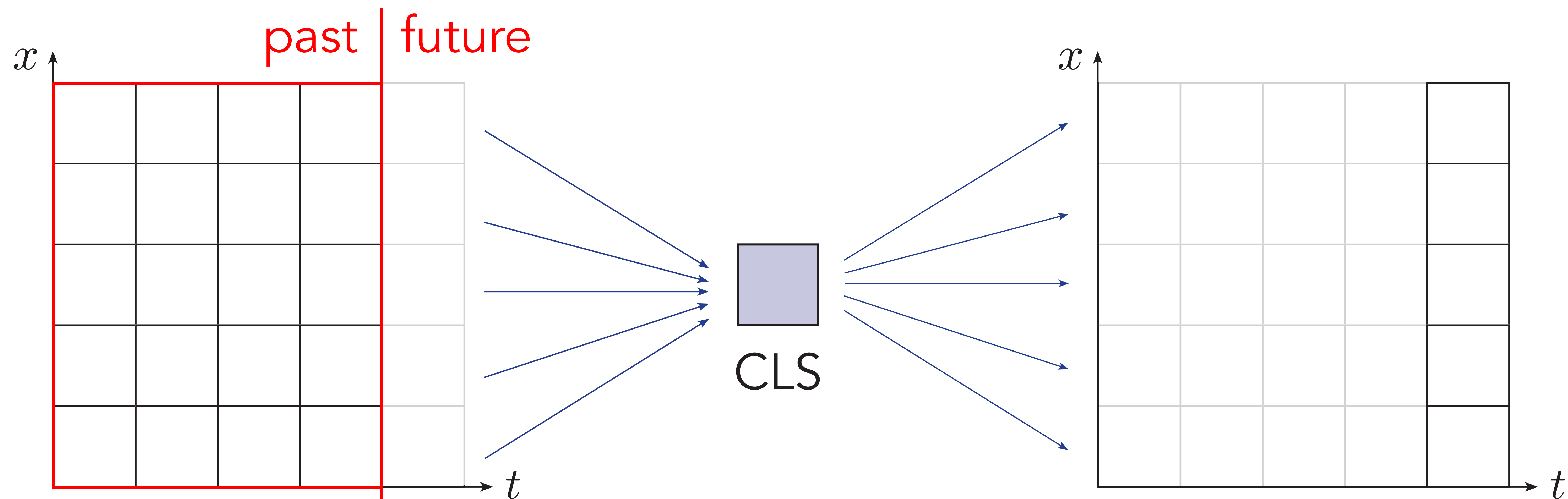
Zero shot evaluation

Embedding-Forecast

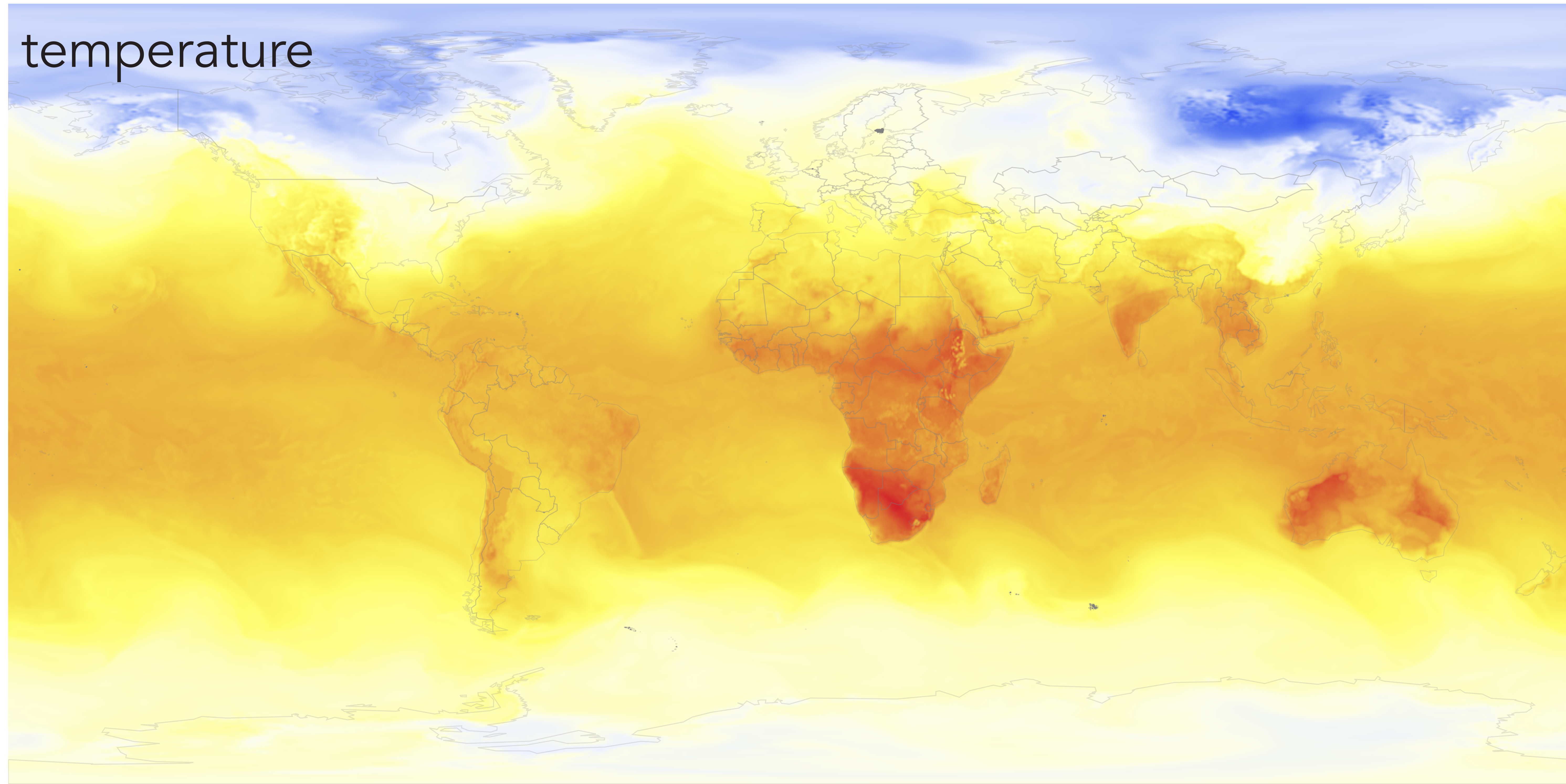


Zero shot evaluation

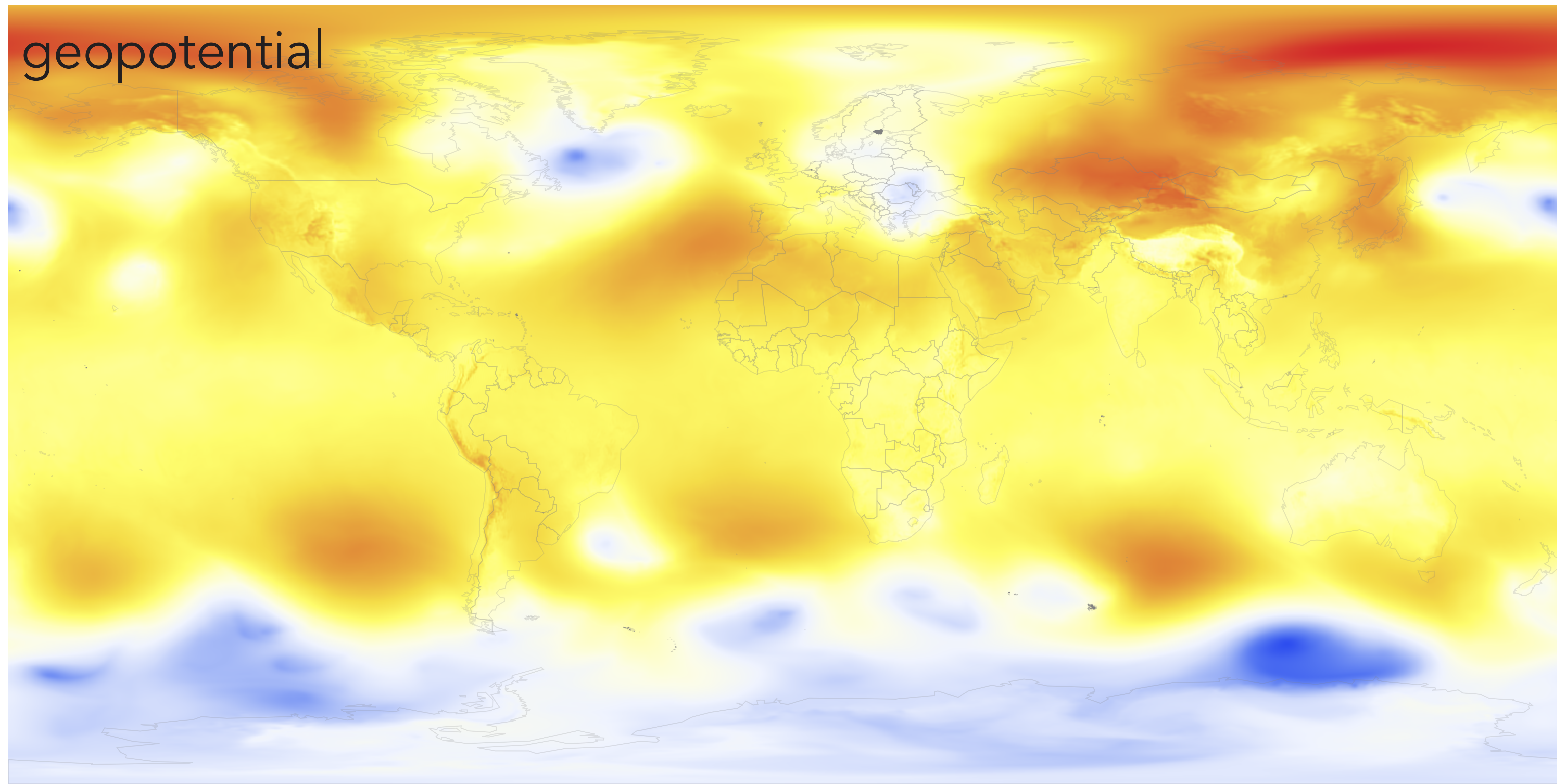
Embedding-Forecast



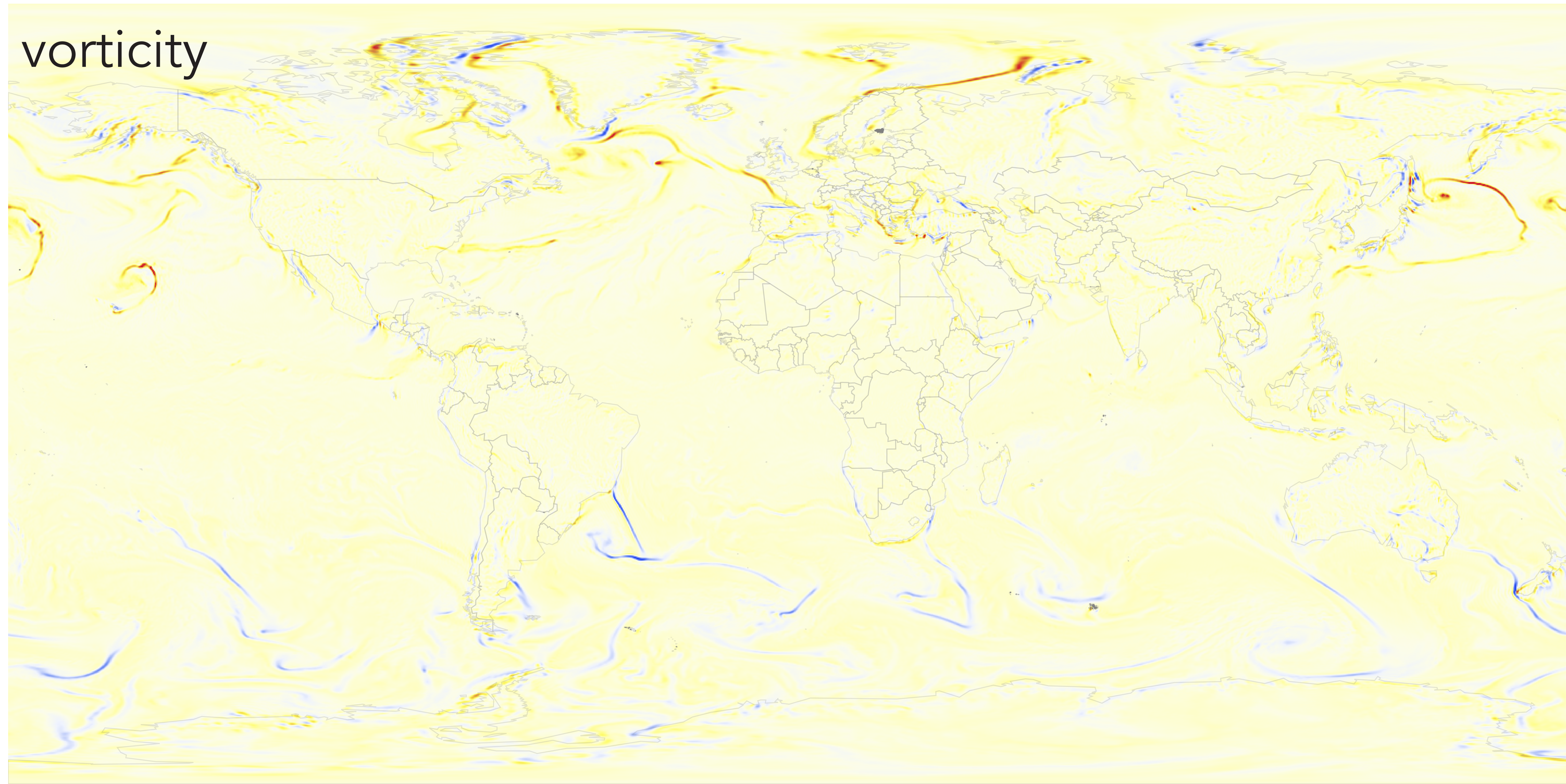
AtmoRep data



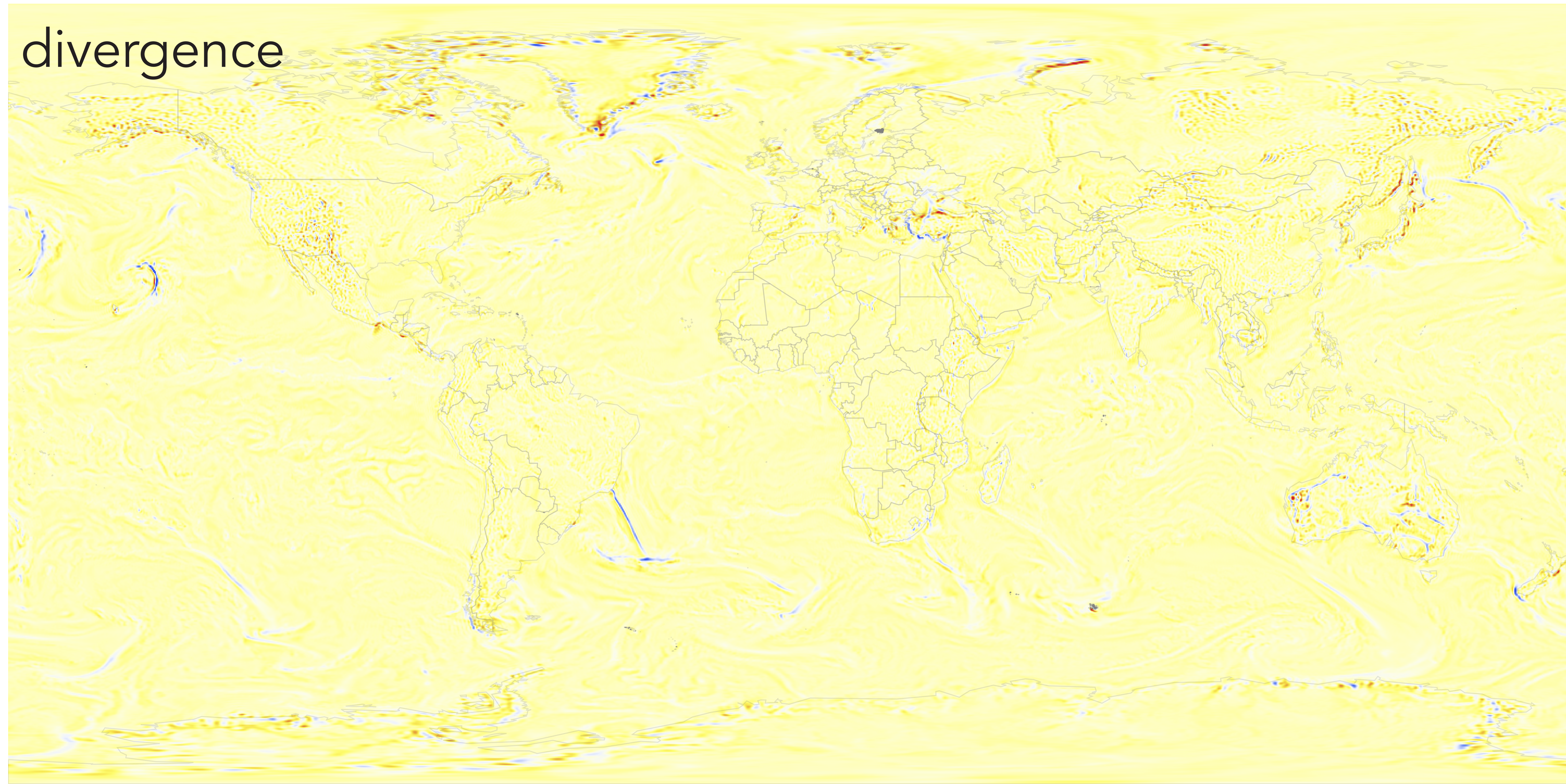
AtmoRep data



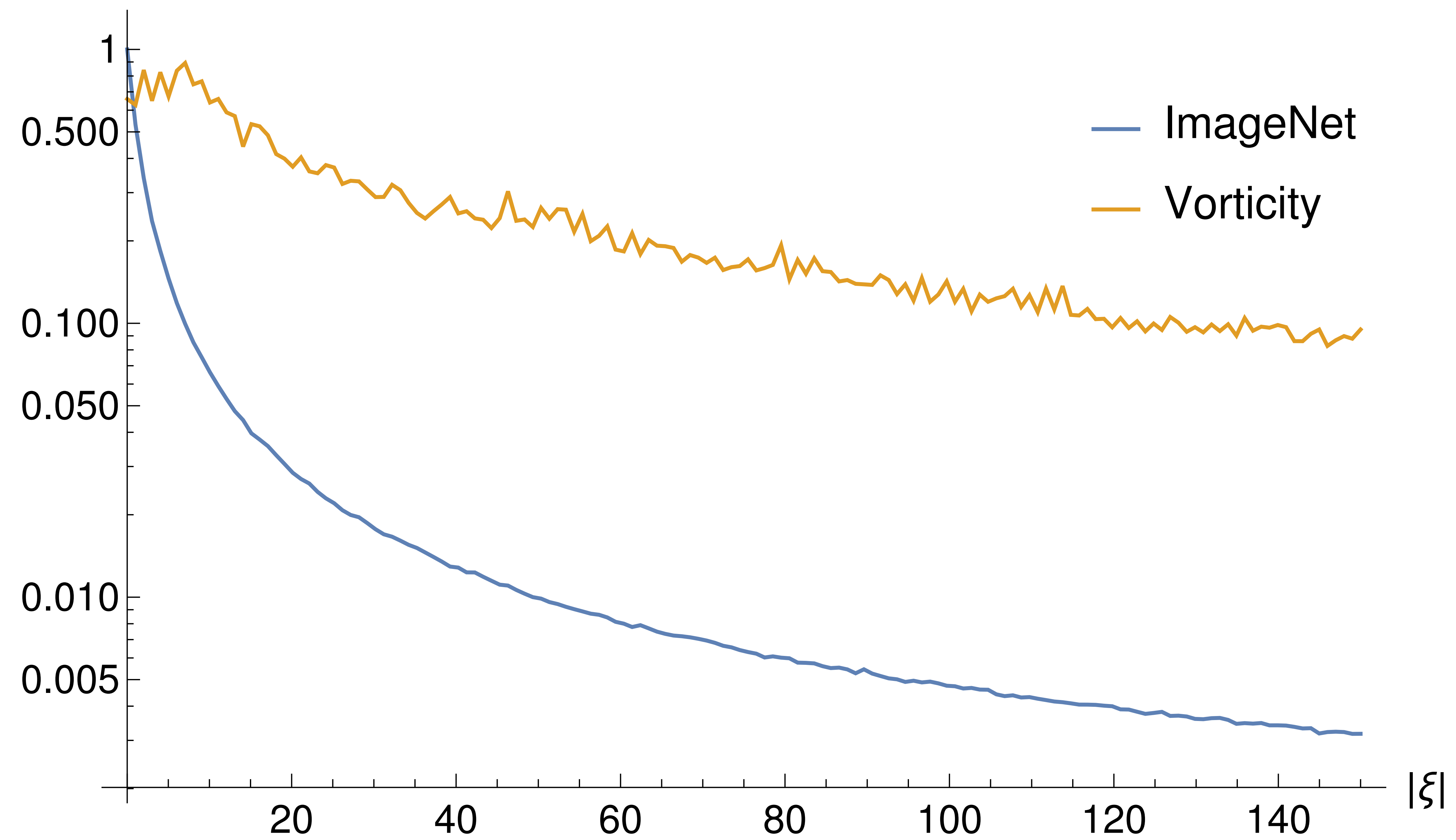
AtmoRep data



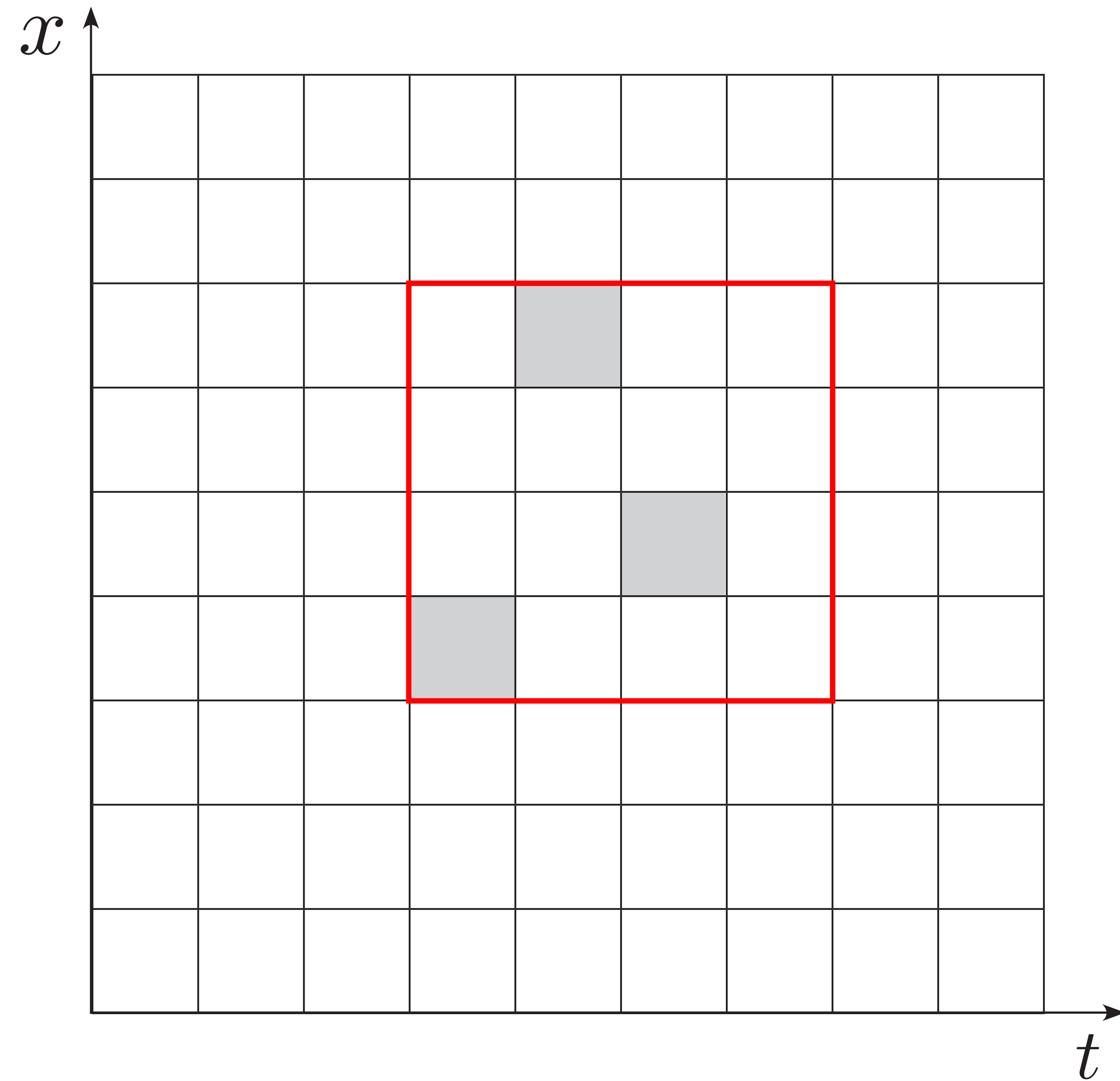
AtmoRep data



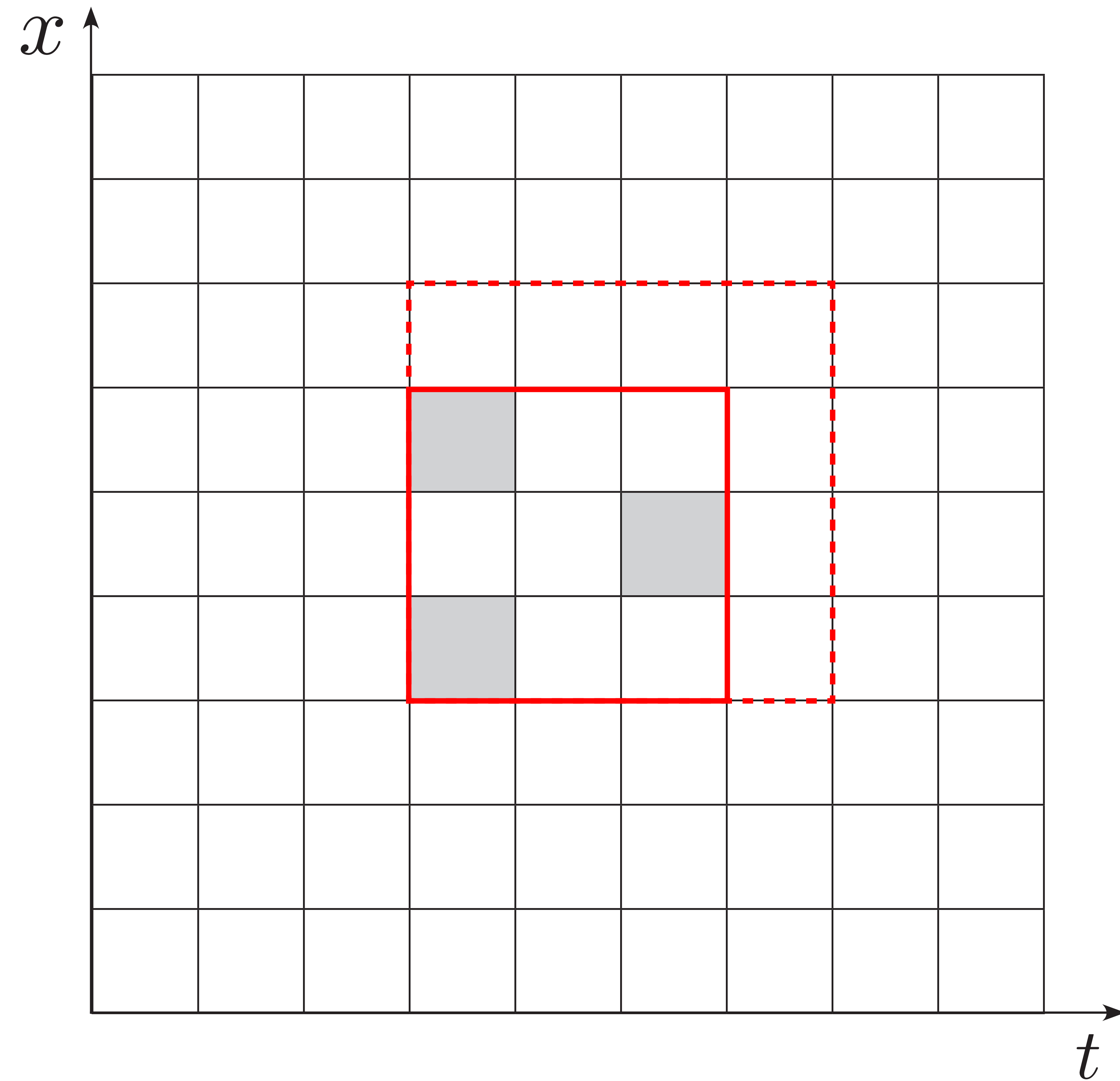
AtmoRep data



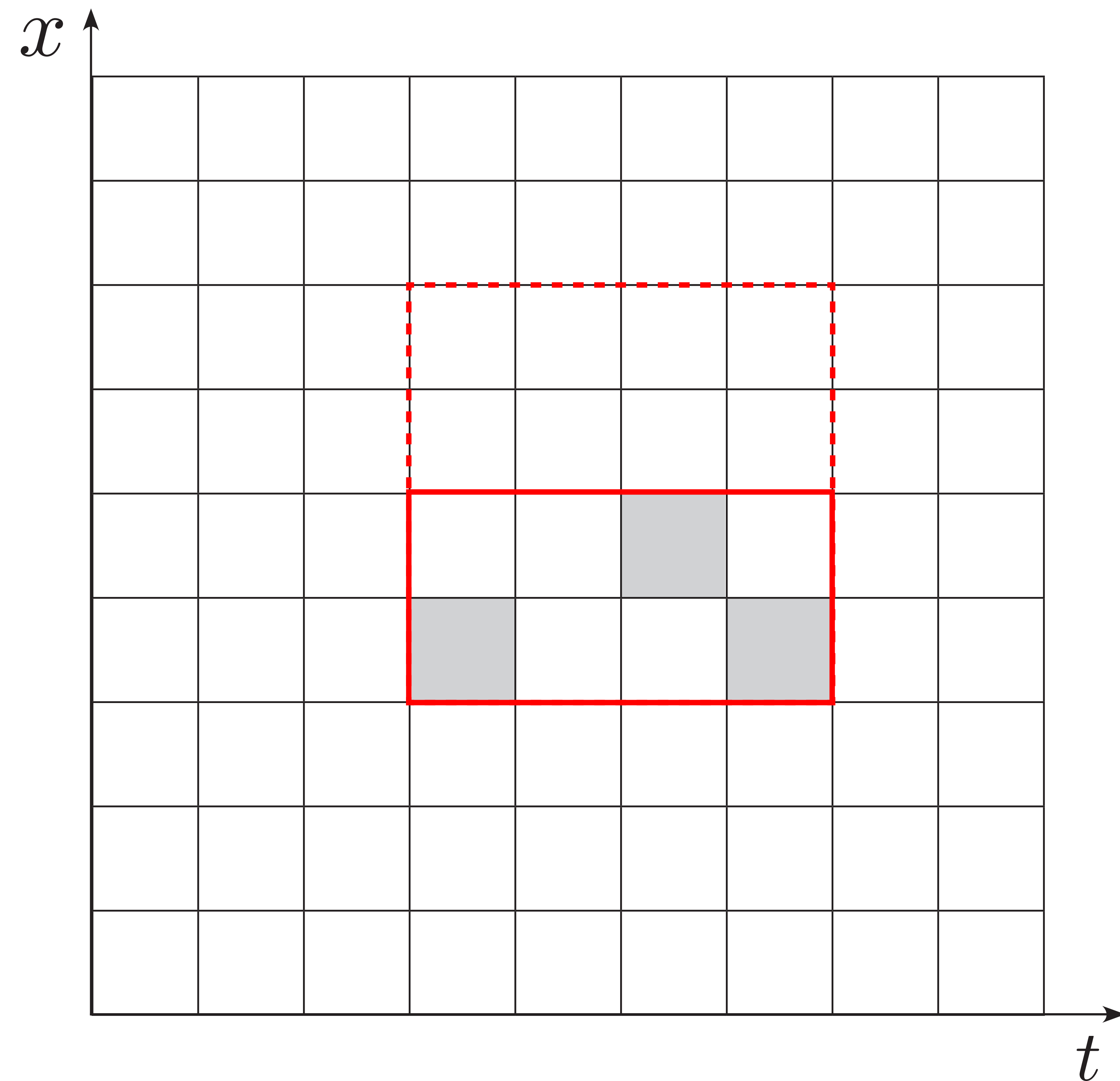
Spatio-temporal BERT



Spatio-temporal BERT

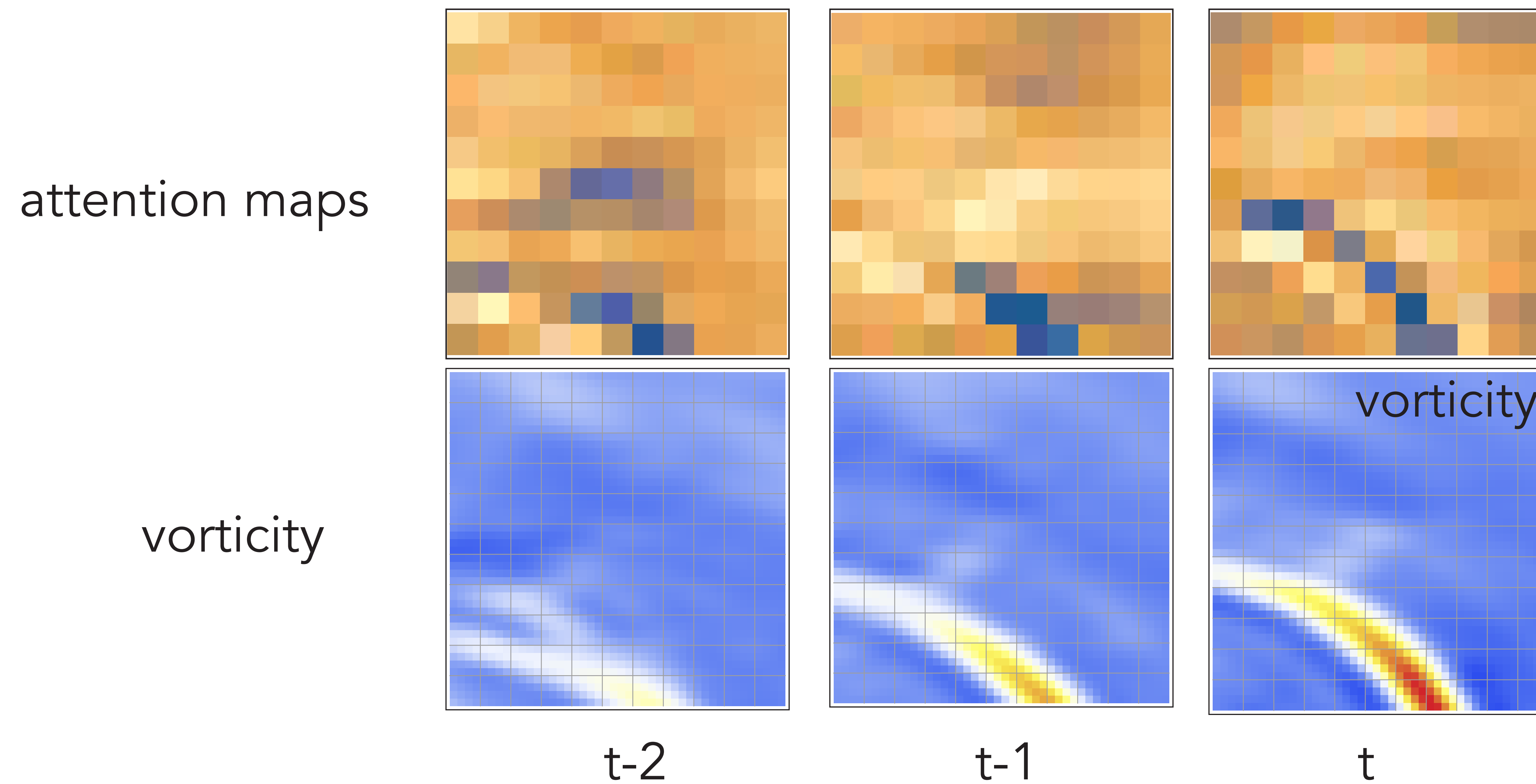


Spatio-temporal BERT

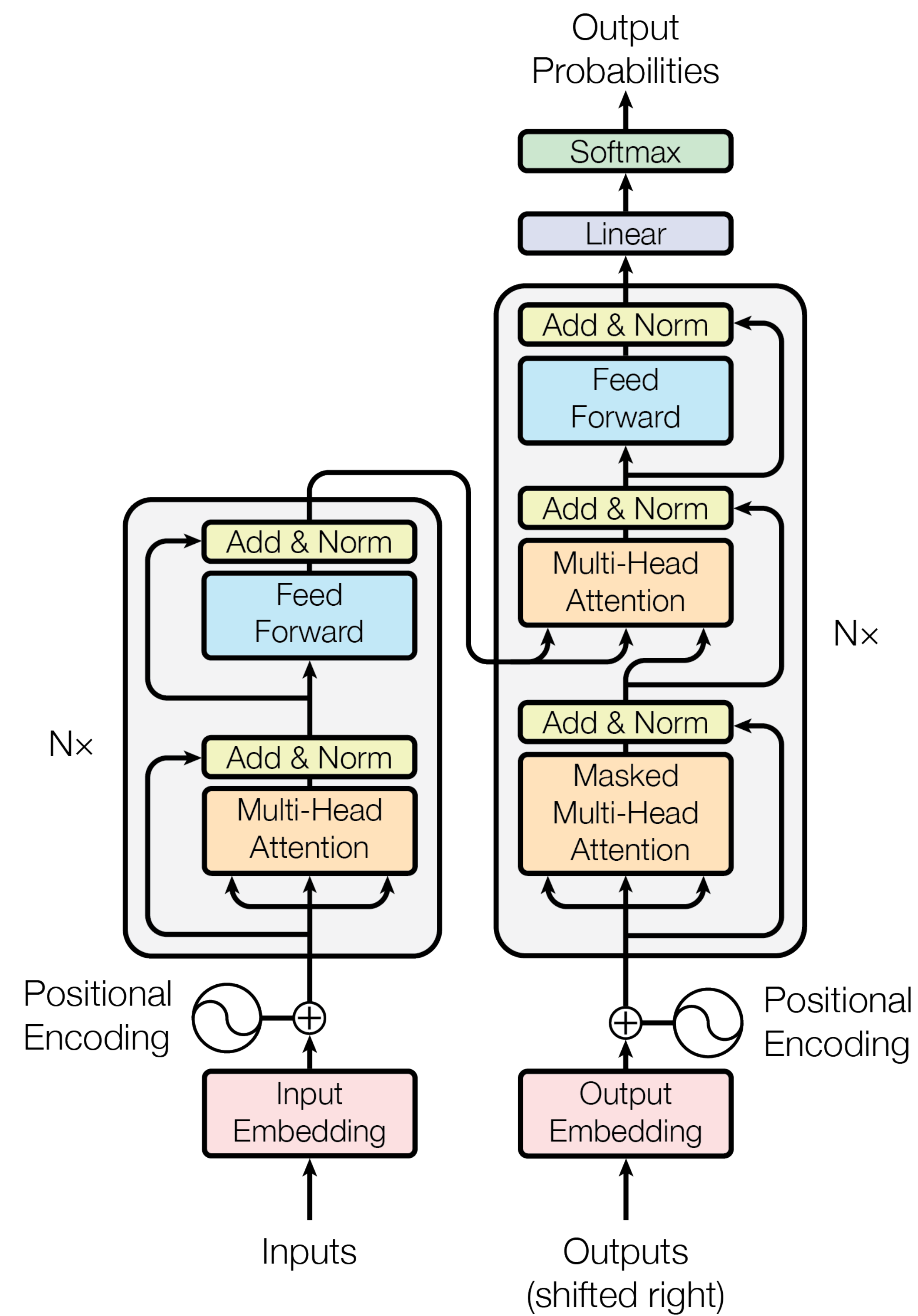


Statistical loss

- Attention maps:

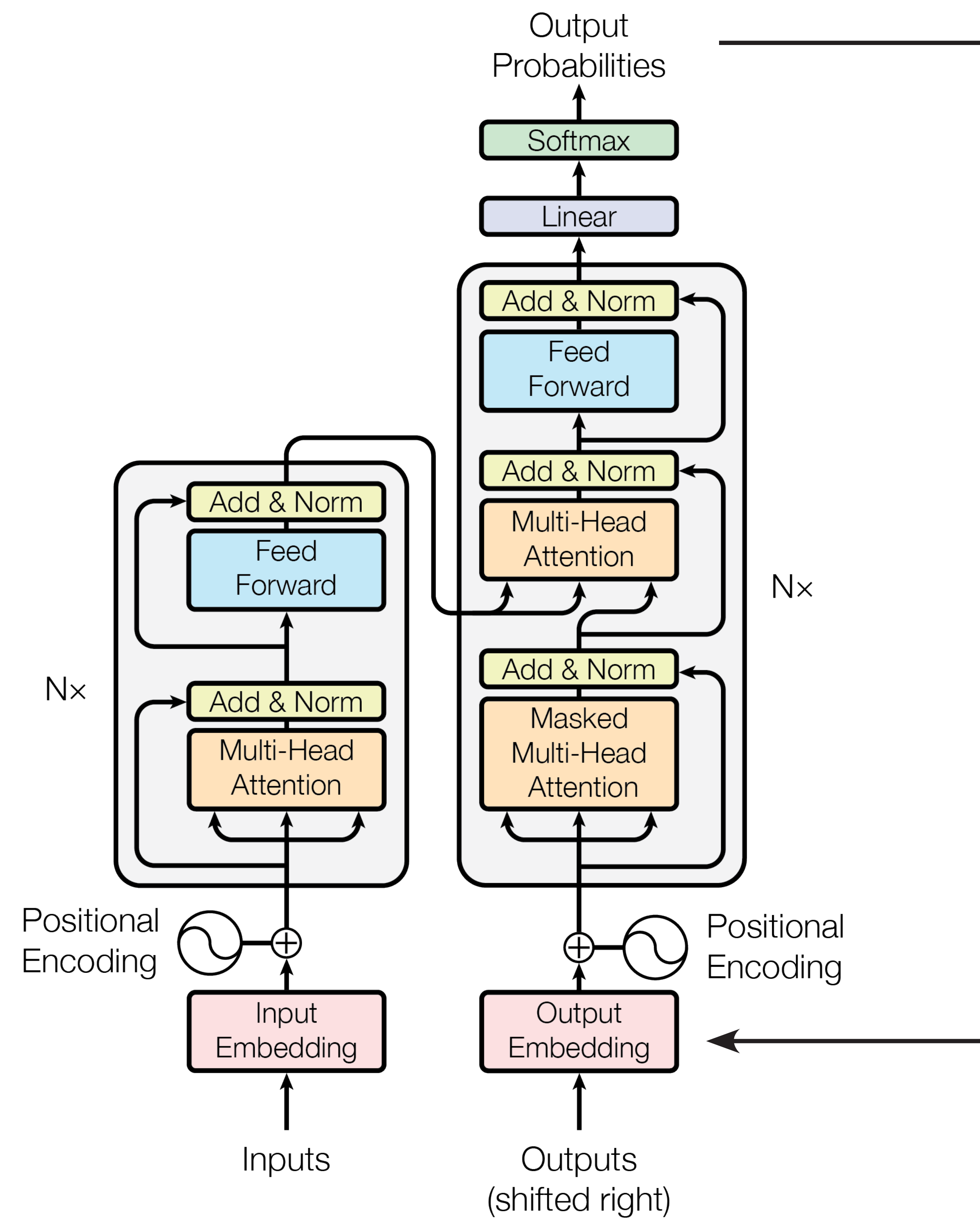


Forecasting / projections



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.

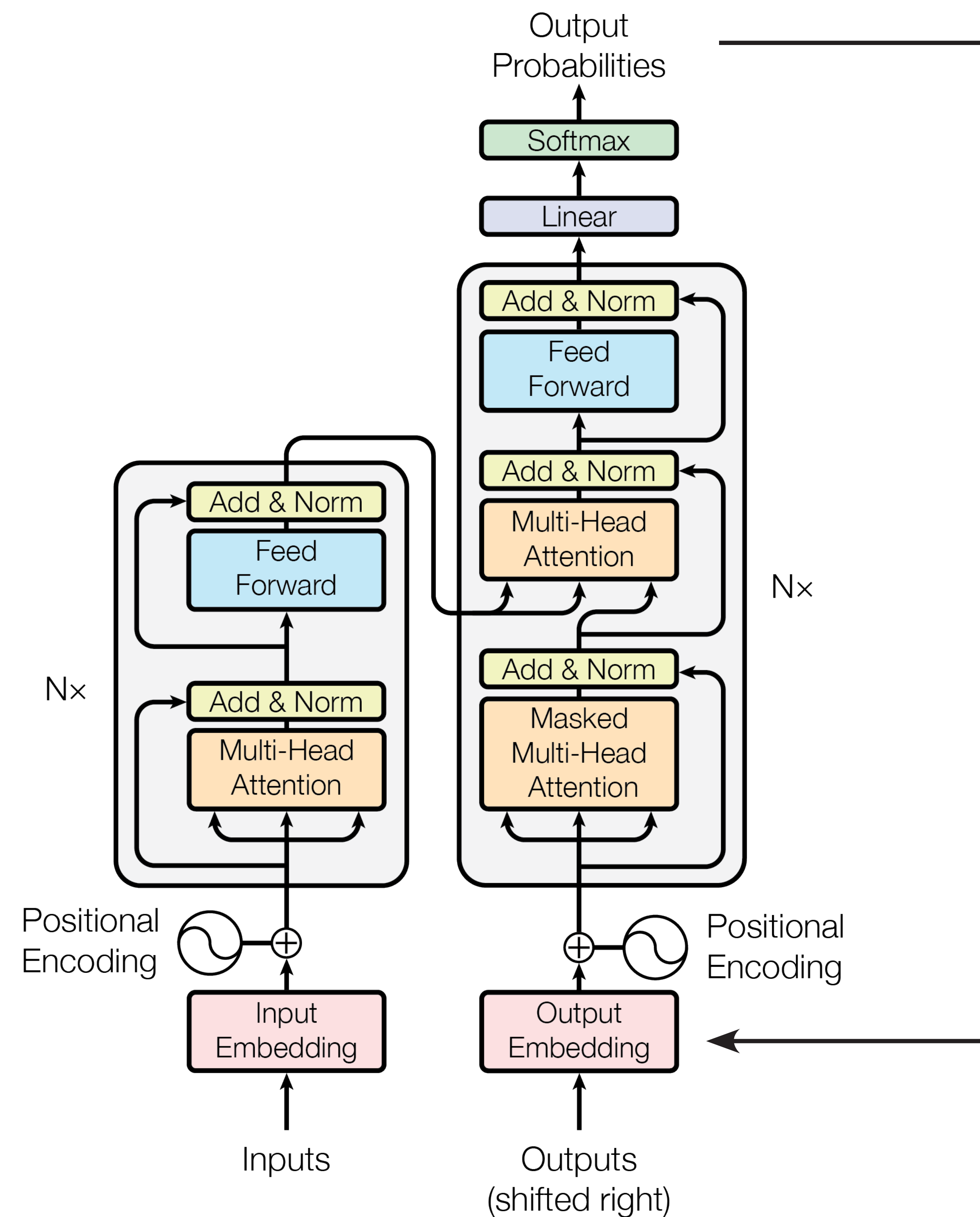
Forecasting / projections



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



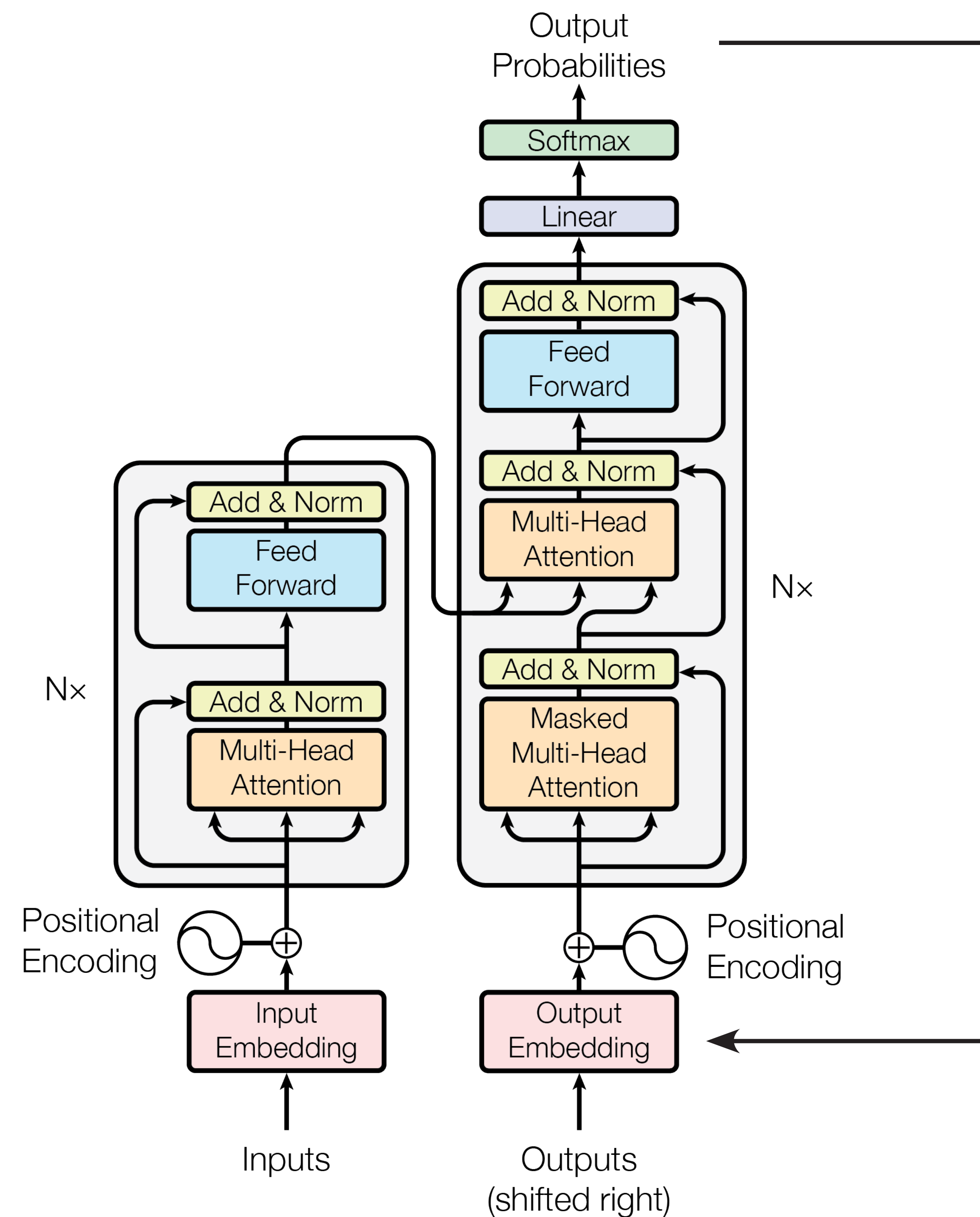
autoregressive,
generative modeling

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



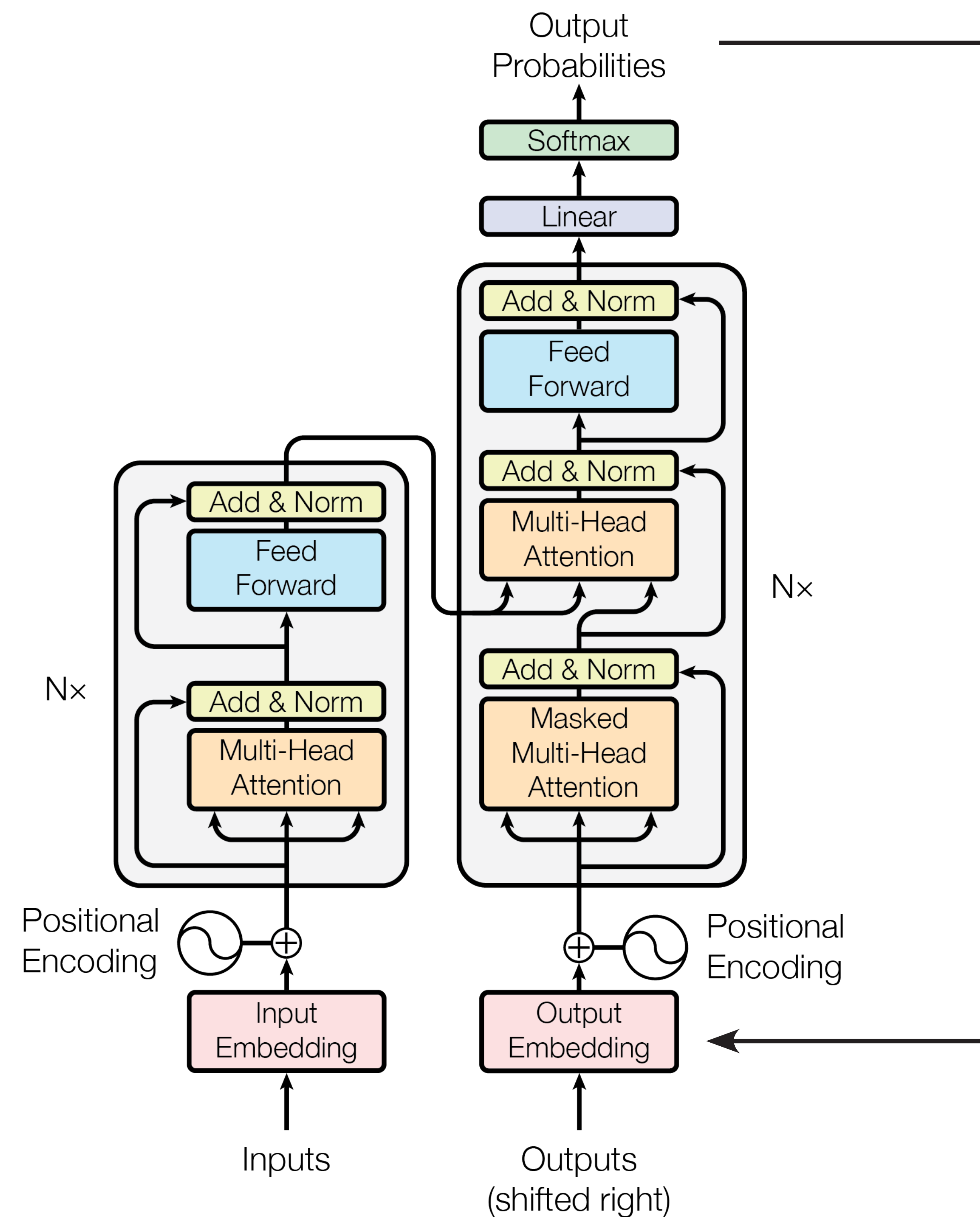
autoregressive,
generative modeling

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

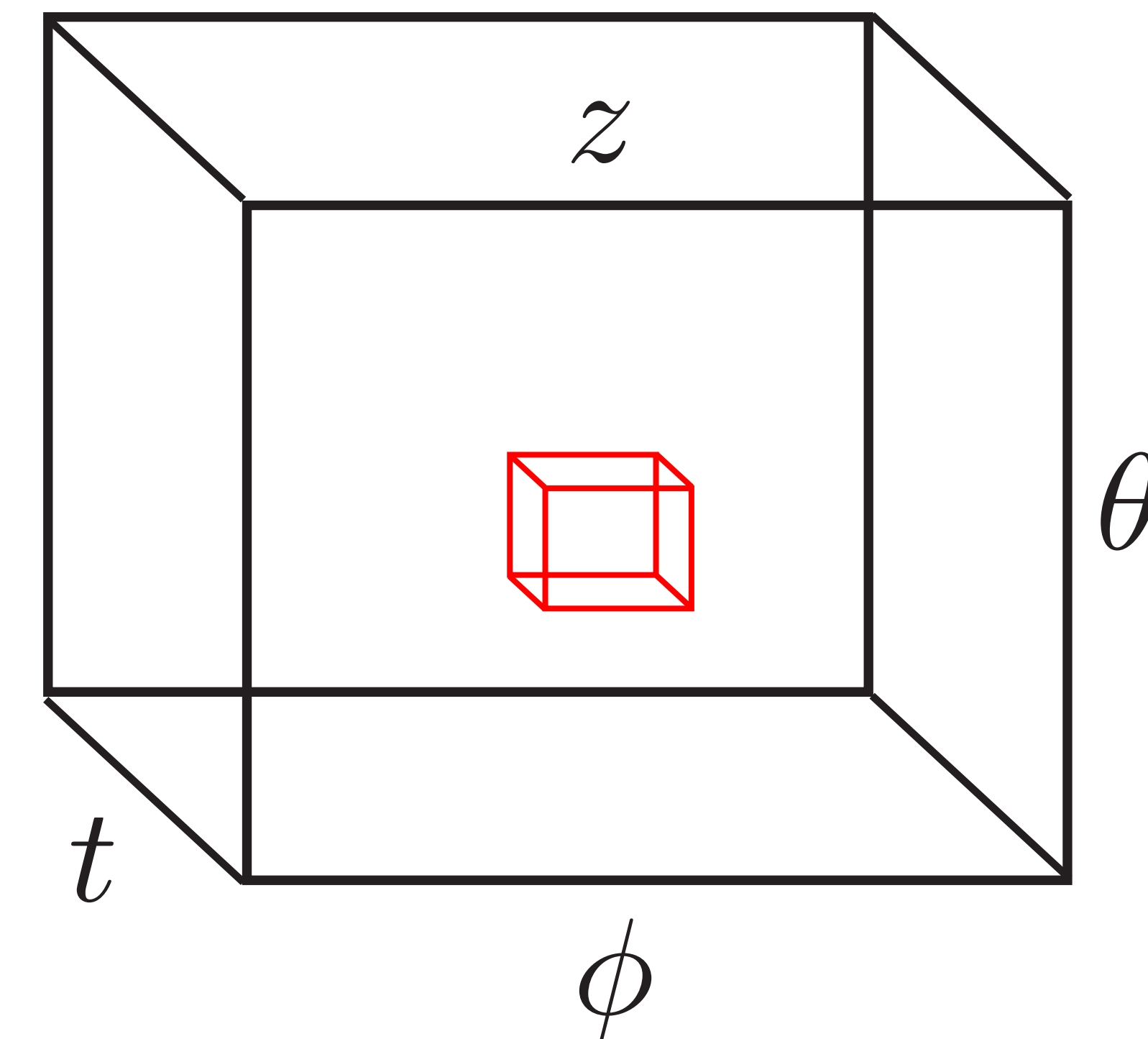
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

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

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 structure



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 structure
- Token size is field-dependent

