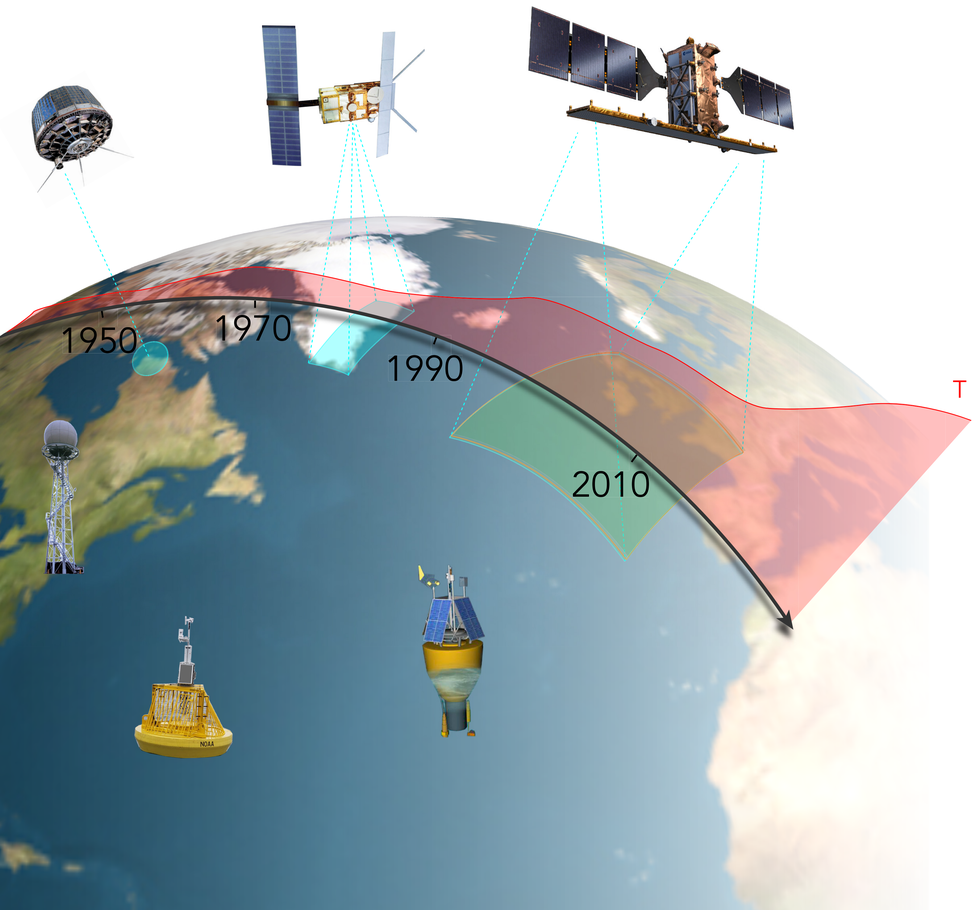
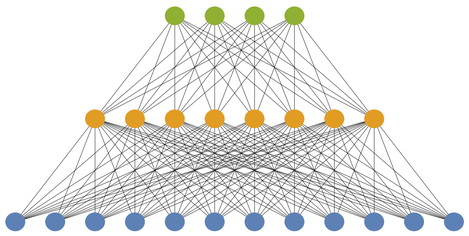


AtmoRep

Large scale representation learning of atmospheric dynamics

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



address climate change

large scale machine learning

scientific insight

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

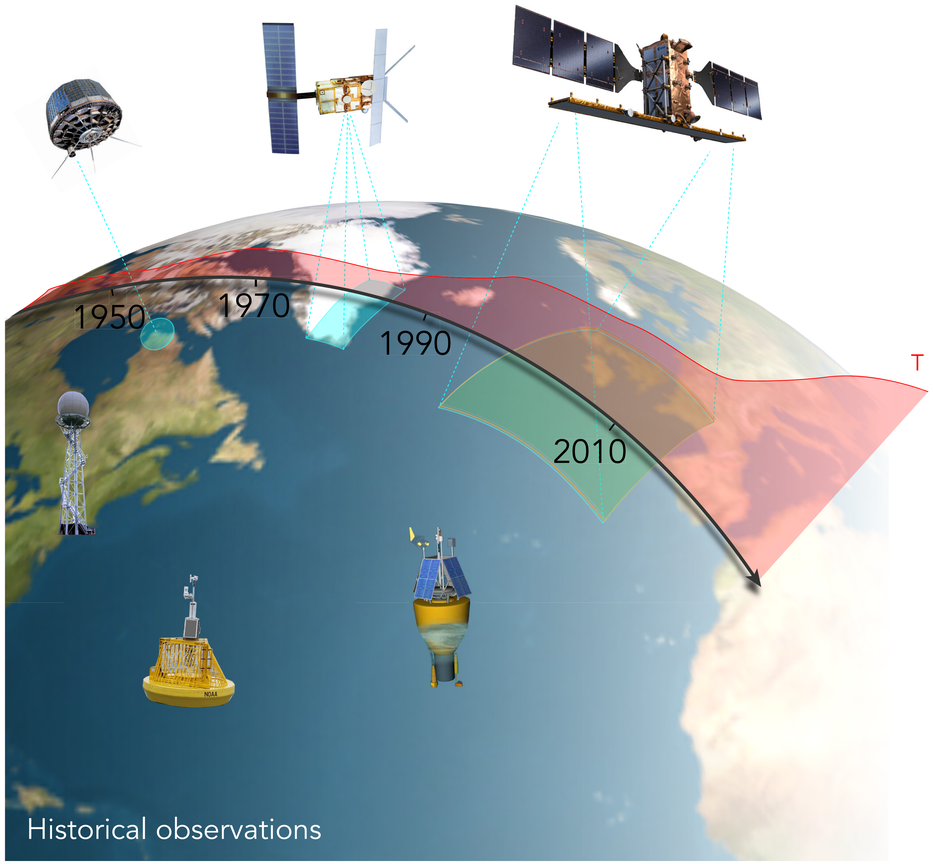
Representation learning for the Earth sciences?

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

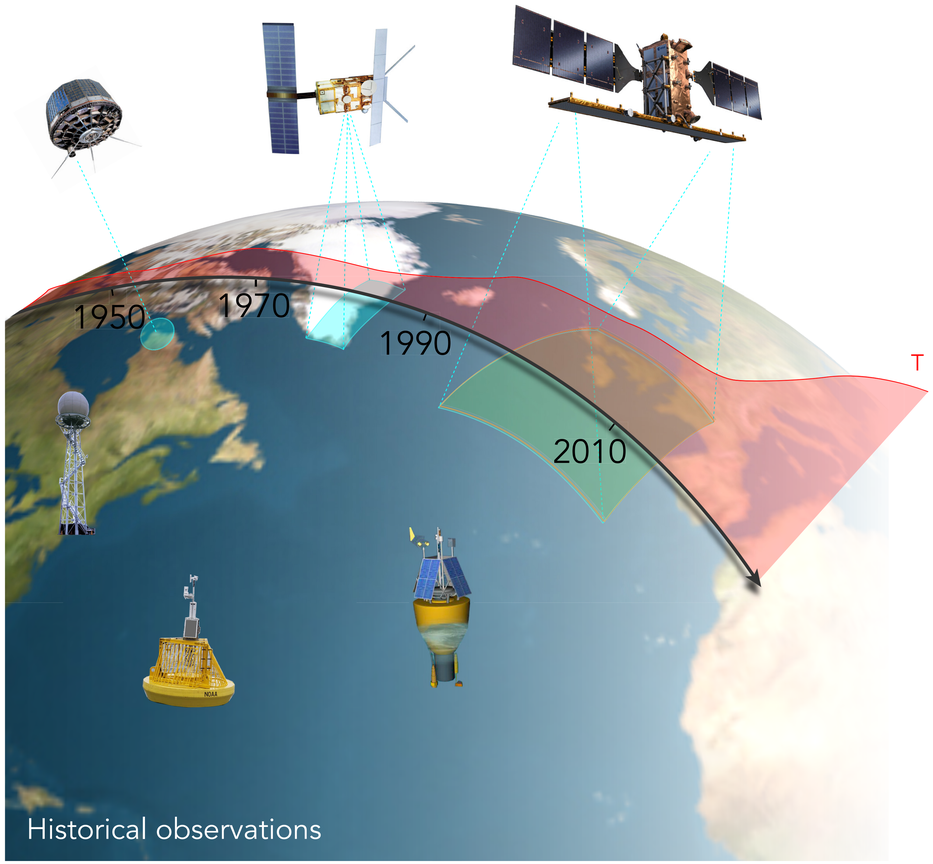
AtmoRep

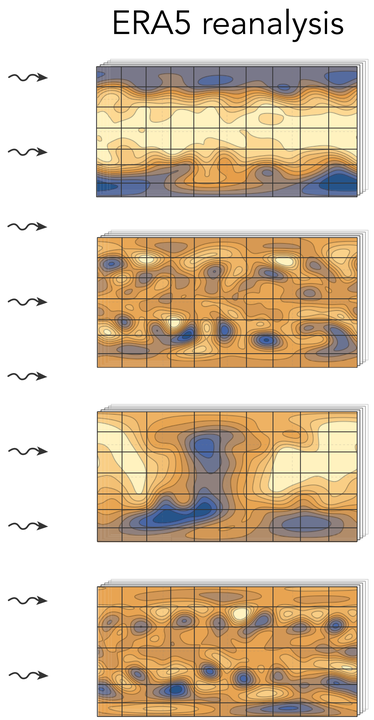
Large scale representation learning of atmospheric dynamics

AtmoRep

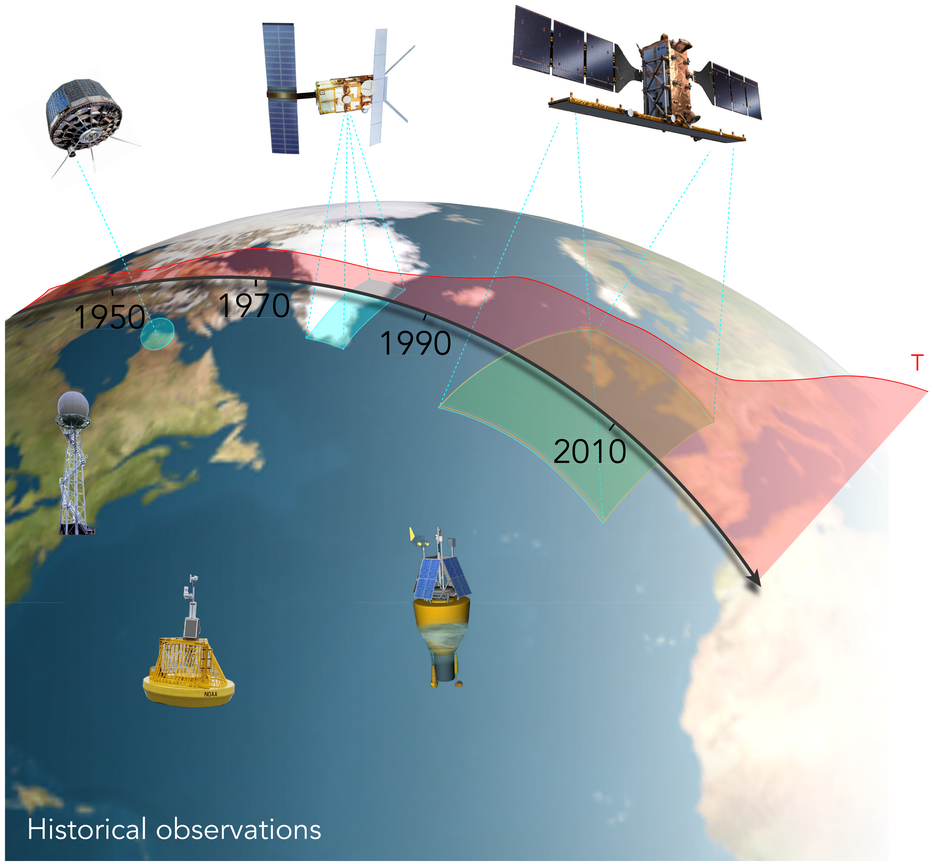


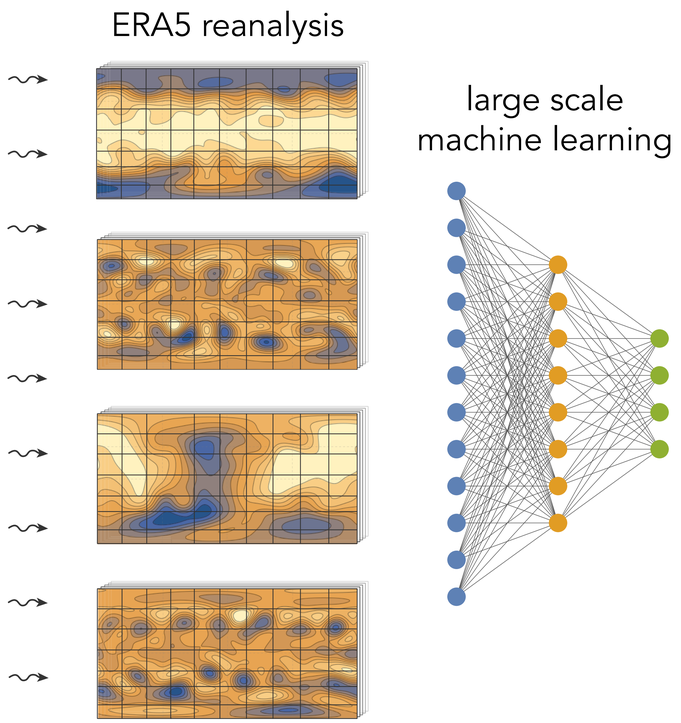
AtmoRep





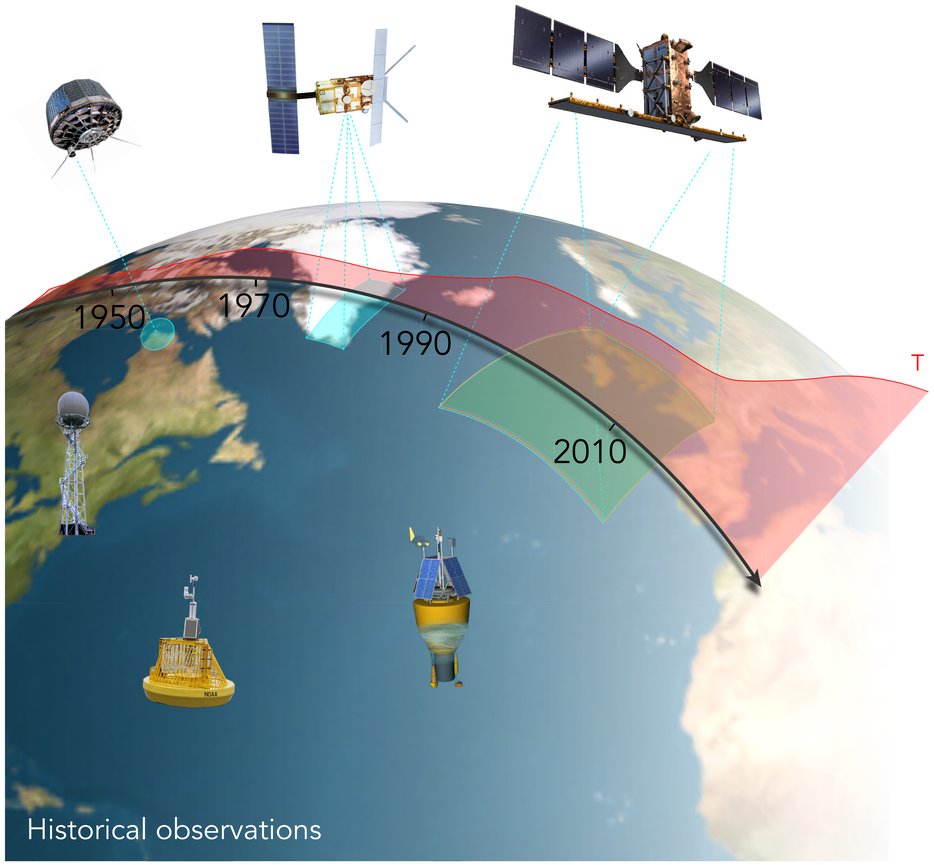
AtmoRep

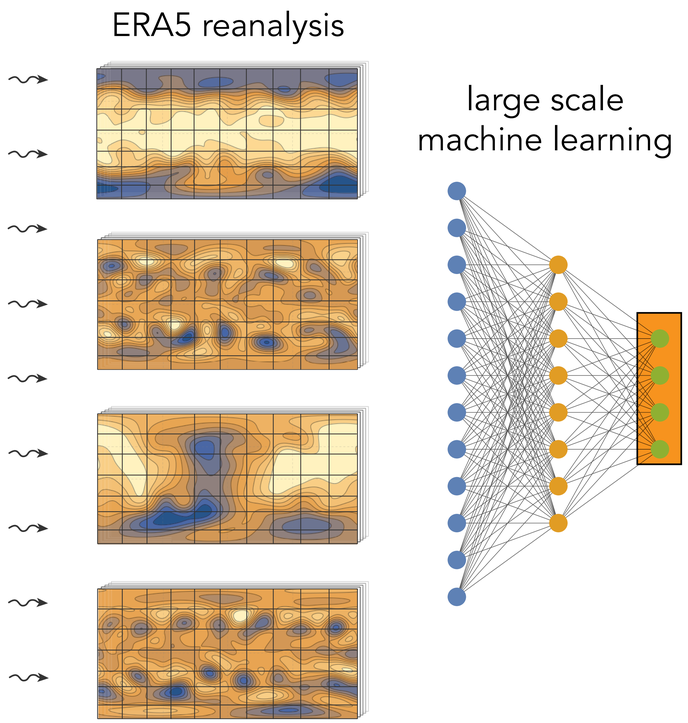




large scale
machine learning

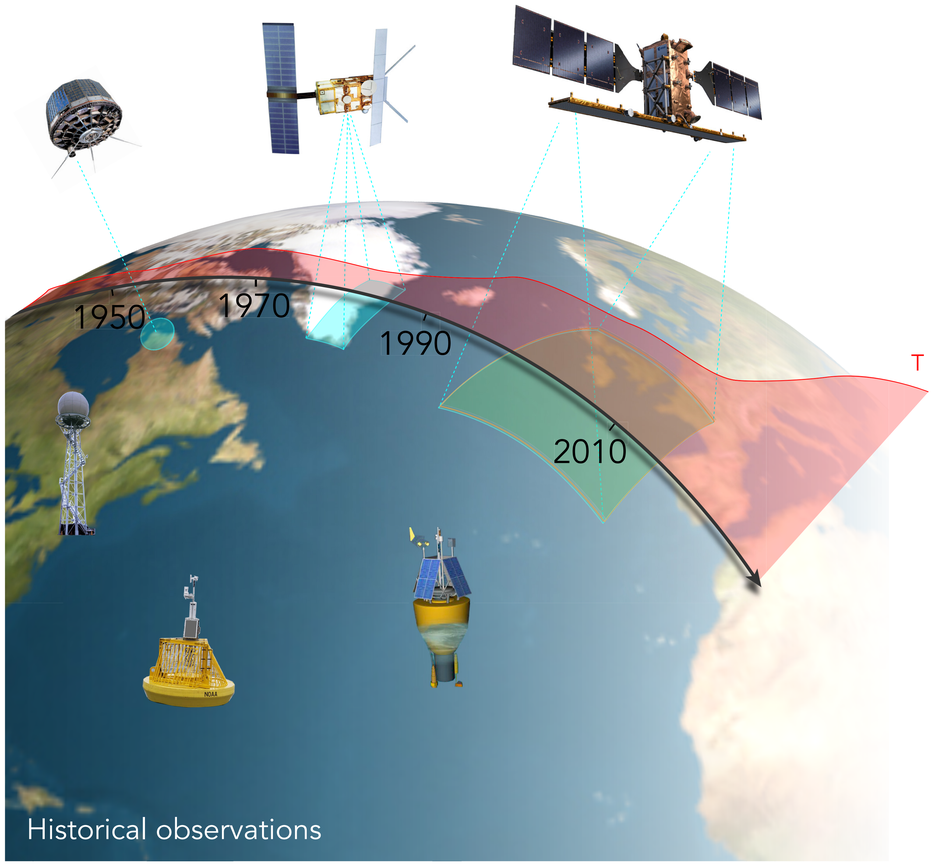
AtmoRep

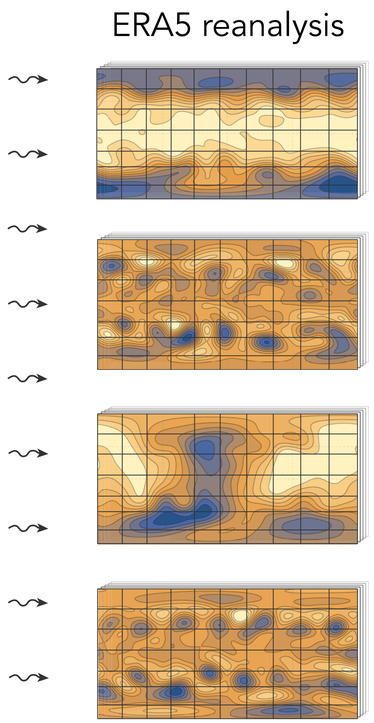


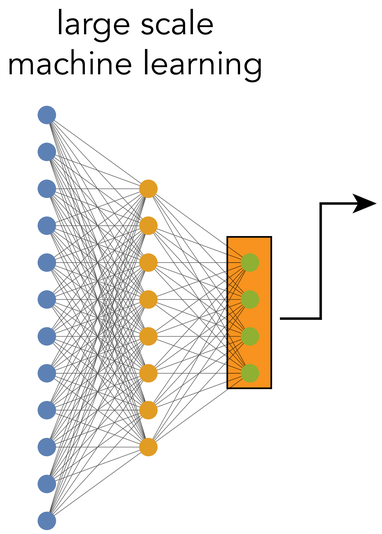


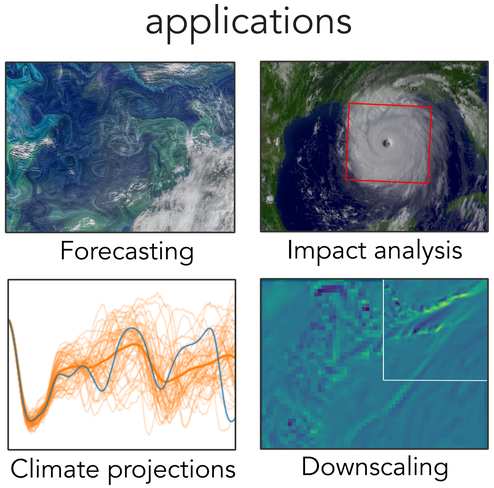
large scale
machine learning

AtmoRep

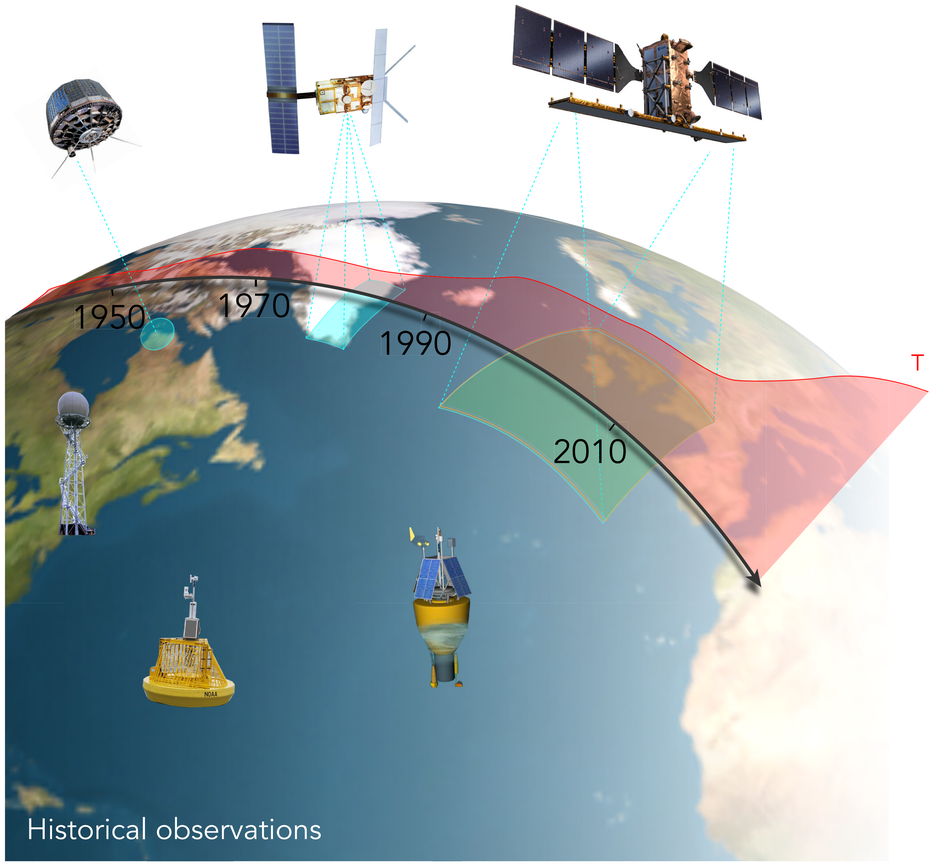


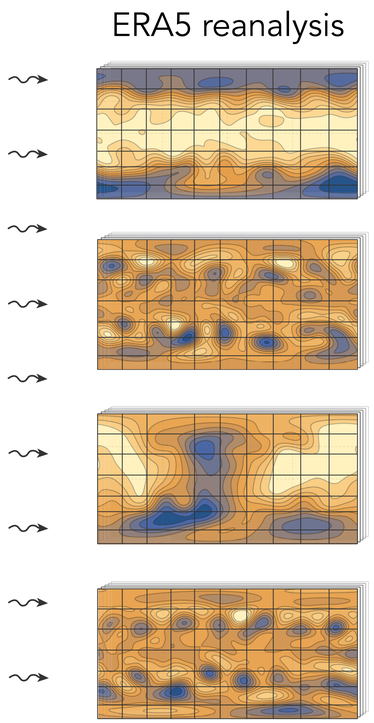


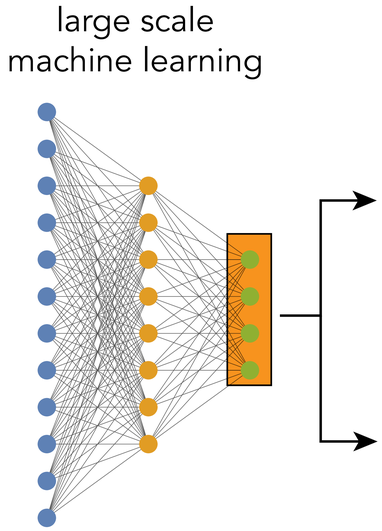


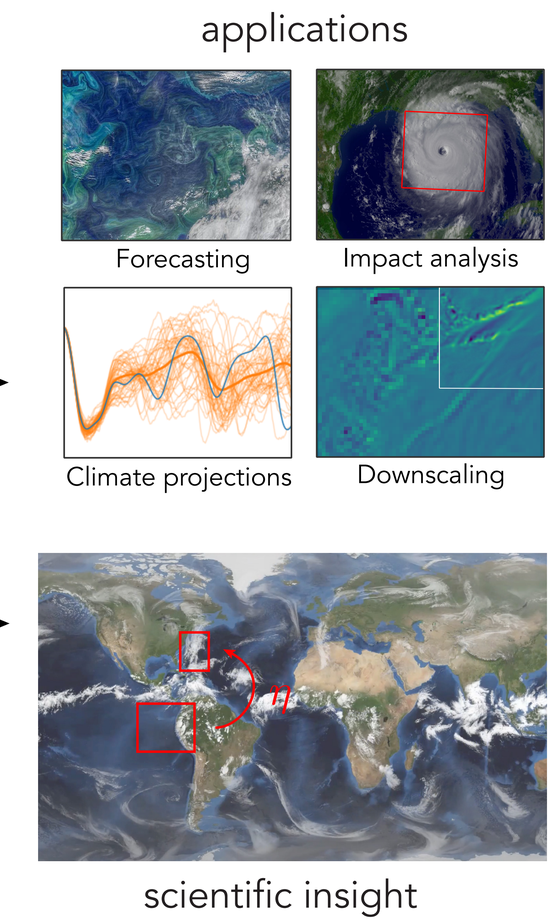


AtmoRep

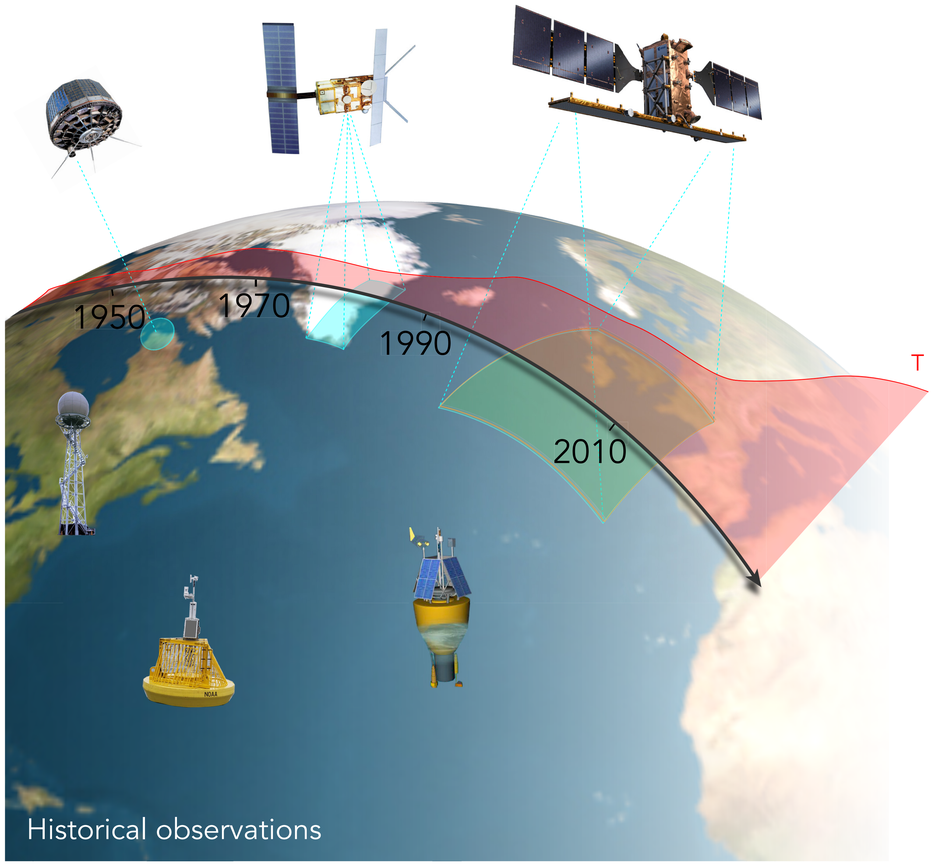


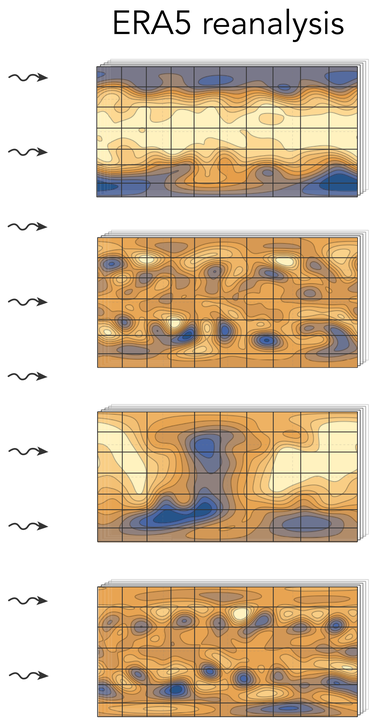


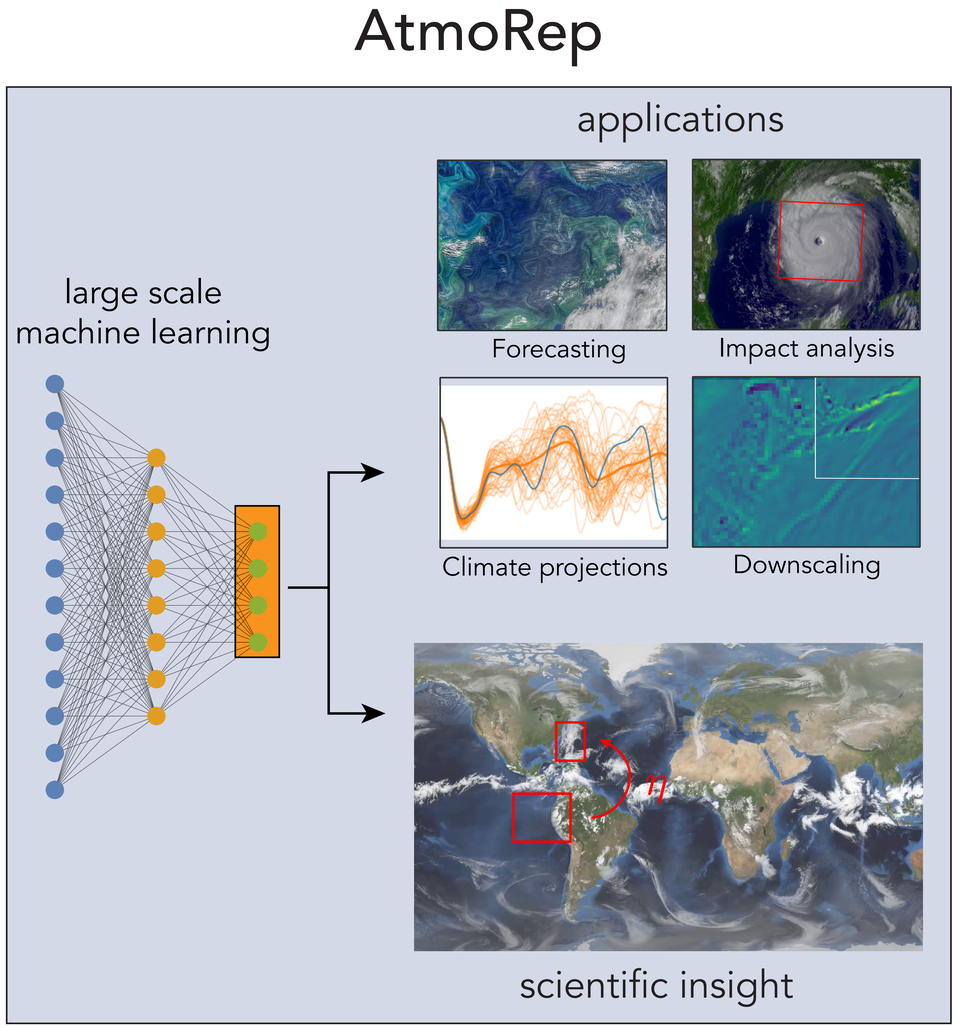




AtmoRep







AtmoRep: a theoretical formulation

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

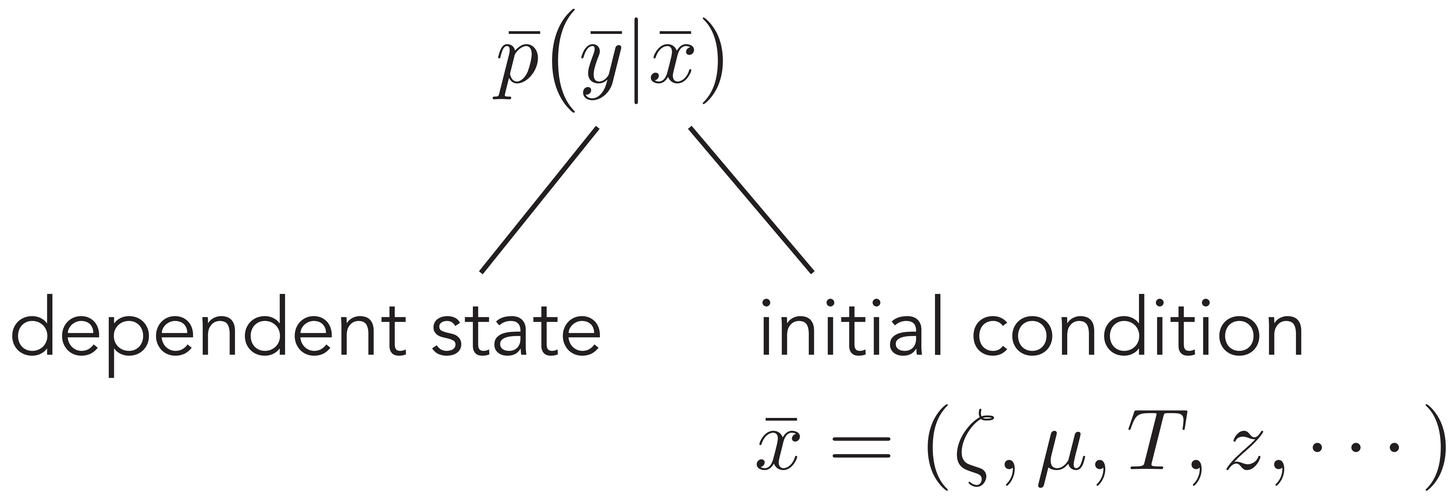
$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

initial condition

$$\bar{x} = (\zeta, \mu, T, z, \cdots)$$

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:



AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$p_\theta\left(y|x,\alpha\right)$$

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\big(\bar{y}|\bar{x}\big)$$

- Neural network model:

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

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$p_\theta\left(y|x,\alpha\right)$$

approx. initial condition
$x = (\zeta, \mu, T, z, \dots)$

auxiliary information
(e.g. global time)

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

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

Standard formulation for generative models, e.g., large language models, Dall-E, diffusion models

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

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

- forecasting
- downscaling
- model correction
- ...

Standard formulation for generative models, e.g., large language models, Dall-E, diffusion models

AtmoRep: a theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - › General model allows to include all effects in the data

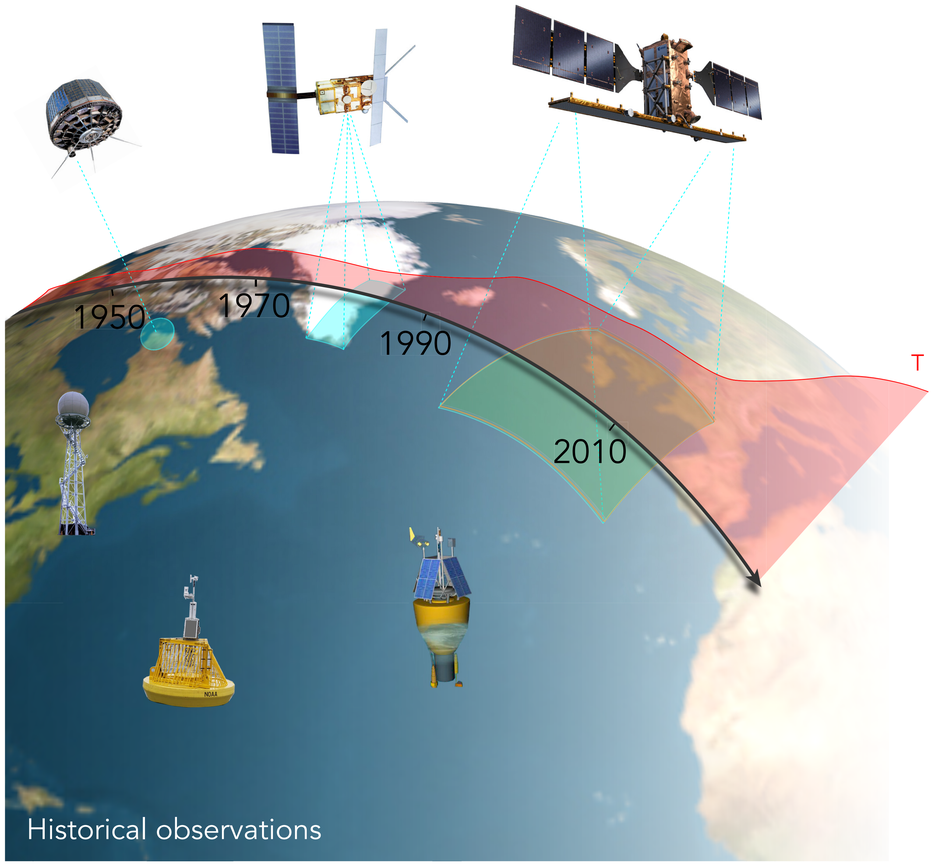
AtmoRep: a theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - › General model allows to include all effects in the data
- Intrinsically statistical/probabilistic formulation
  - › Fits the statistical/chaotic nature of the atmosphere

AtmoRep: a theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - › General model allows to include all effects in the data
- Intrinsically statistical/probabilistic formulation
  - › Fits the statistical/chaotic nature of the atmosphere
- Neural network model as factorization of $p_\theta\left(y|x,\alpha\right)$
  - › Loss (expectation maximization, ELBO, ...) has clear connection to physical model

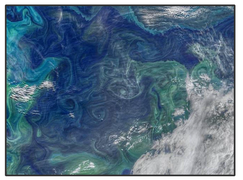
AtmoRep

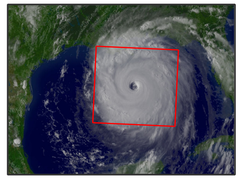


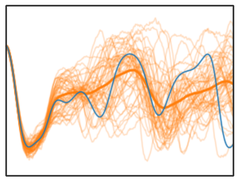
ERA5 reanalysis

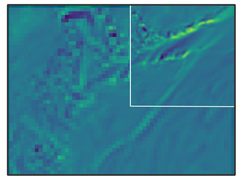
large scale machine learning

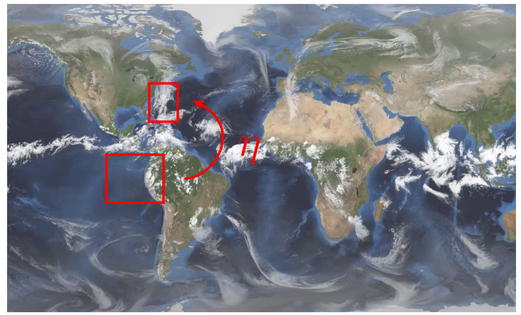
applications


Forecasting


Impact analysis

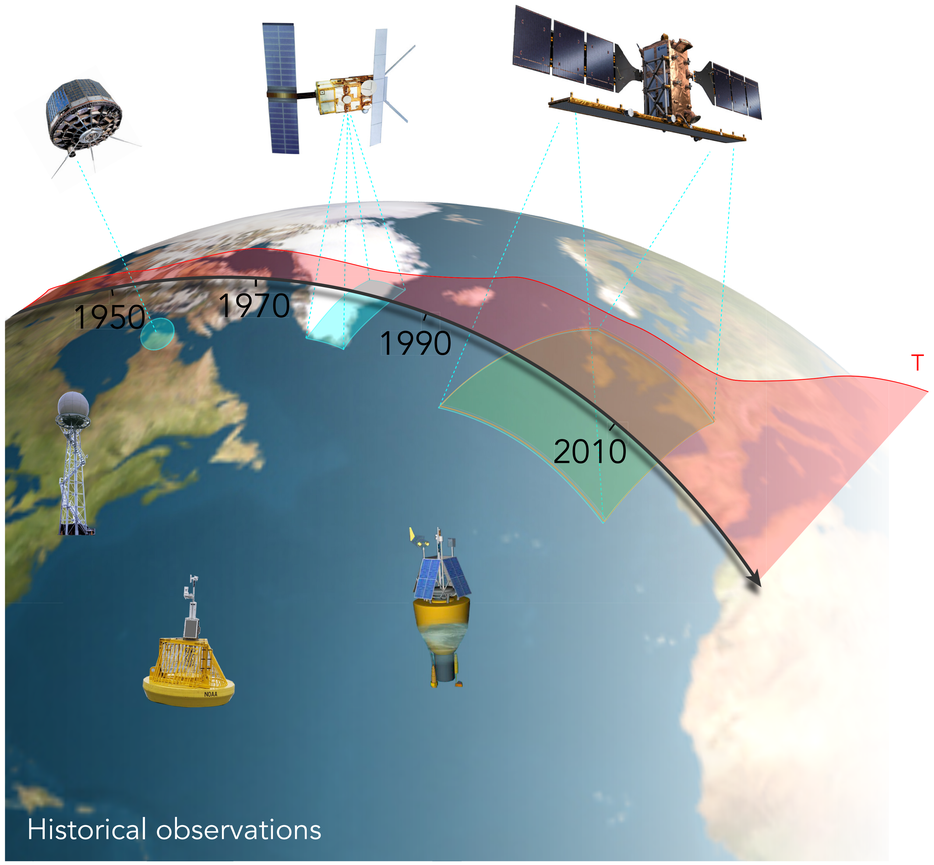

Climate projections

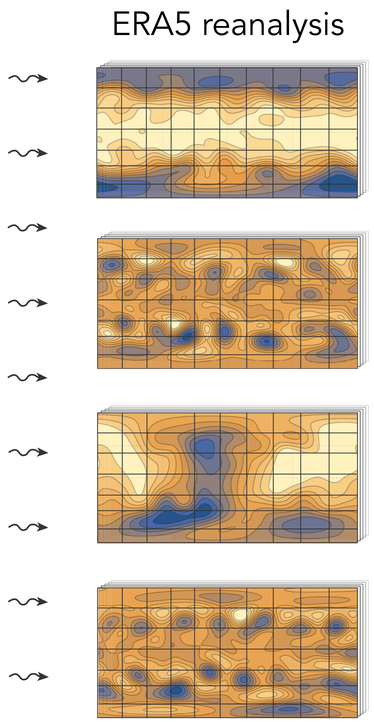

Downscaling



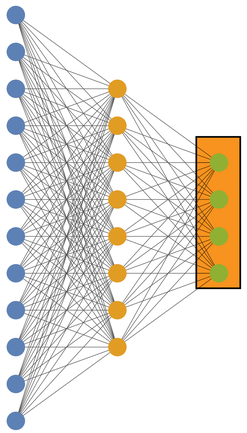
scientific insight

AtmoRep

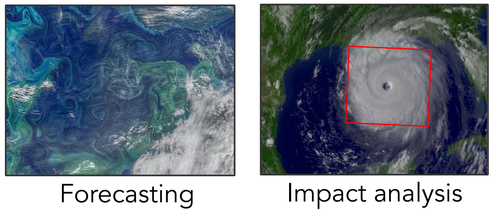


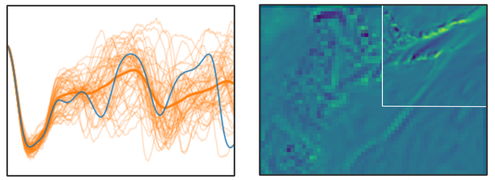


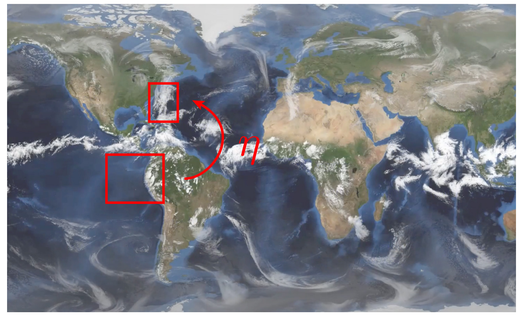
$p_\theta\left(y|x,\alpha\right)$



$p_\theta\left(y|x,\alpha,\tau\right)$





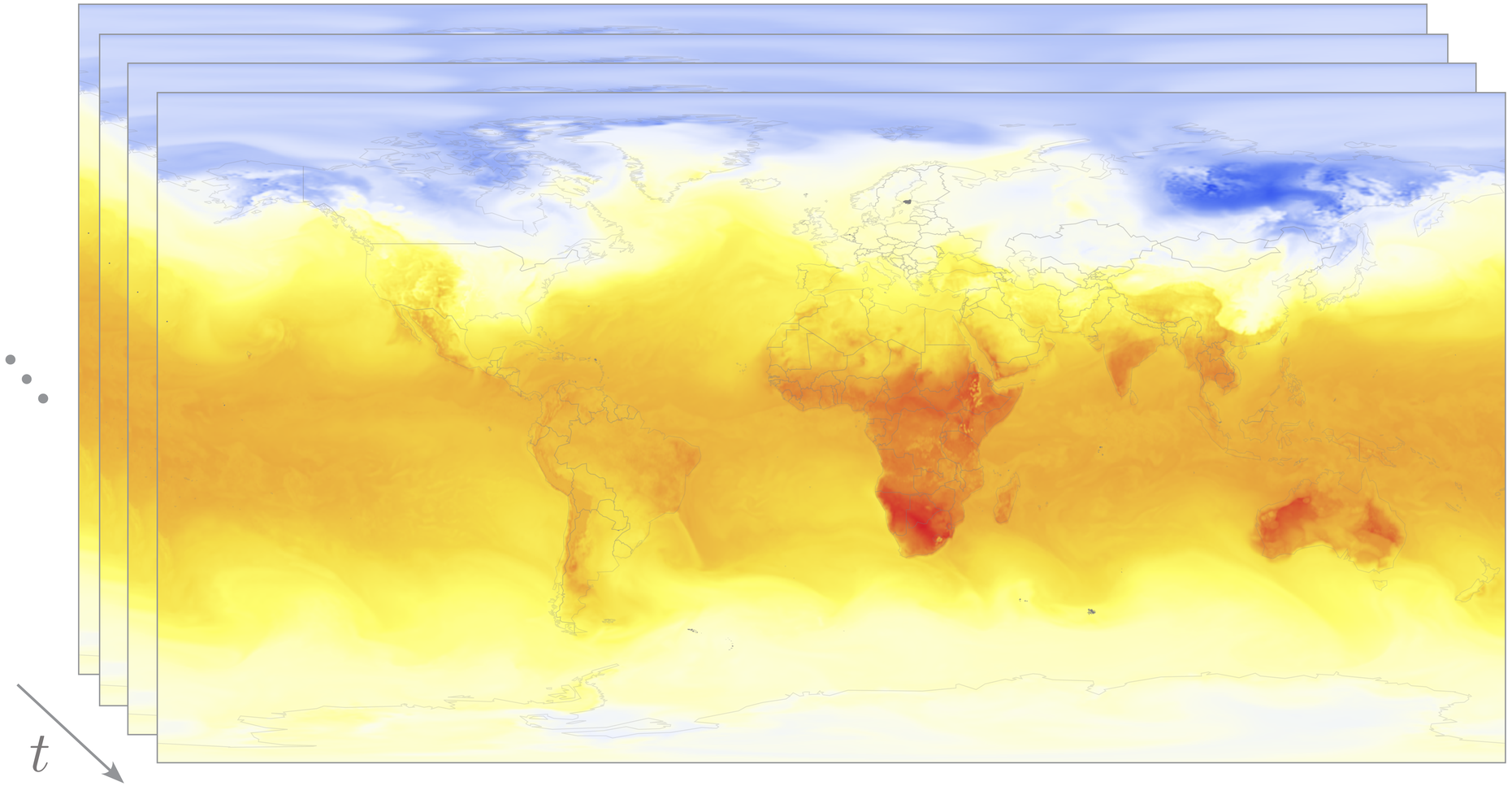


$p_\theta\left(y|x,\alpha\right) \approx \bar{p}\left(\bar{y}|\bar{x}\right)$

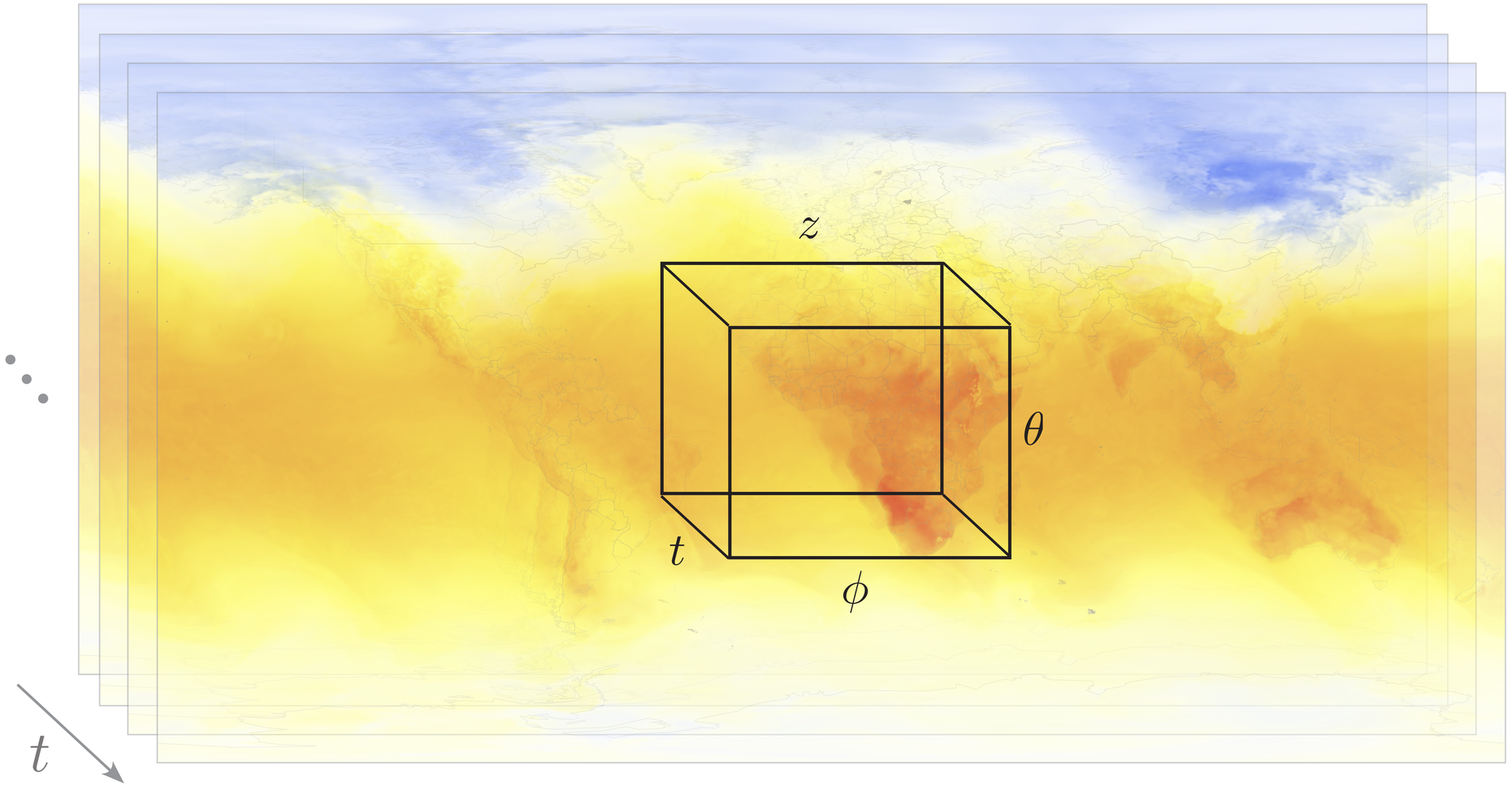
AtmoRep network architecture

- 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

$$p_\theta\left(y|x,\alpha\right)$$

approx. initial condition
$x = (\zeta, \mu, T, z, \dots)$

auxiliary information
(e.g. global time)

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

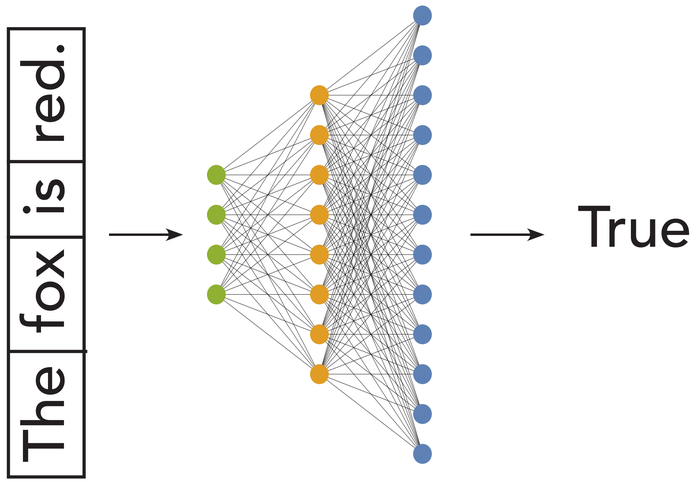
- Transformer as network architecture
  - Scales well to very large data-sets
  - Generative model (with decoder)
  - Attention maps provide (physical) interpretability

What is a token?

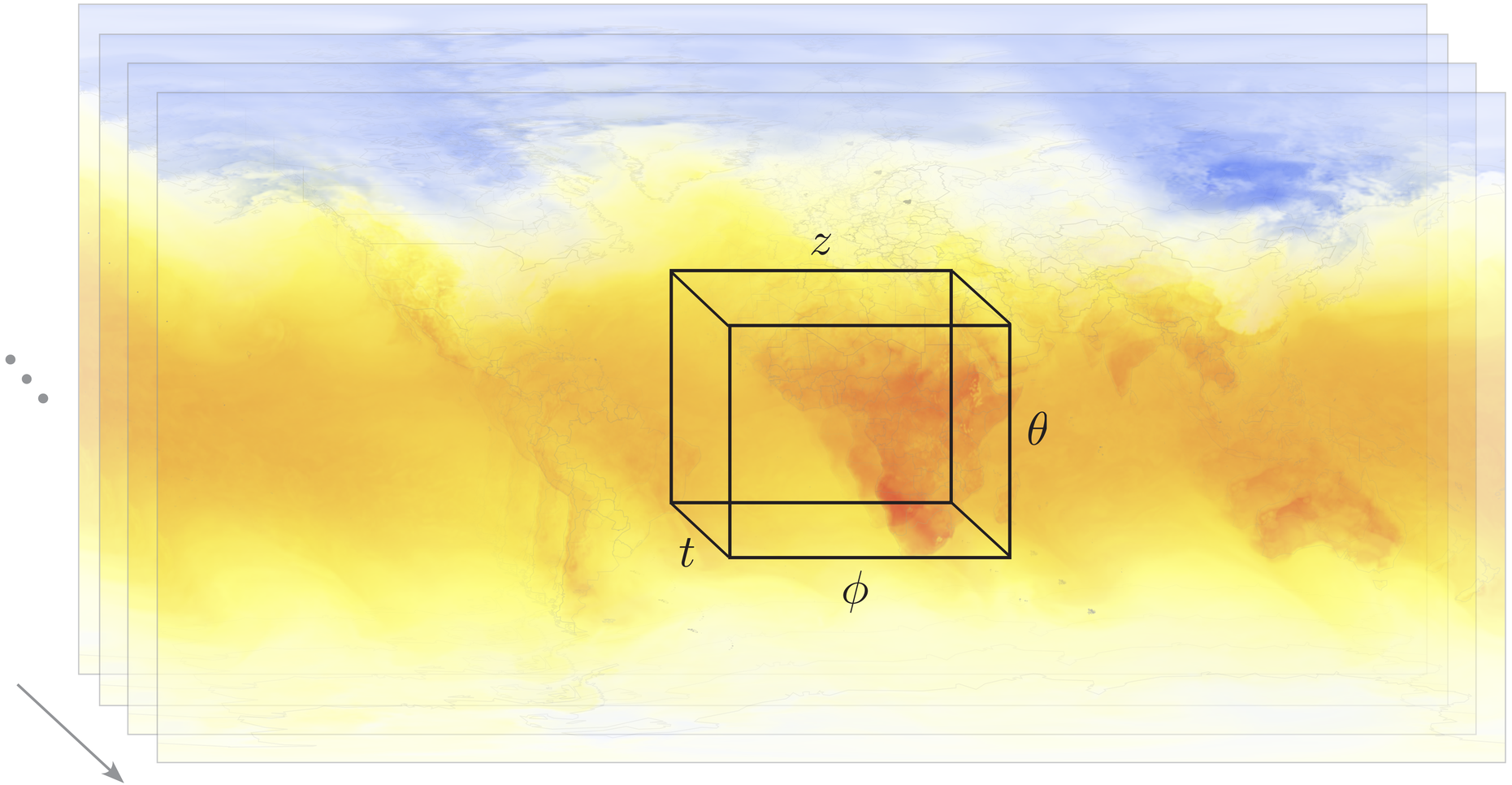
- Key property of transformers: sequence of inputs is processed simultaneously
  - Language: input is sentence with words being tokens
  - Images: small image patches

What is a token?

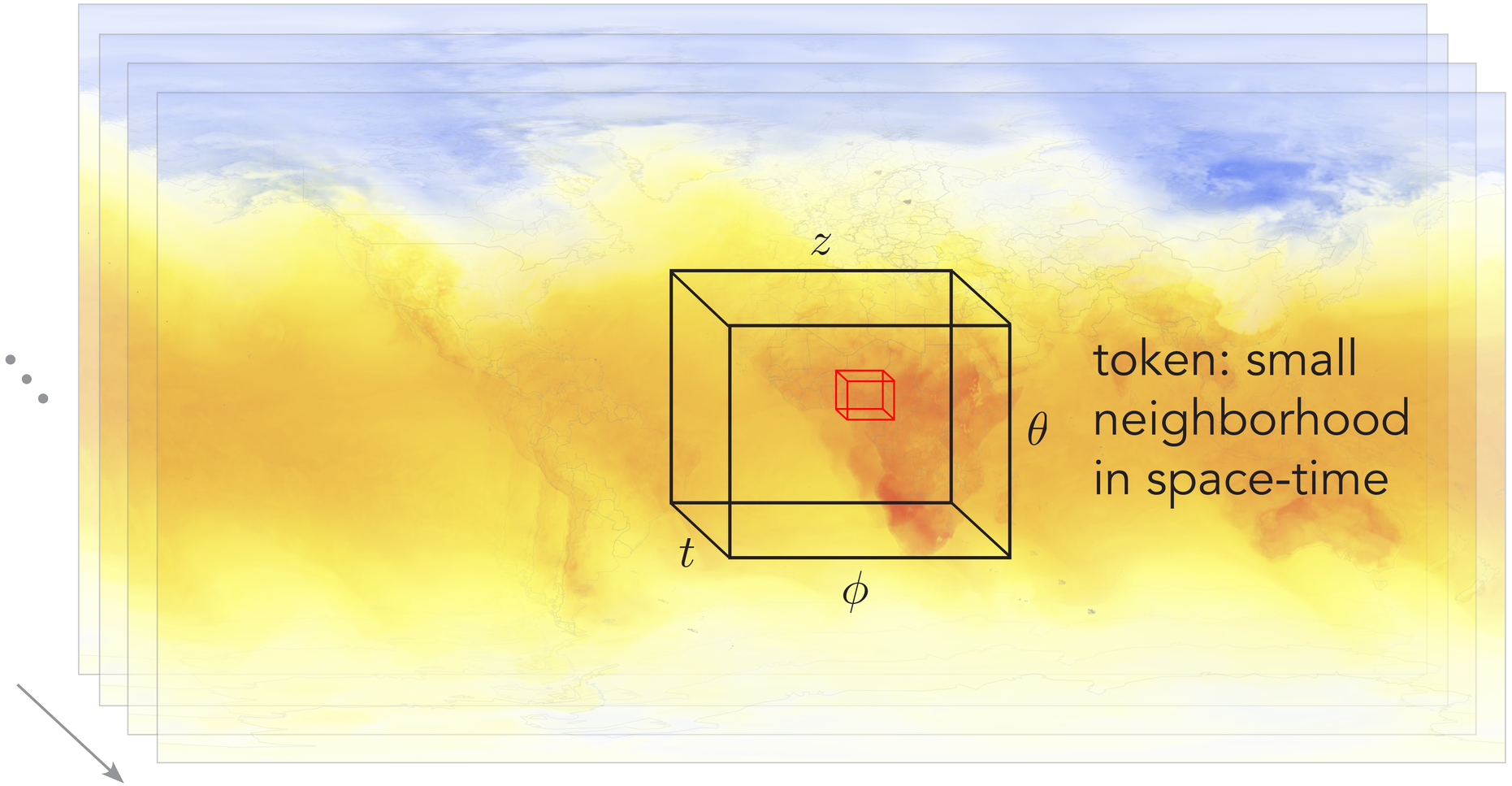
- Key property of transformers: sequence of inputs is processed simultaneously
  - Language: input is sentence with words being tokens
  - Images: small image patches



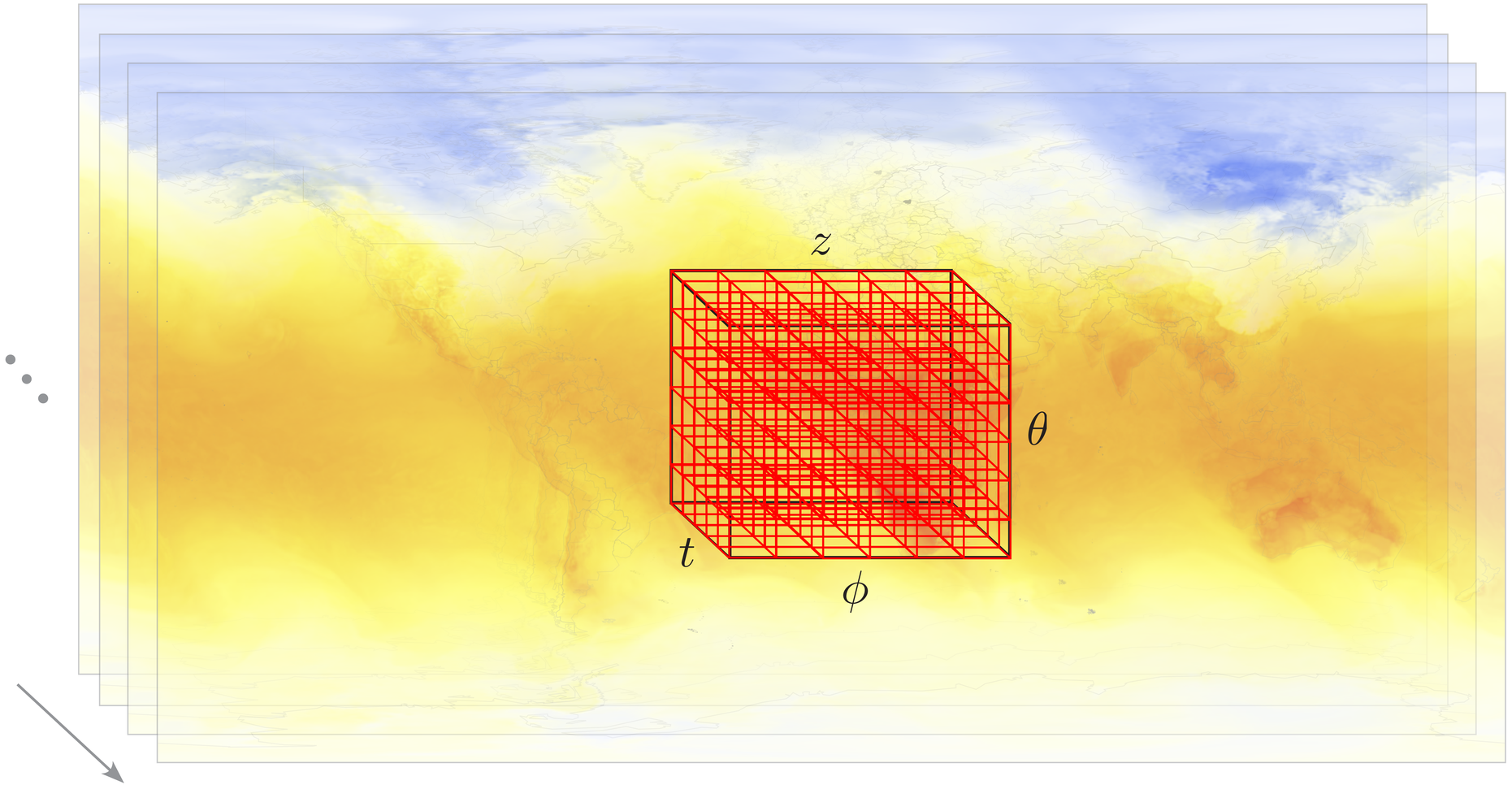
What is a token?



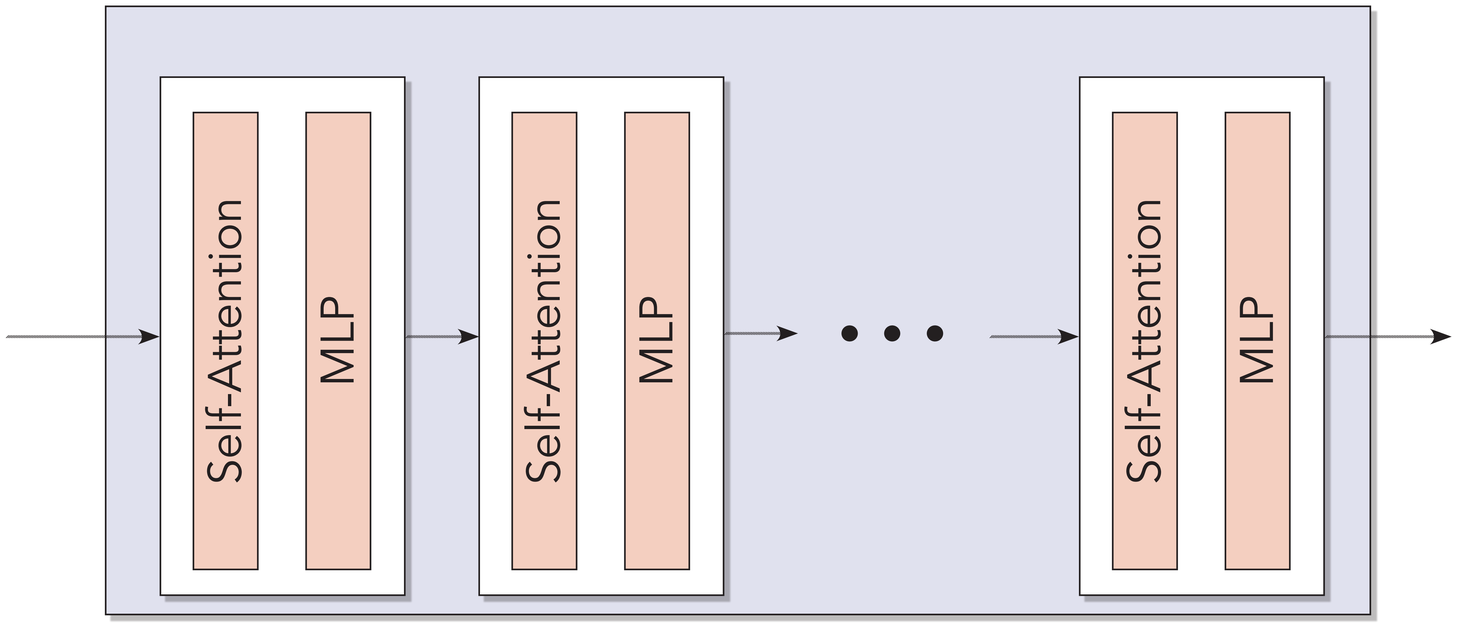
What is a token?



What is a token?



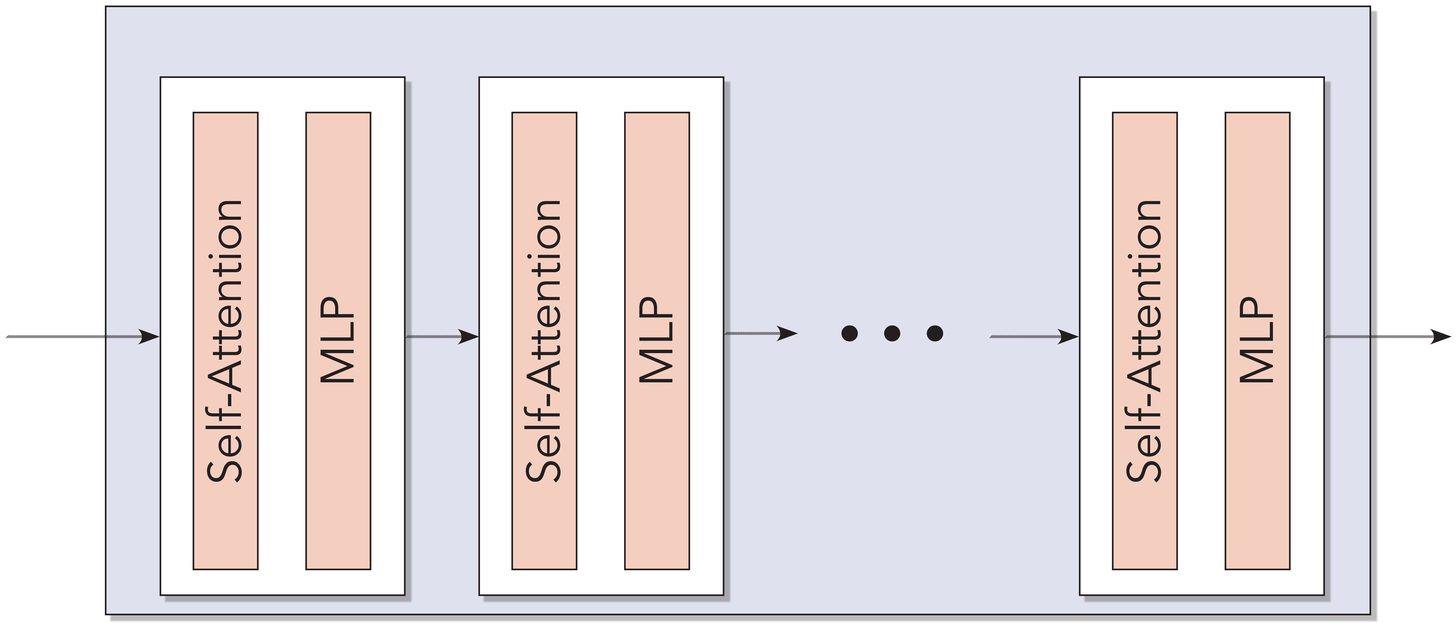
Multiformer



Multiformer

Self
attention

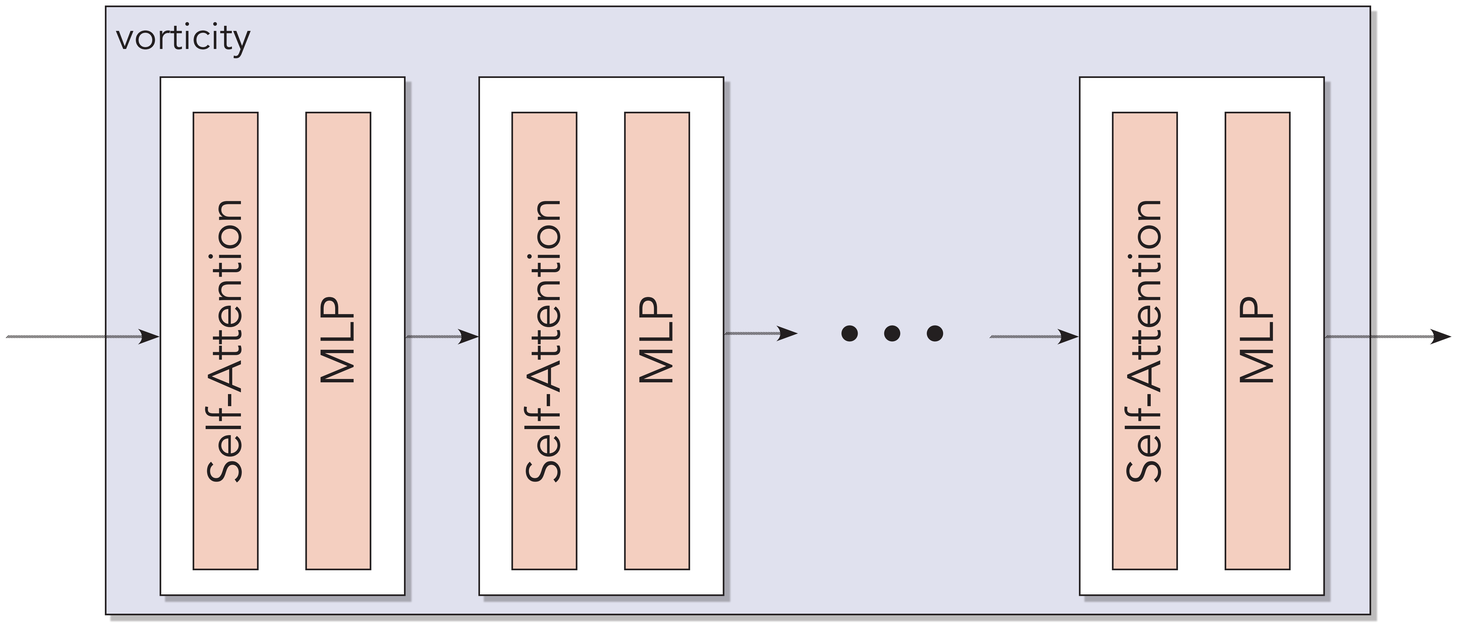
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

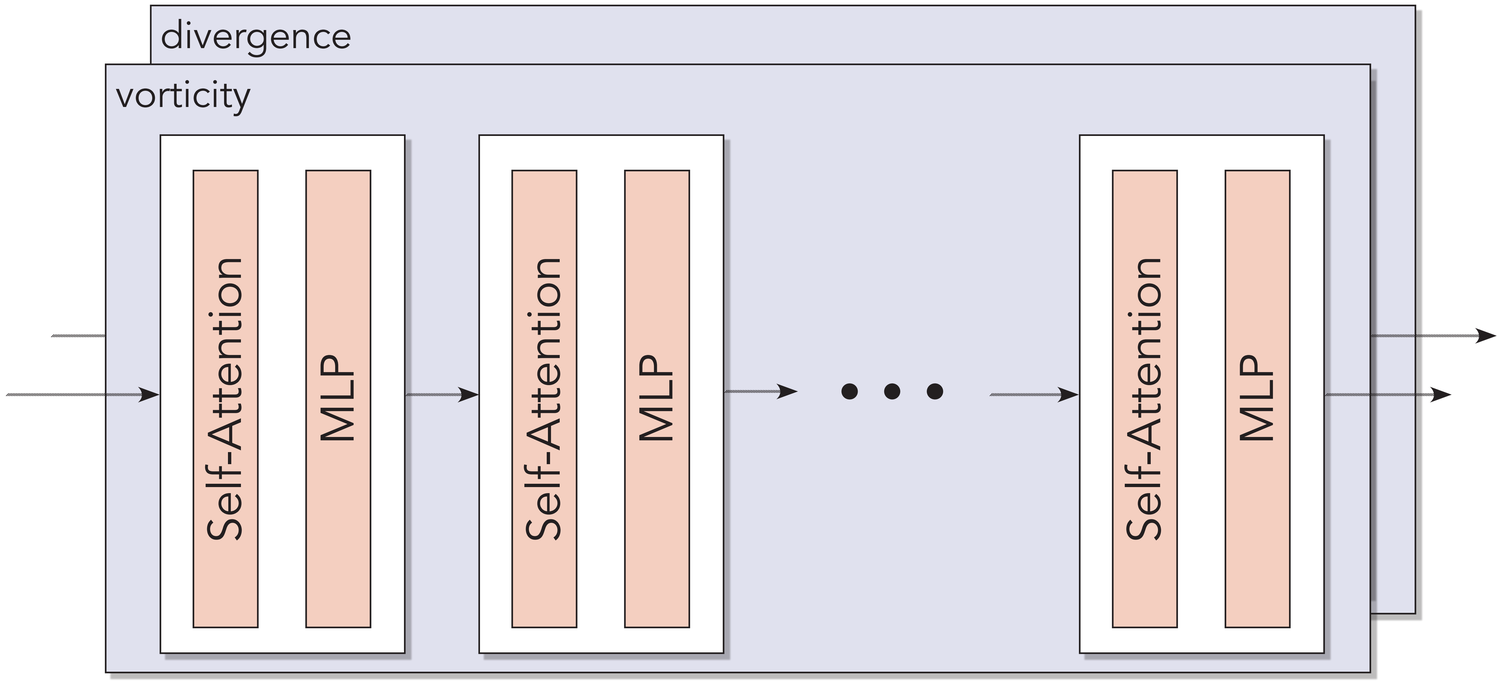
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

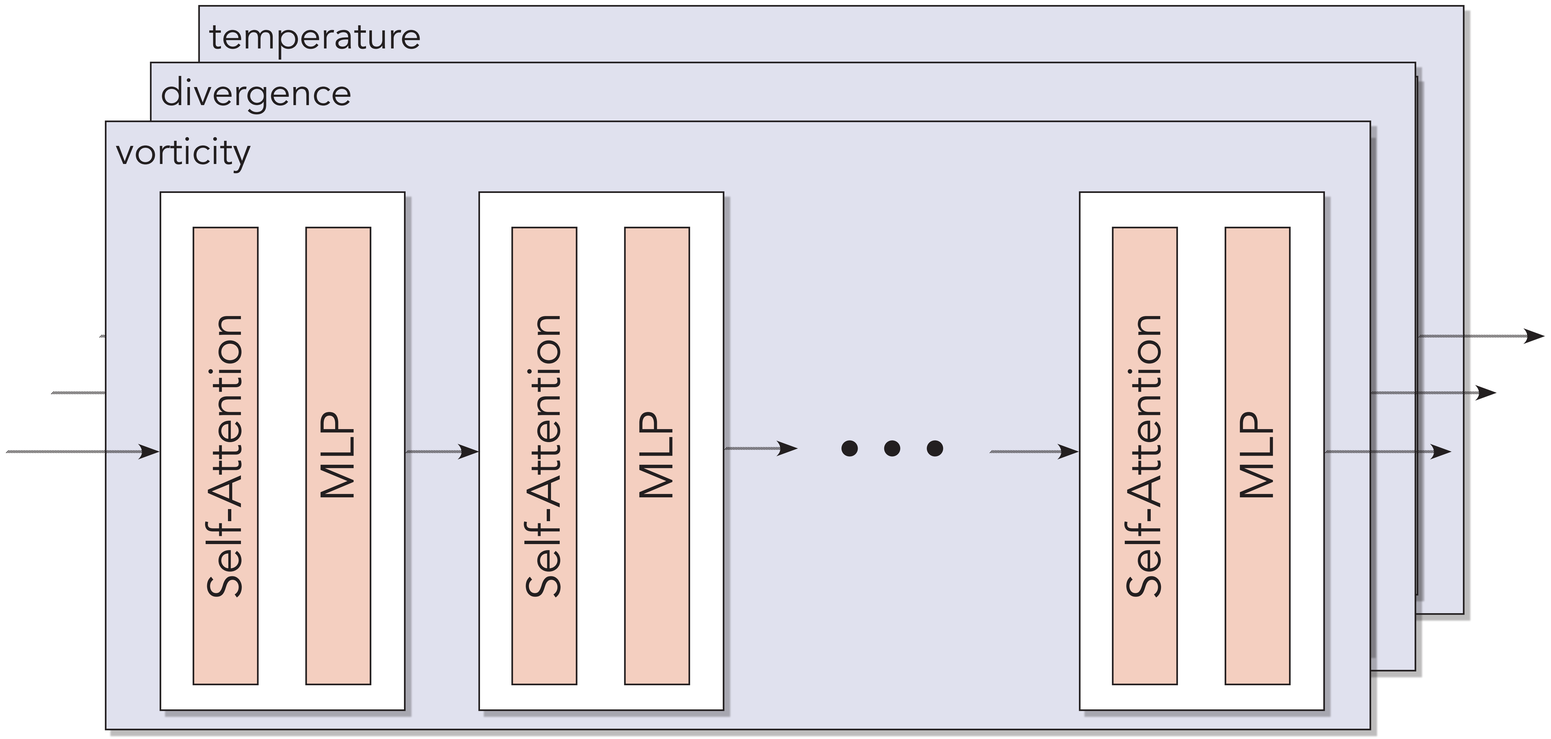
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

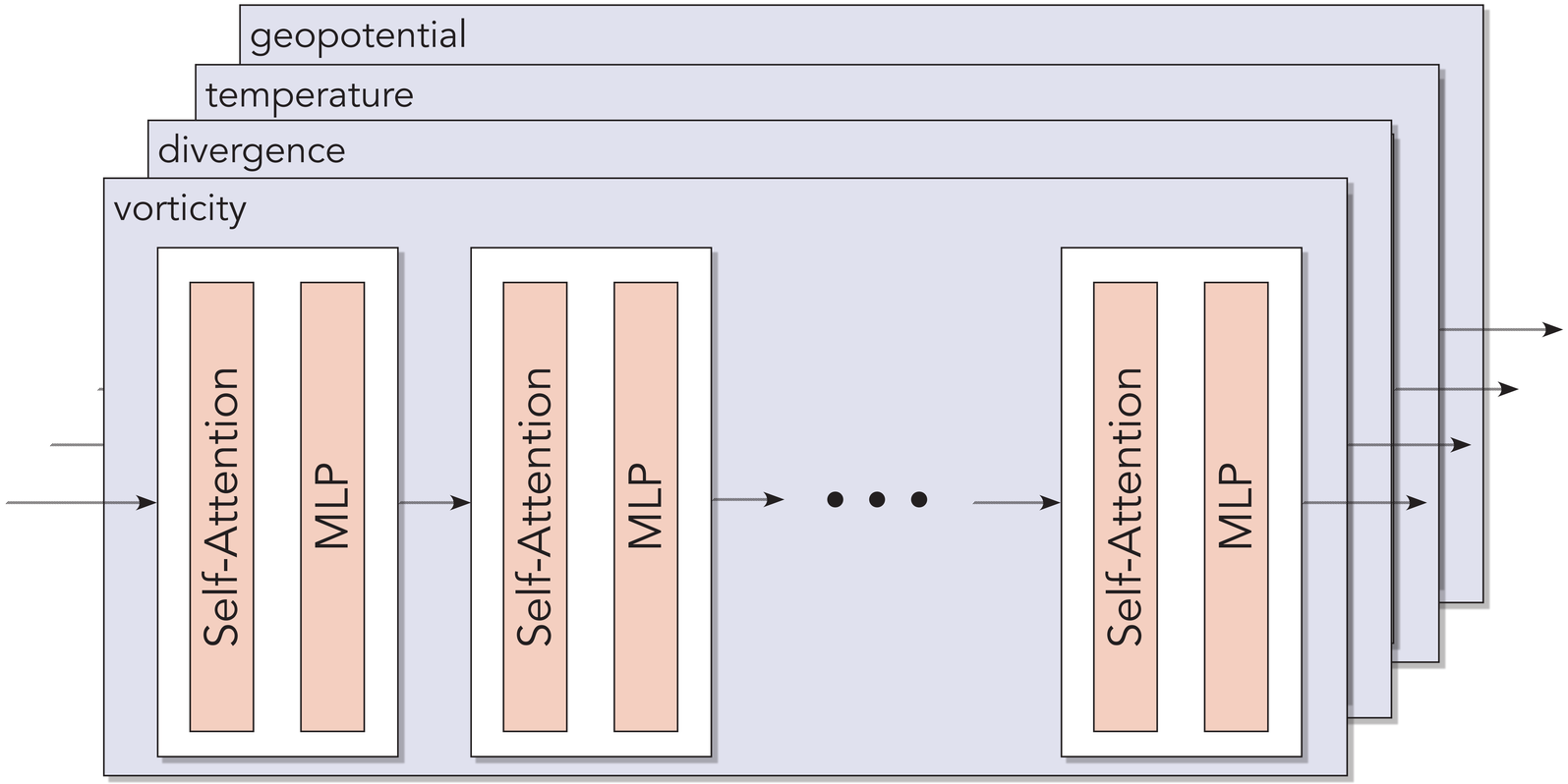
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

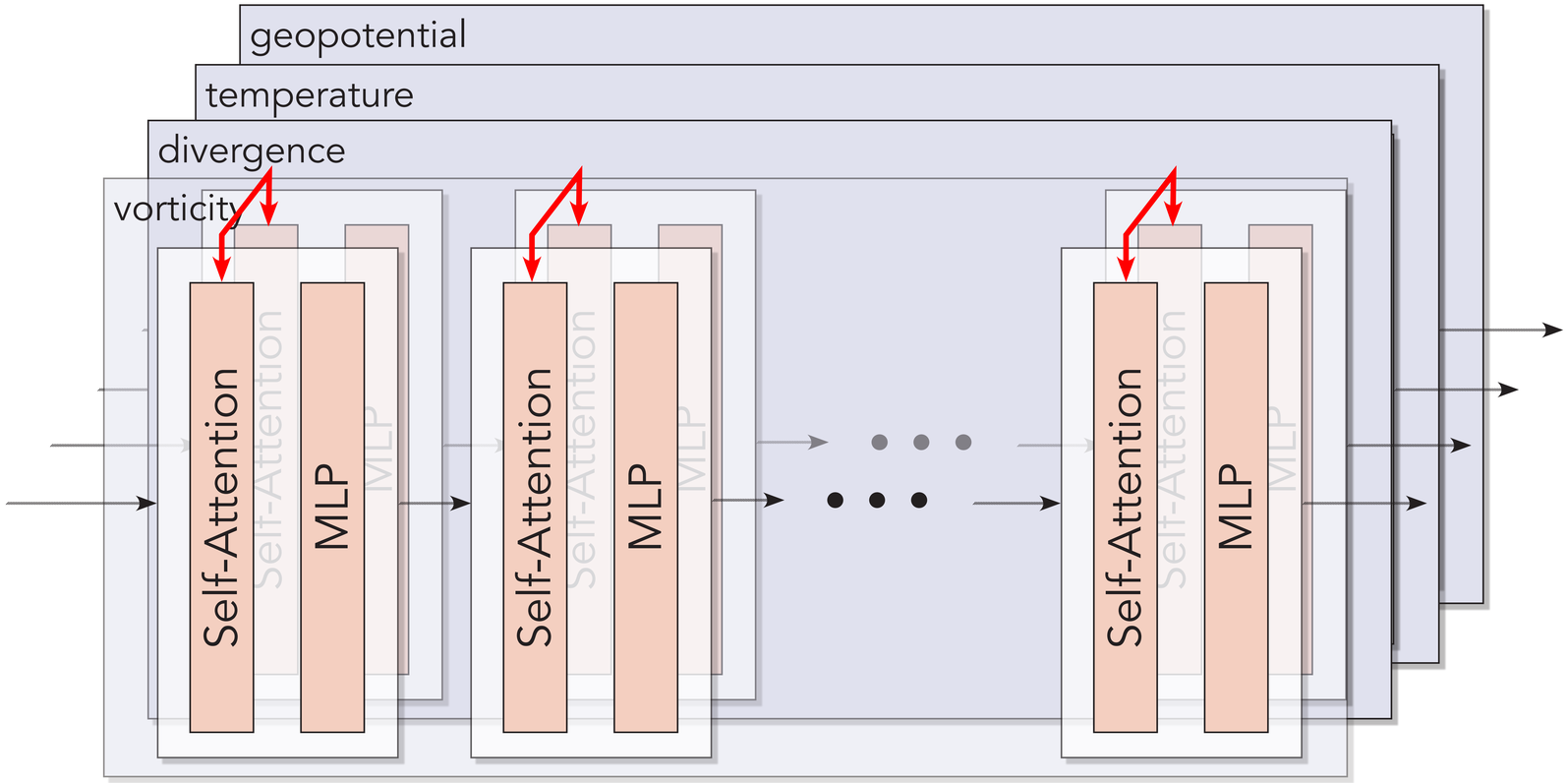
$$\sigma\left(Q\,K^T\right)V$$



Multiformer

Self
attention

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



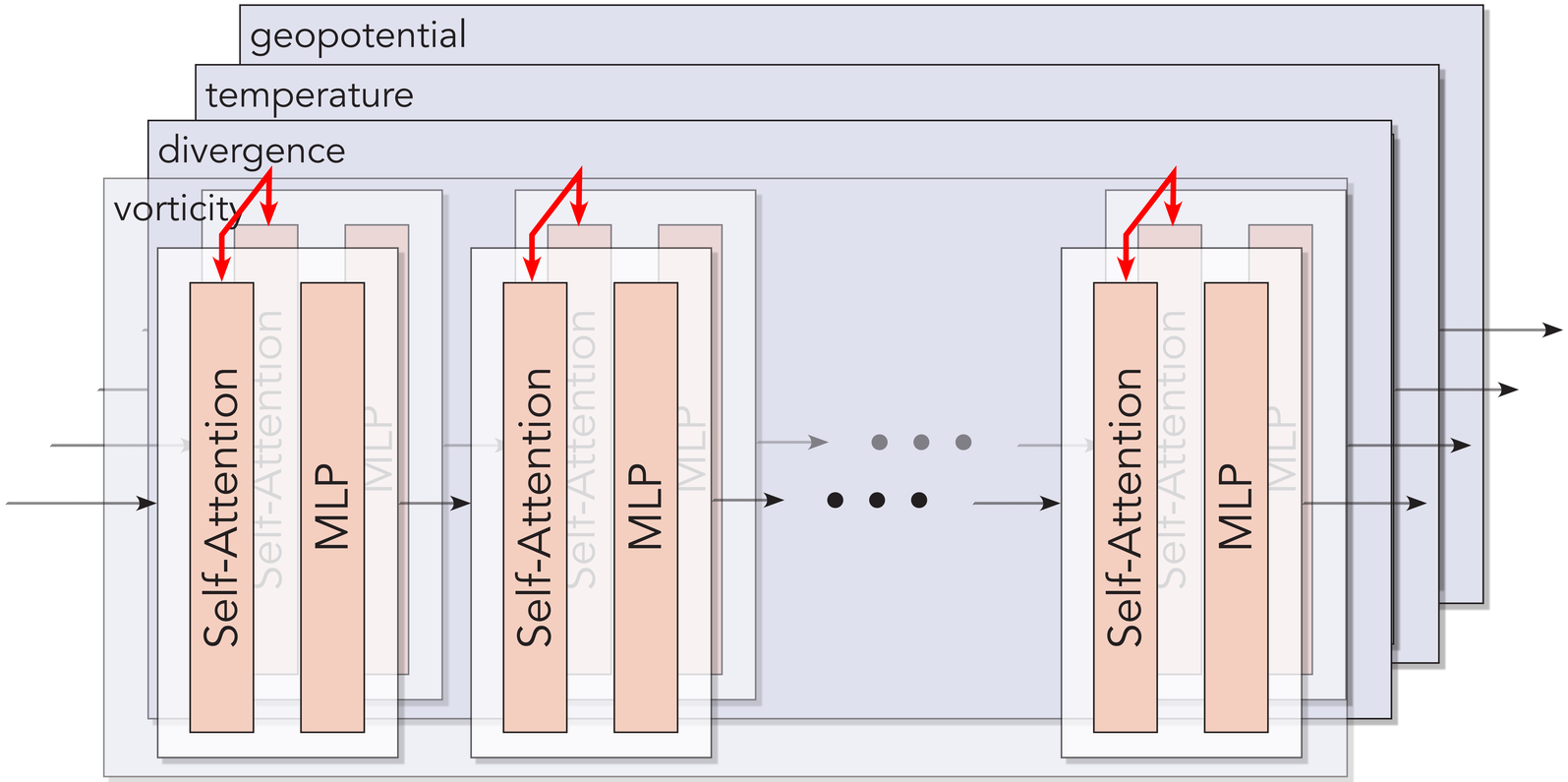
Multiformer

Self attention

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

Cross attention

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



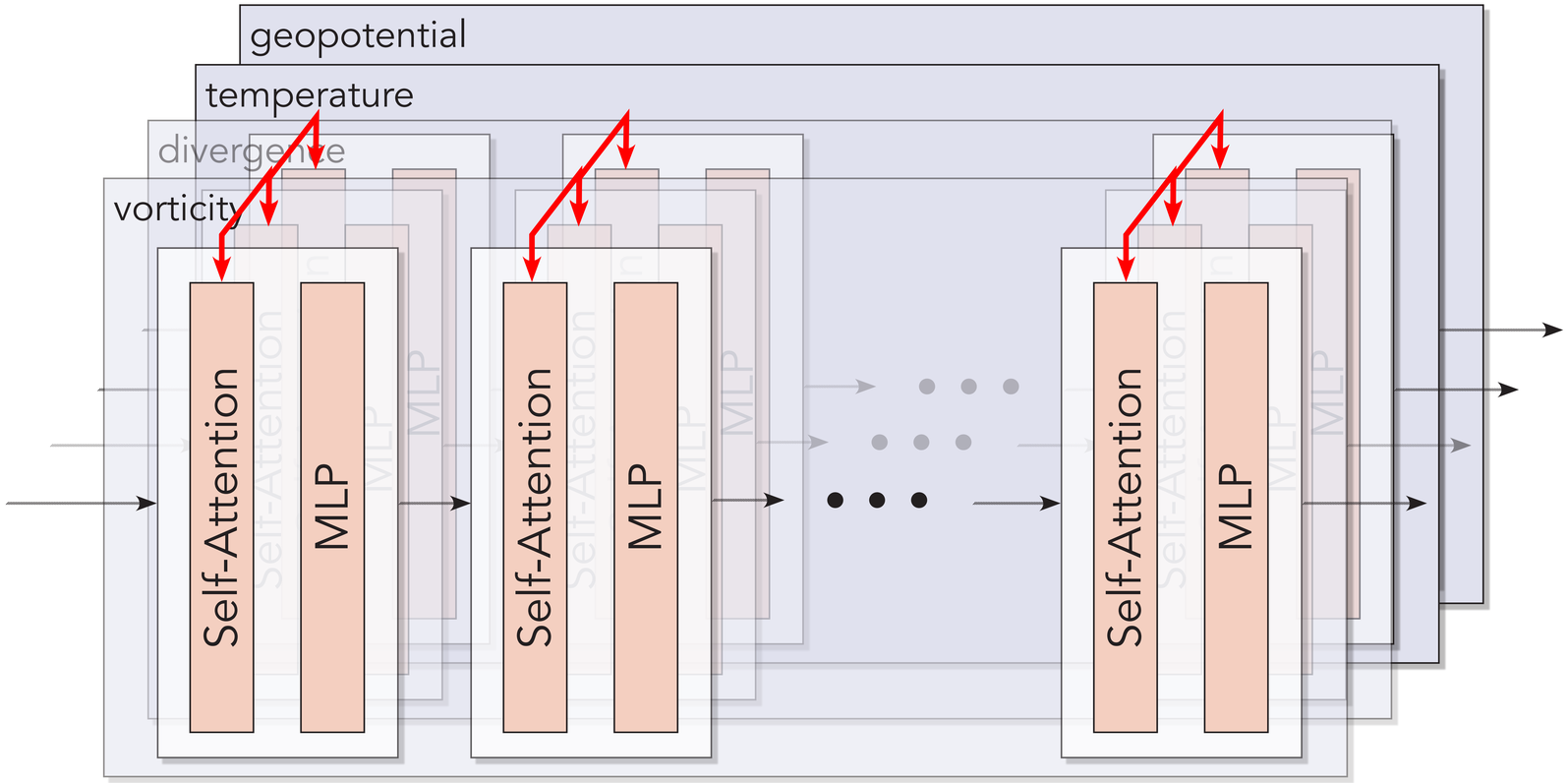
Multiformer

Self attention

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

Cross attention

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



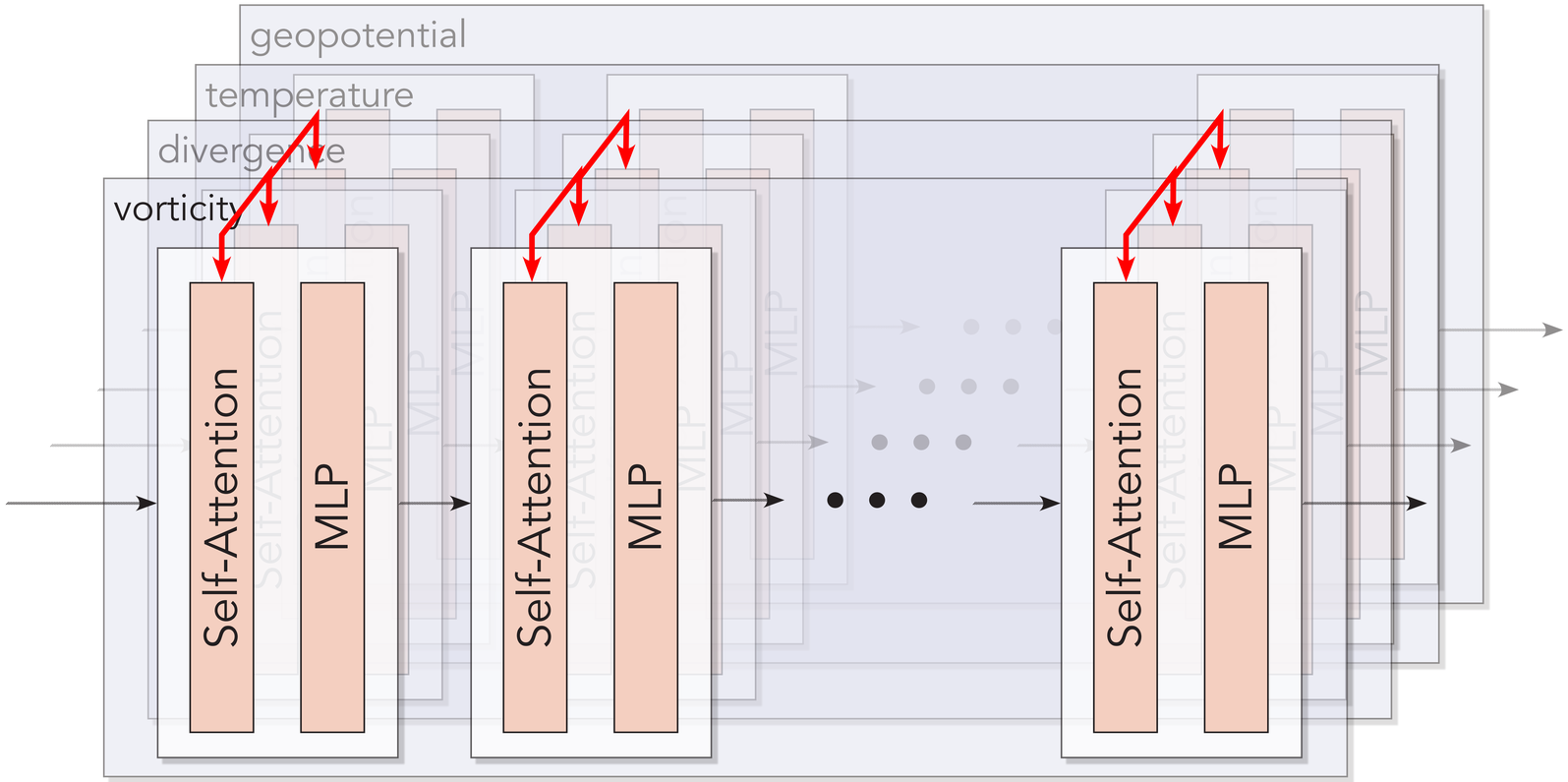
Multiformer

Self attention

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

Cross attention

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



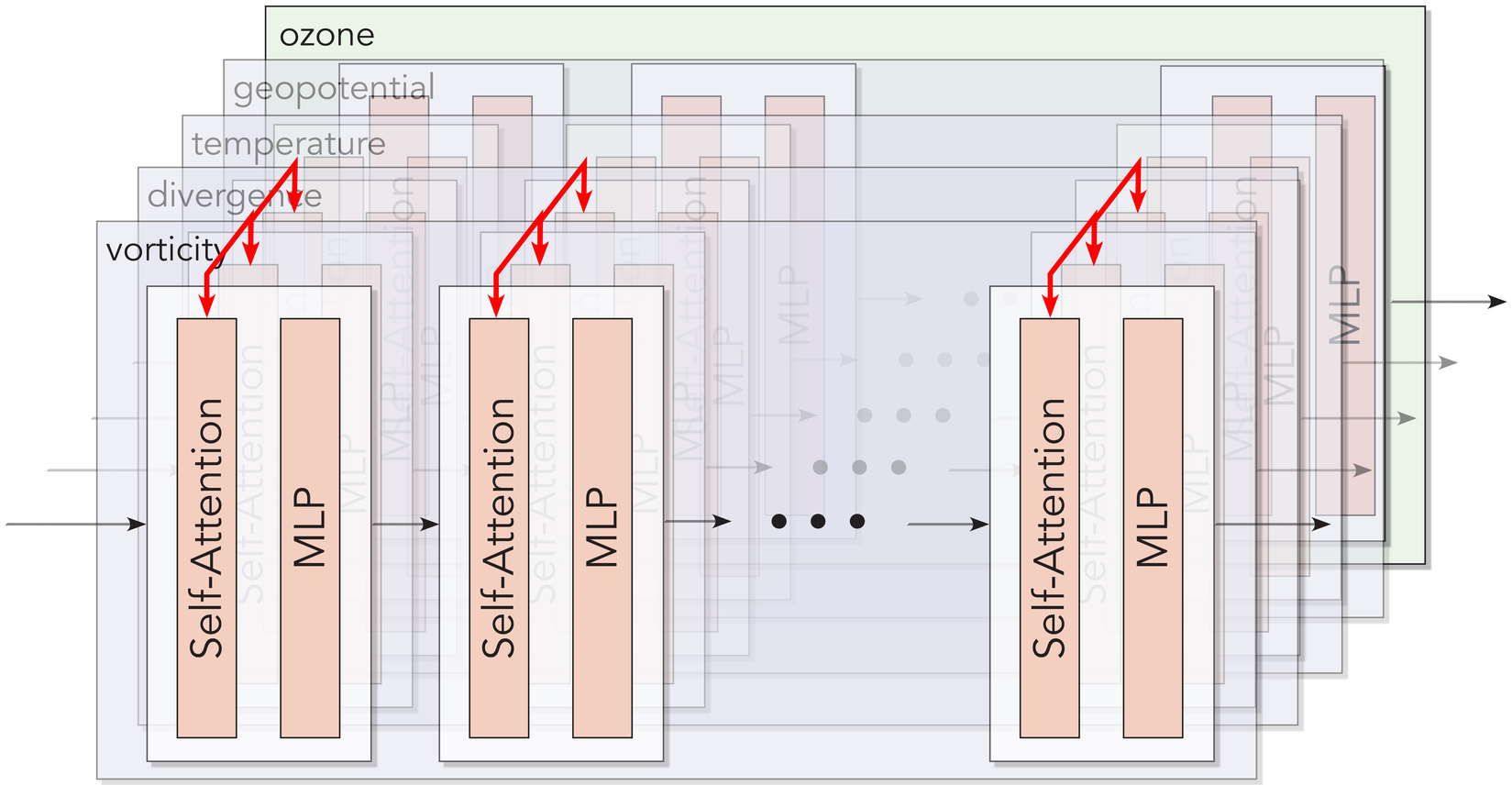
Multiformer

Self attention

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

Cross attention

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



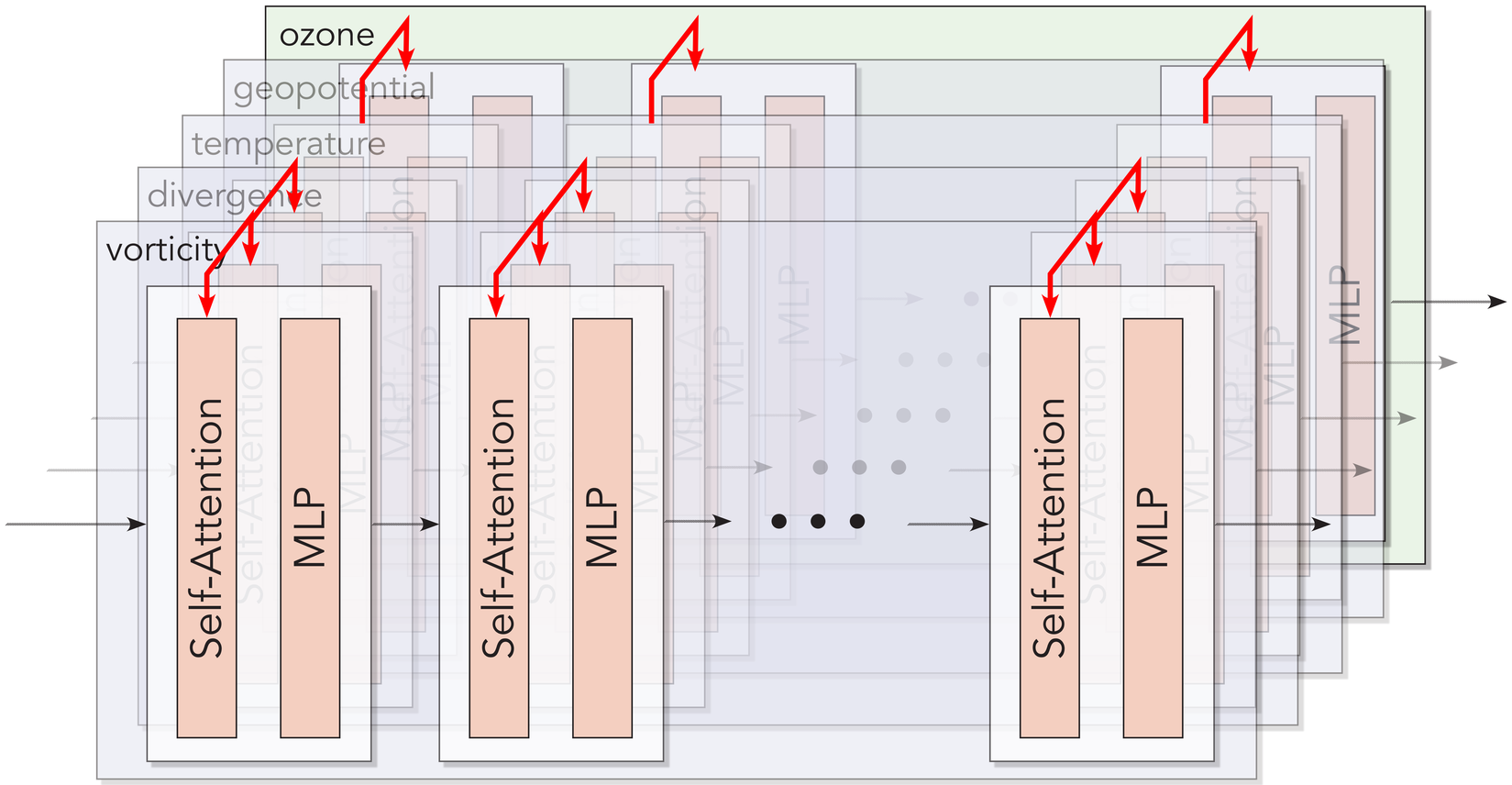
Multiformer

Self attention

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

Cross attention

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



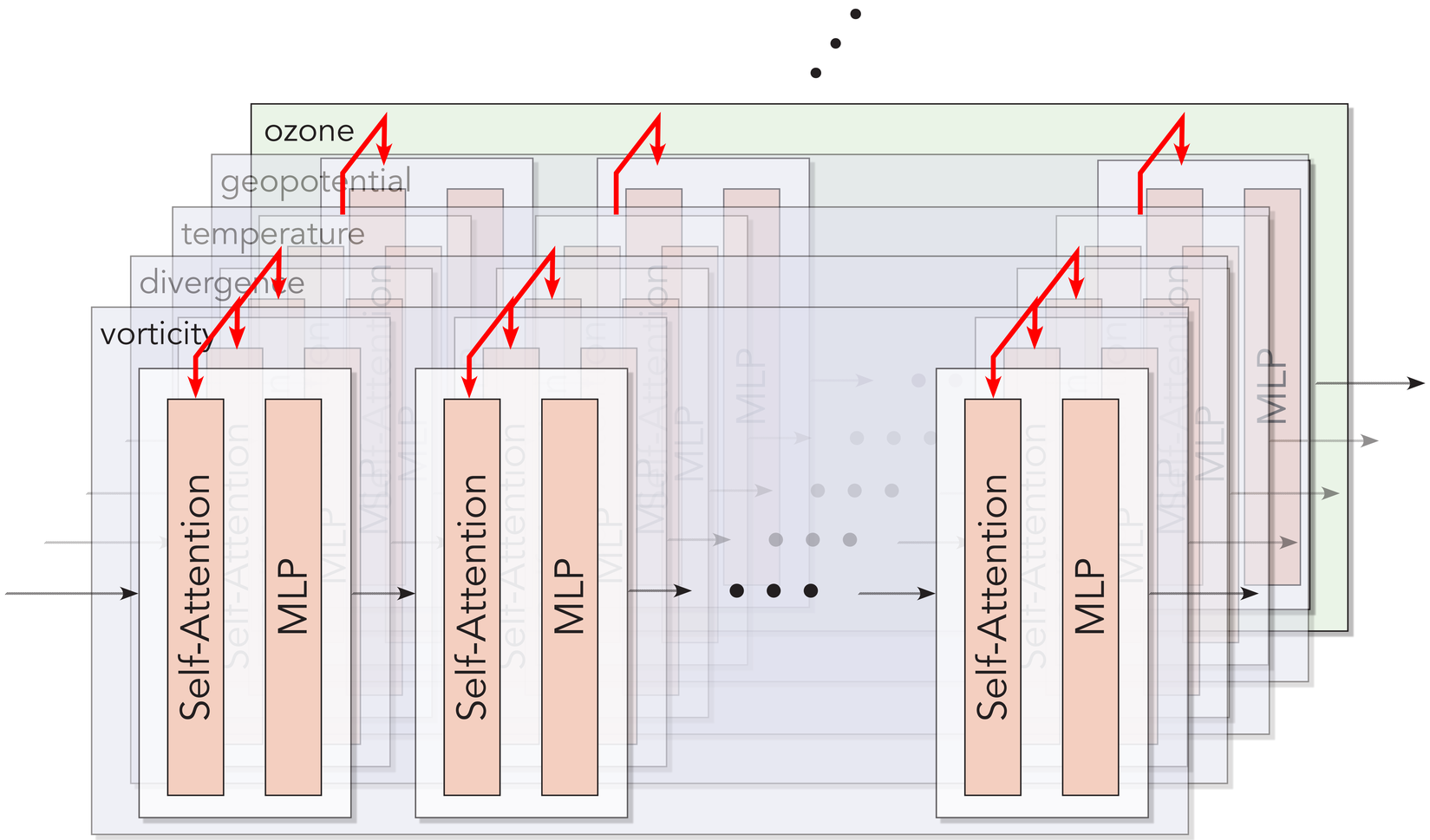
Multiformer

Self attention

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

Cross attention

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



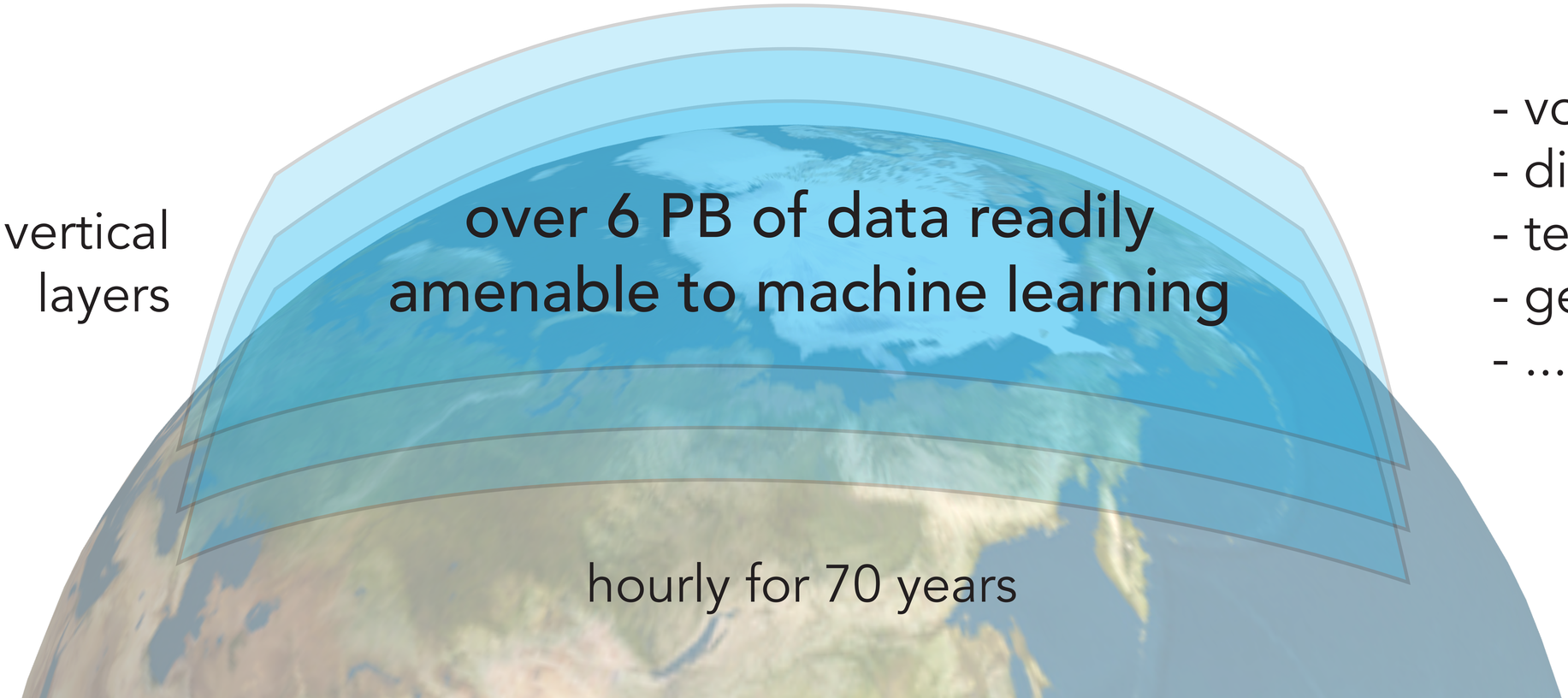
Multiformer

- 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
- 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

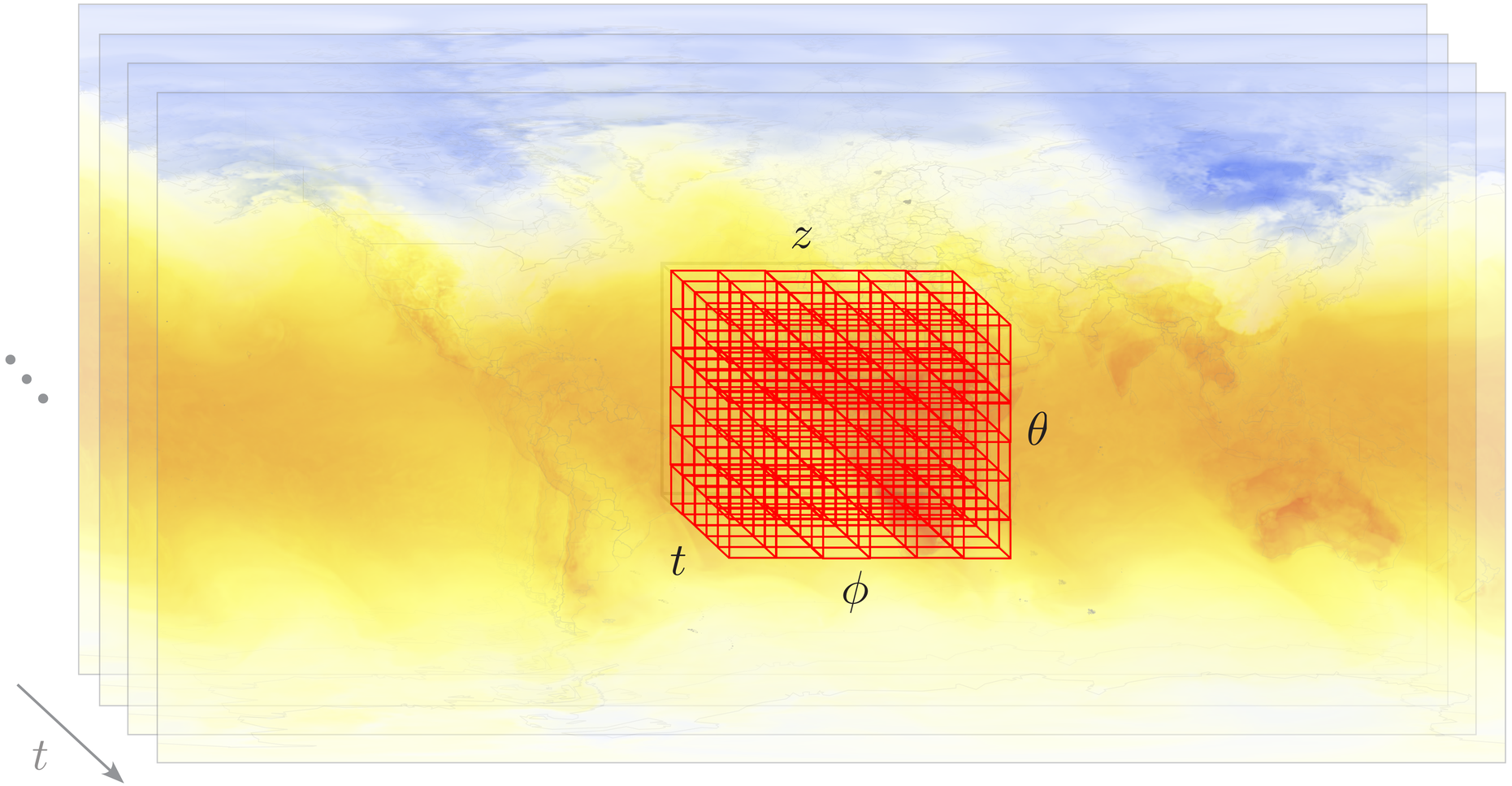
- vorticity
- divergence
- temperature
- geopotential
- ...

hourly for 70 years

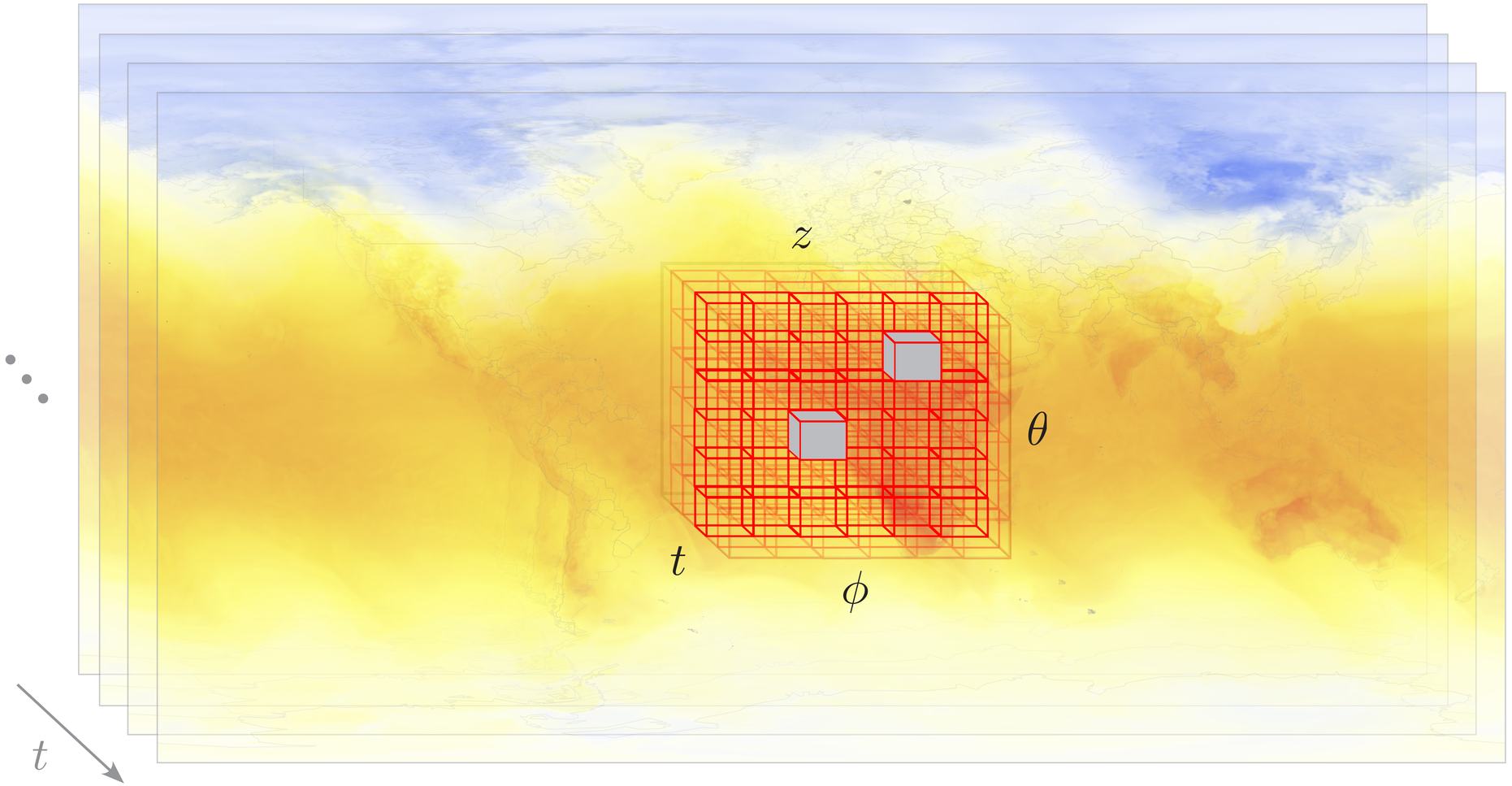
Spatio-temporal BERT

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

Spatio-temporal BERT

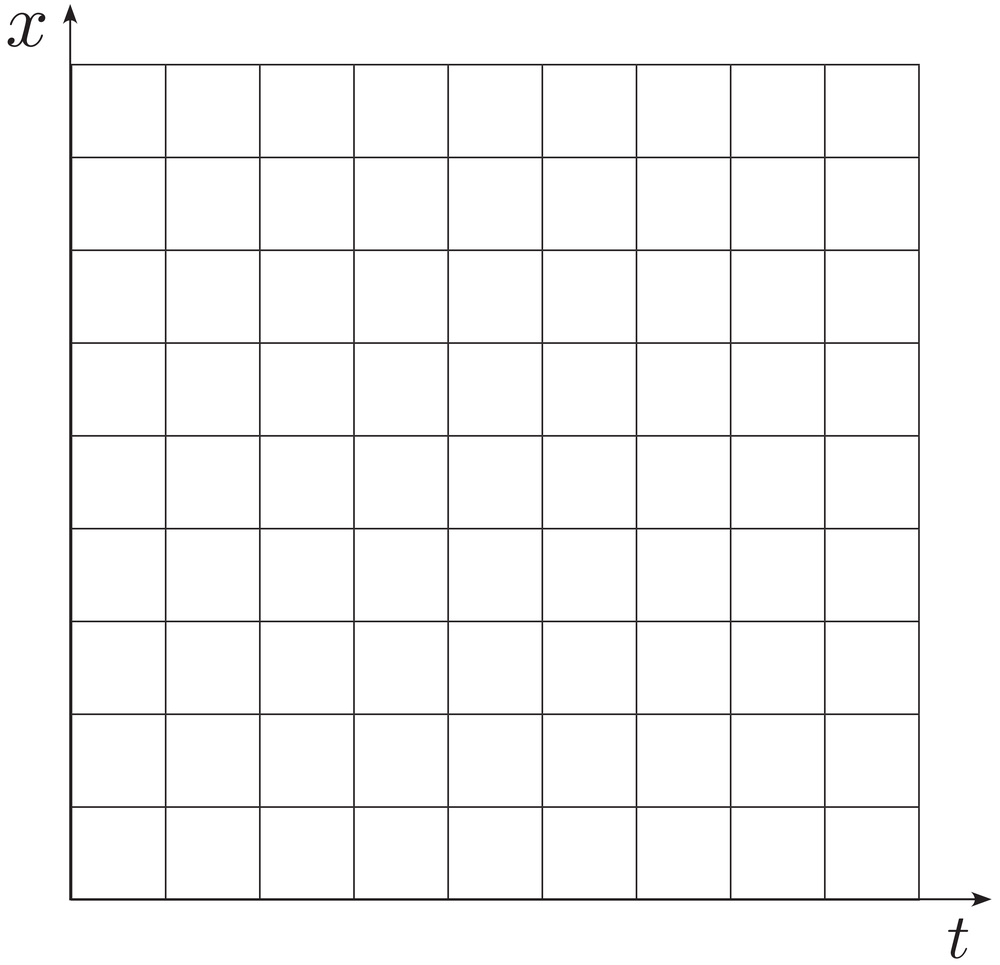


Spatio-temporal BERT



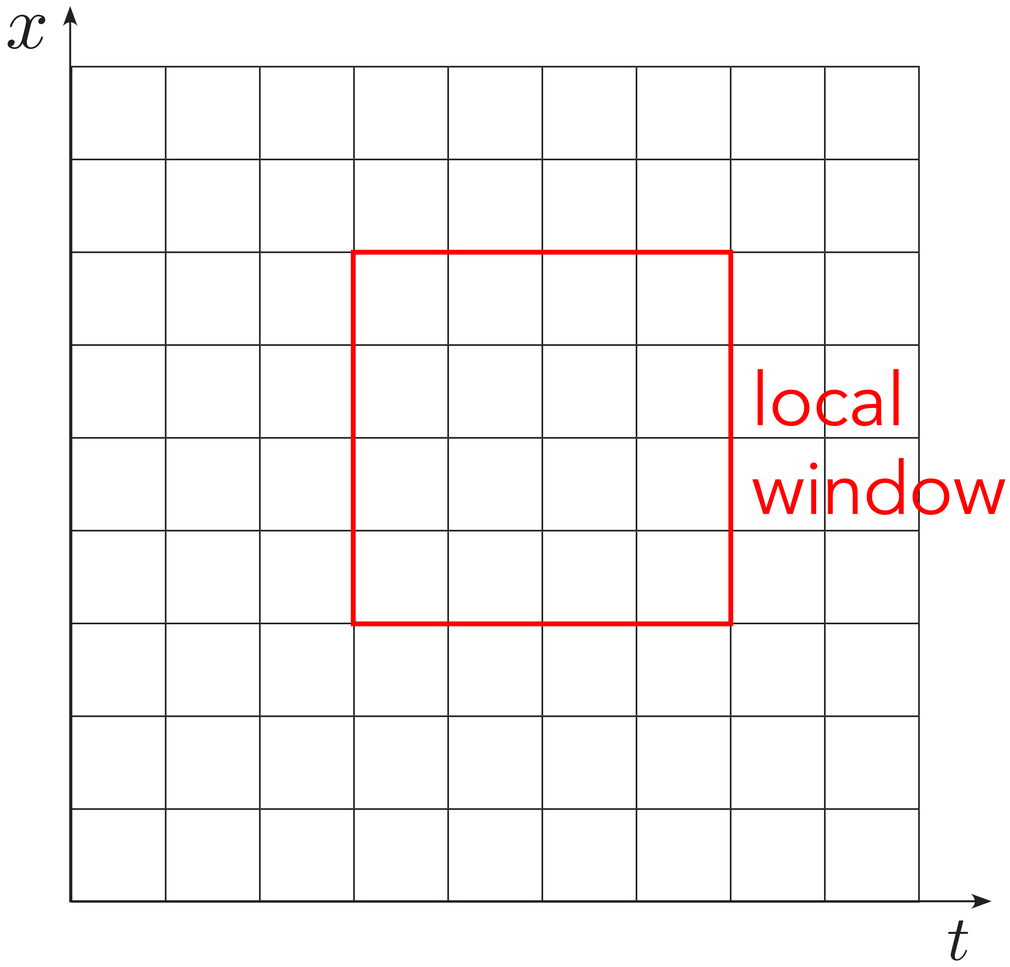
Spatio-temporal BERT

Flatland
view



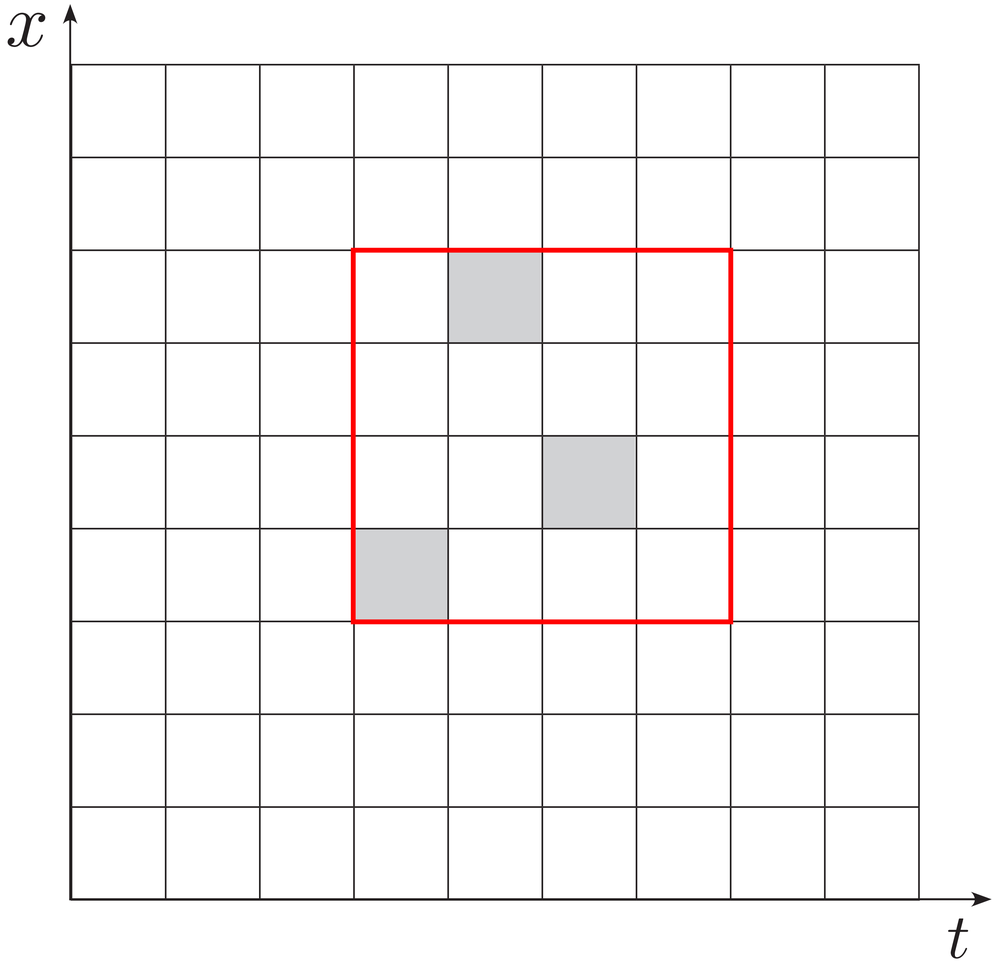
Spatio-temporal BERT

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

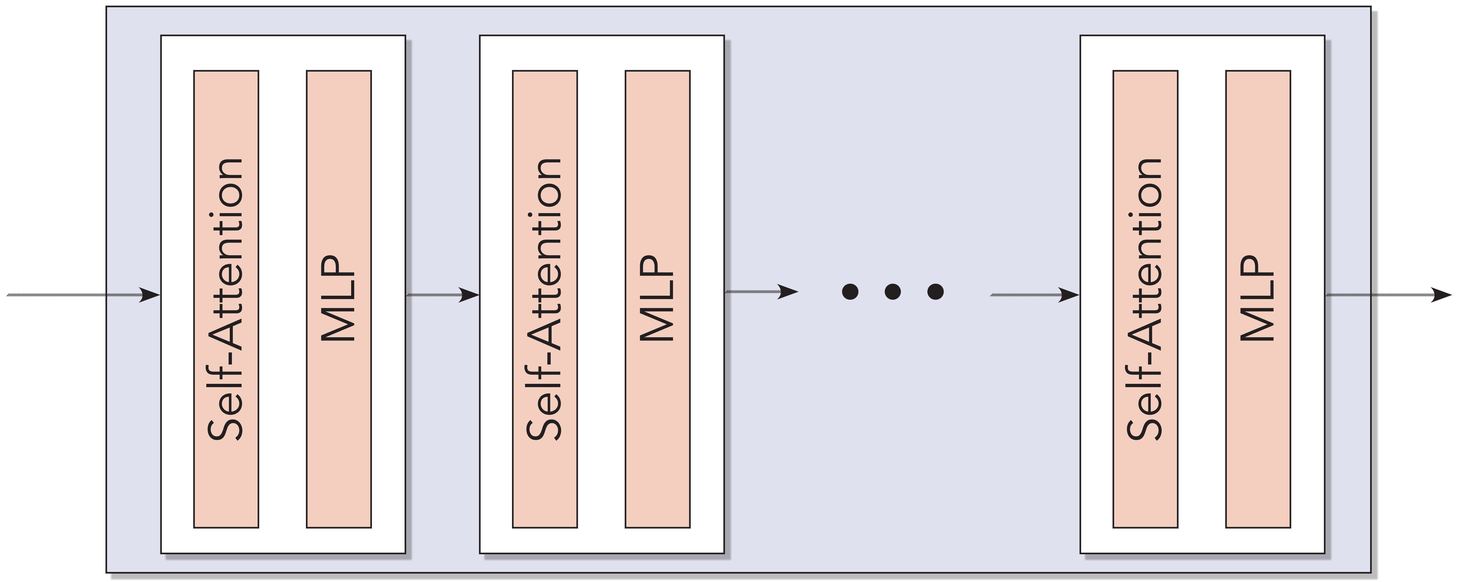
Statistical loss

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

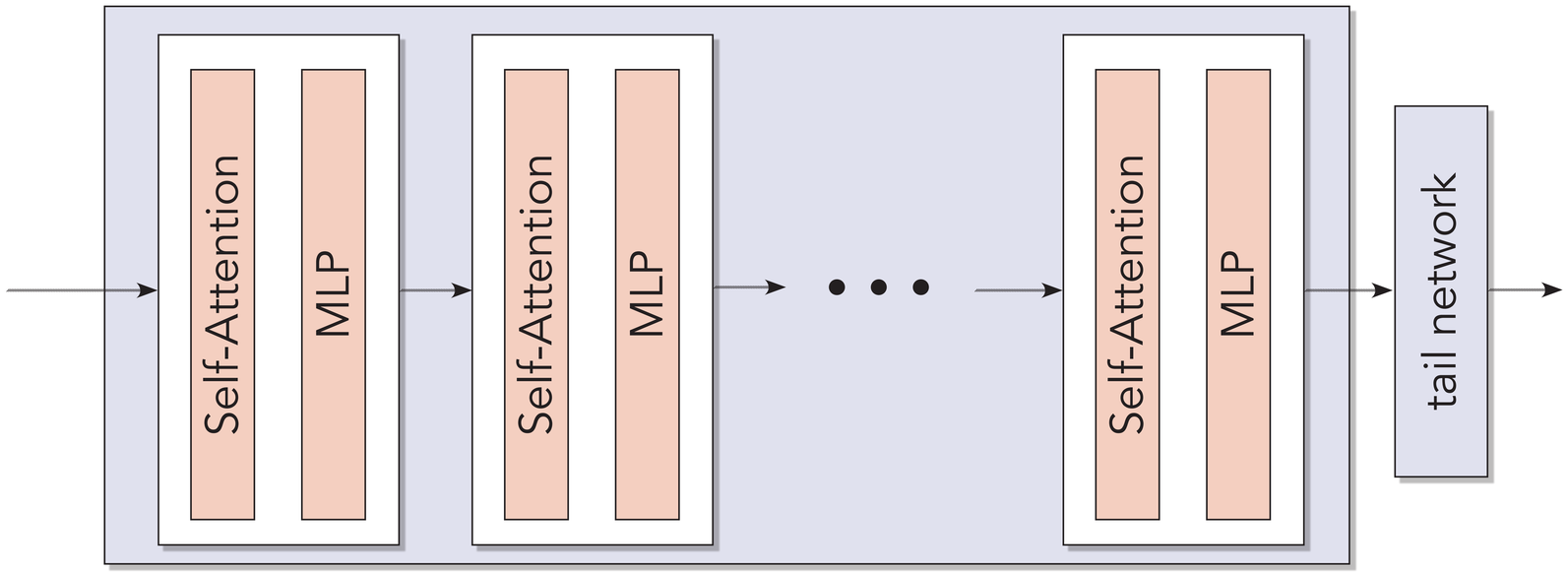
Statistical loss

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

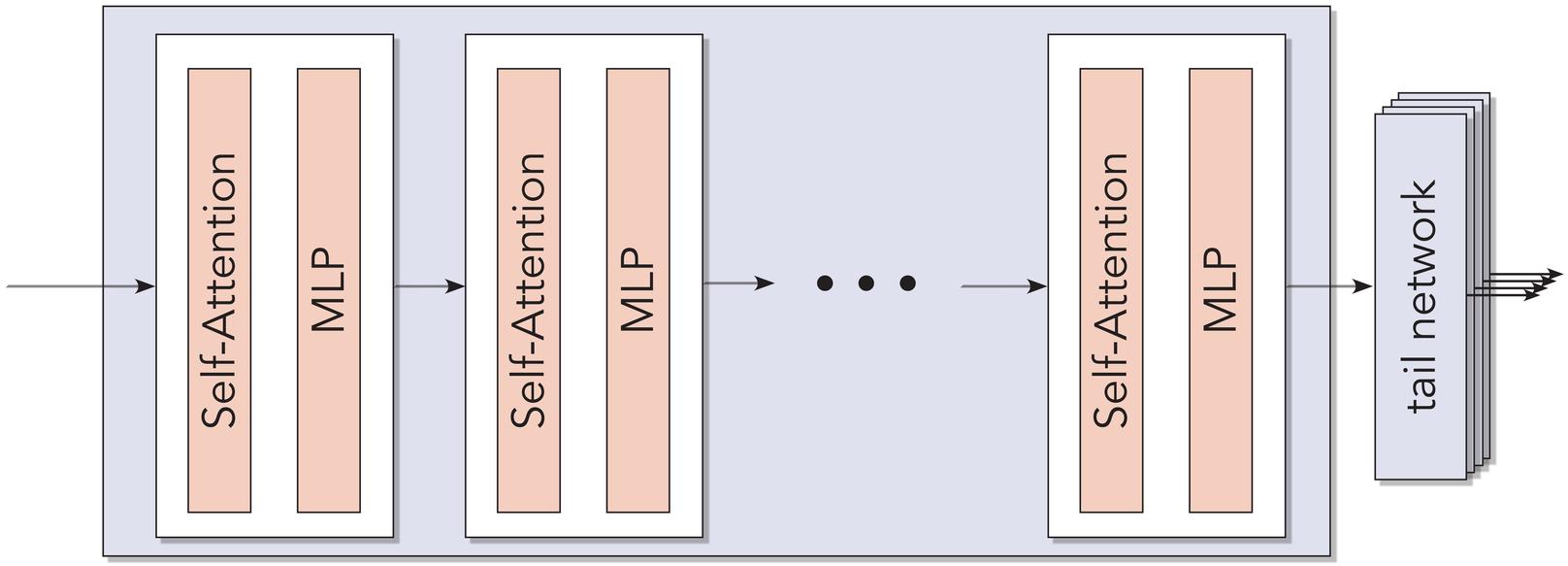
Statistical loss



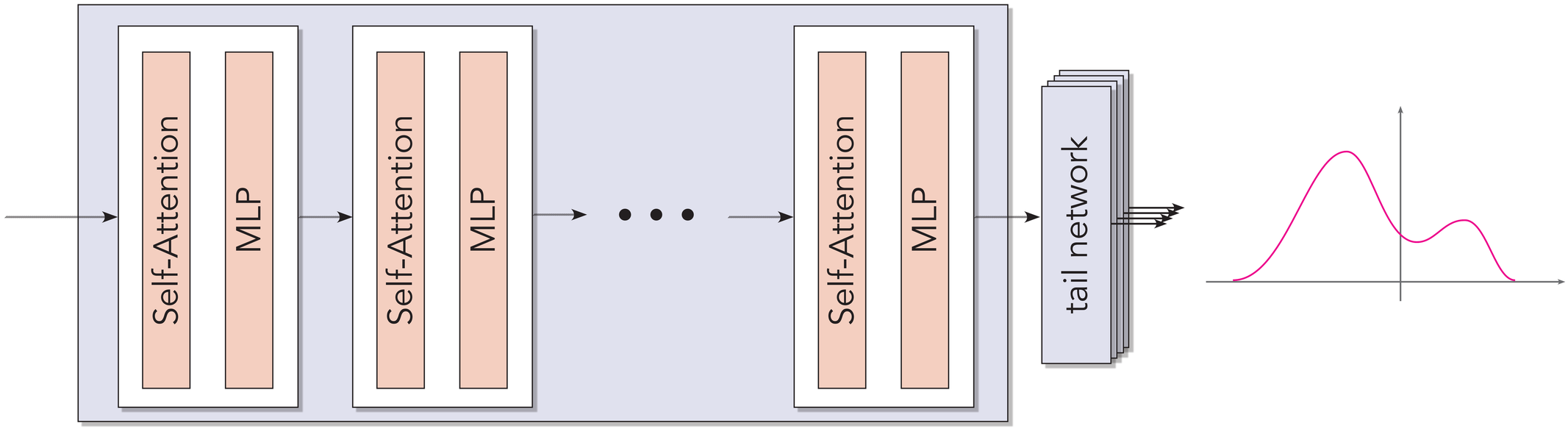
Statistical loss



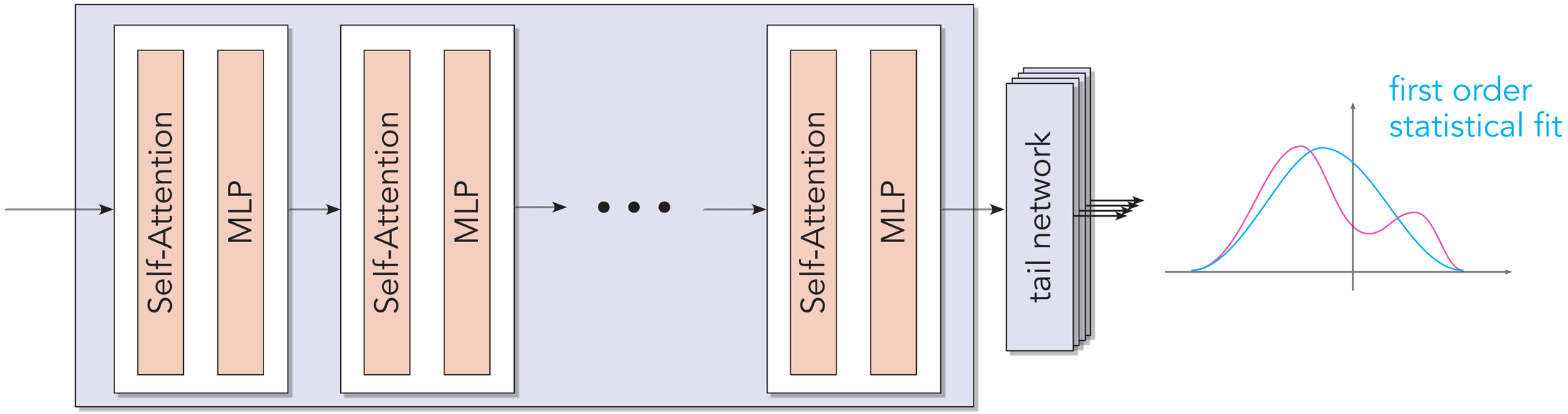
Statistical loss



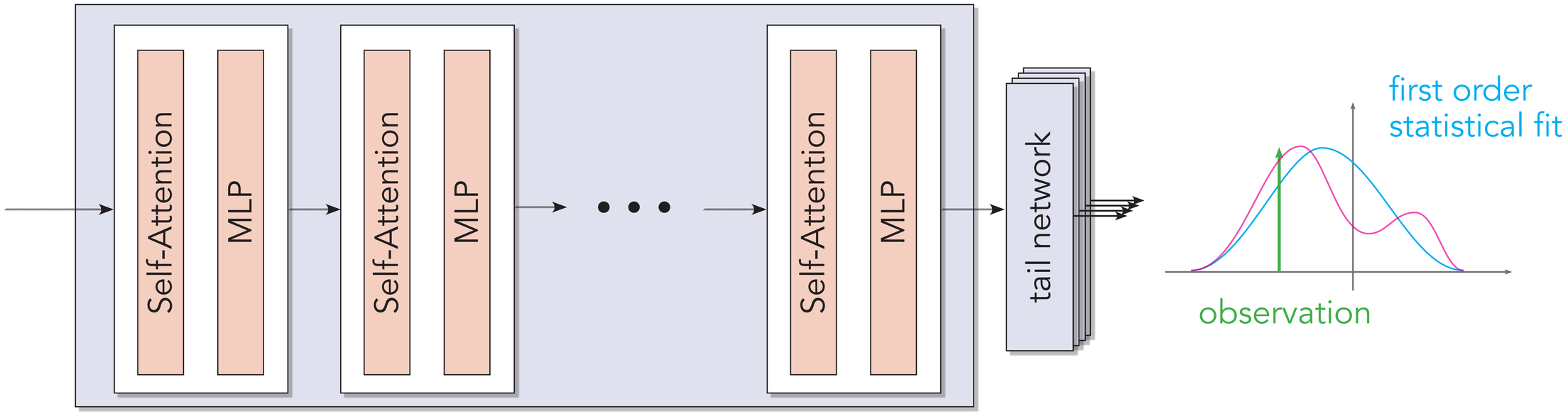
Statistical loss



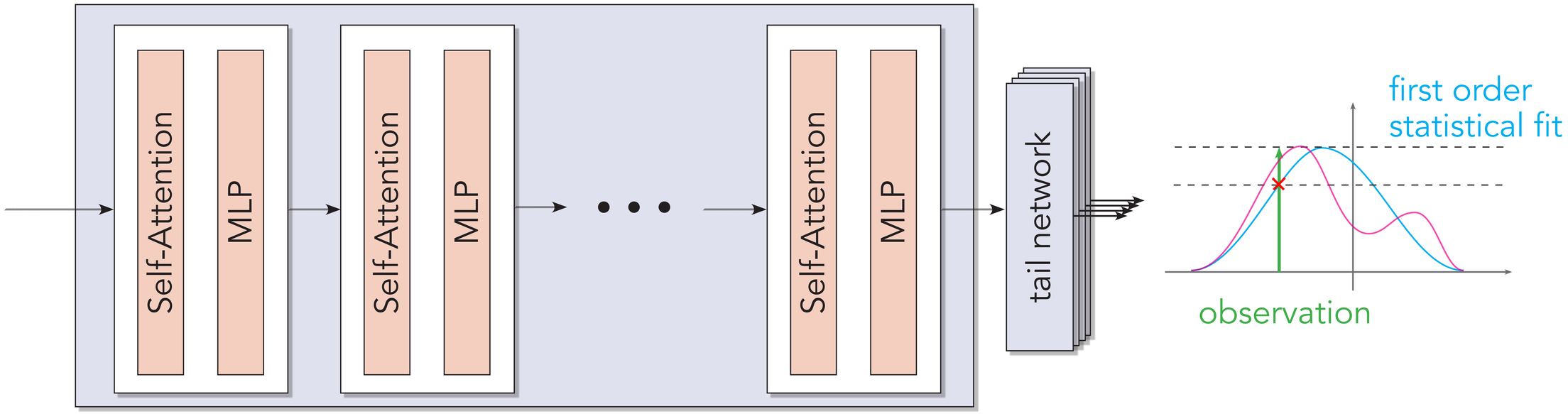
Statistical loss



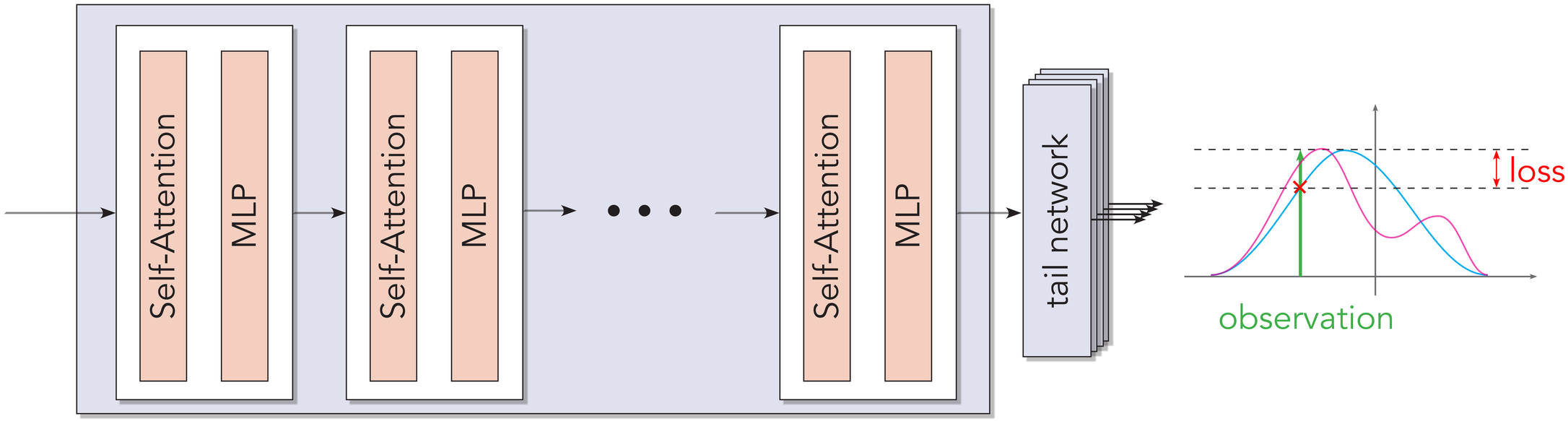
Statistical loss



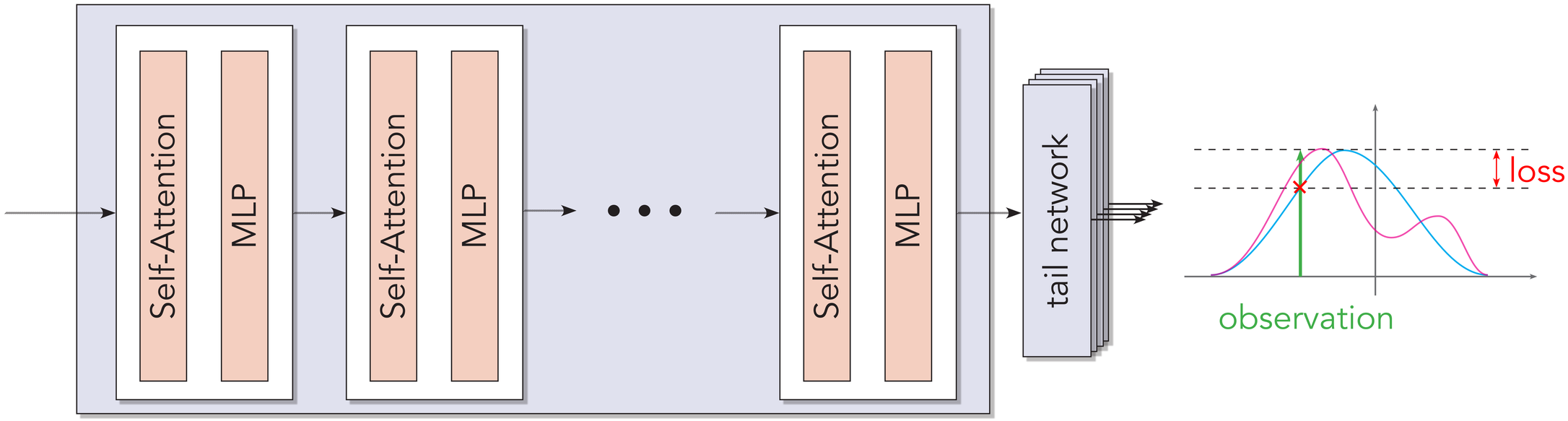
Statistical loss



Statistical loss



Statistical loss

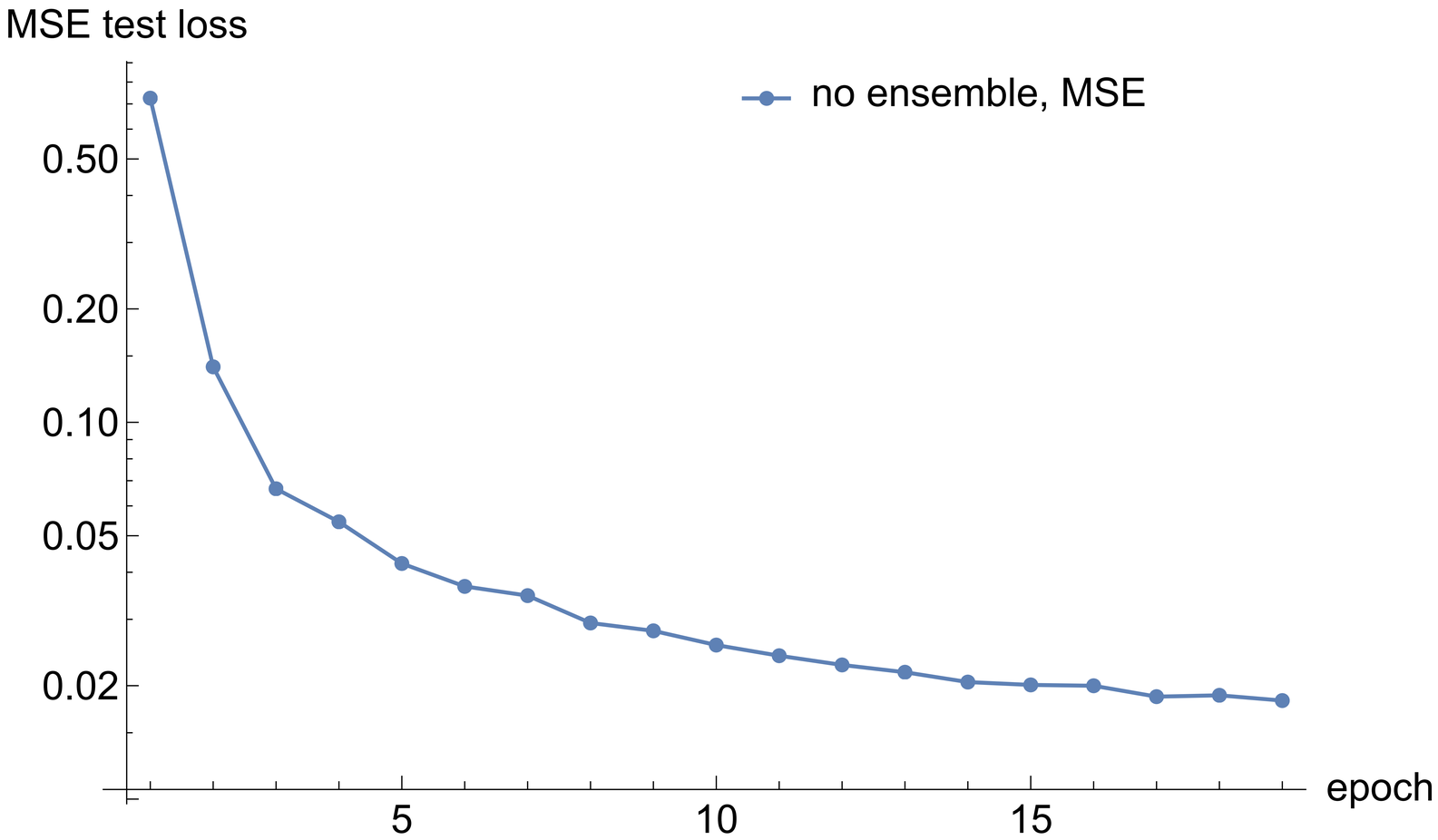


end-to-end training encourages statistical representation

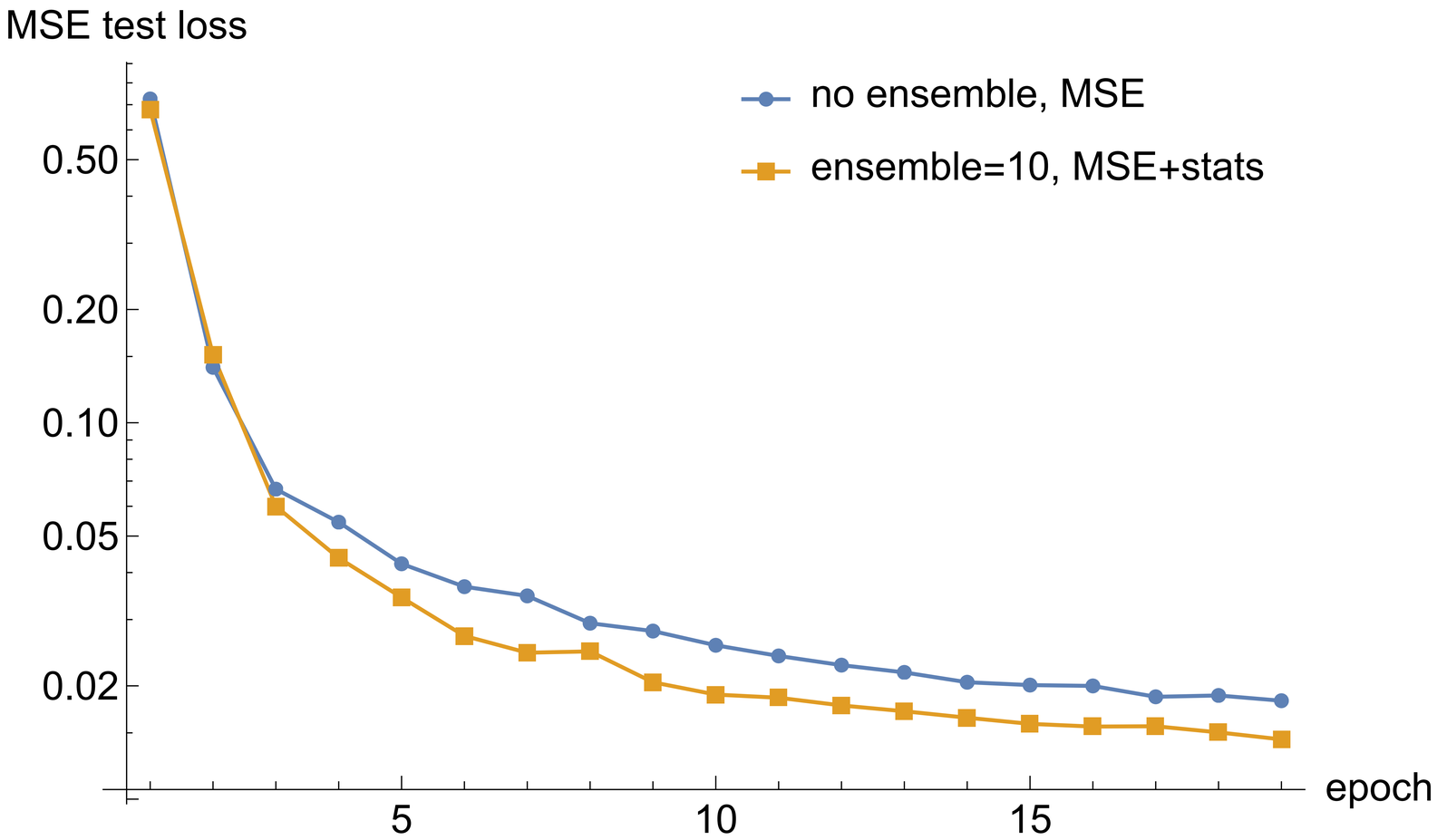
Statistical loss: experiments

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

Statistical loss

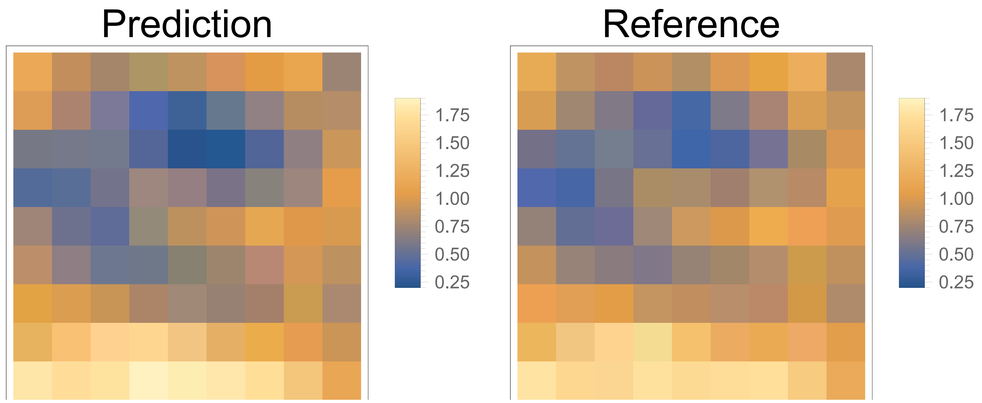


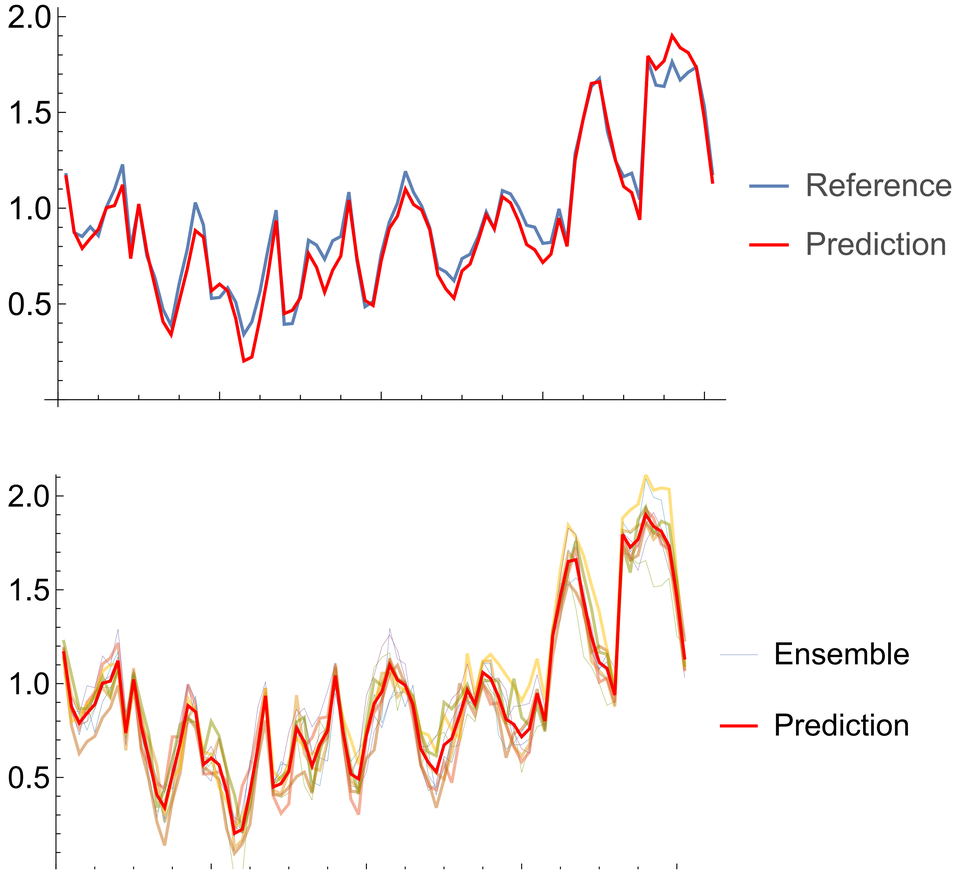
Statistical loss



Statistical loss

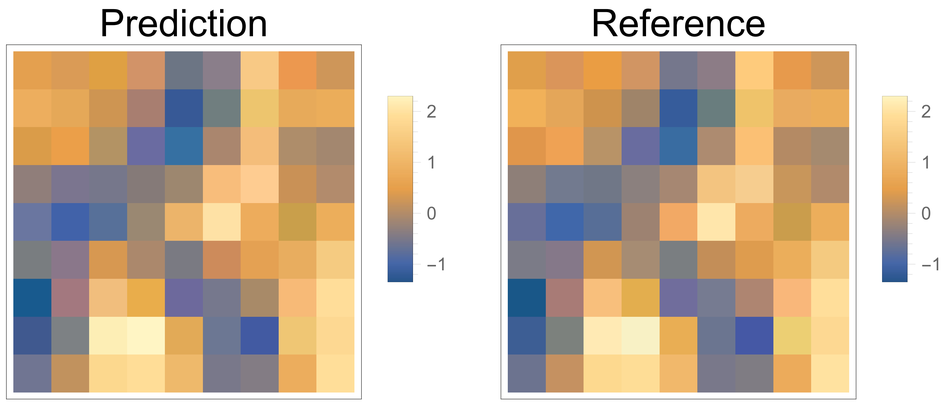
- Predictions:

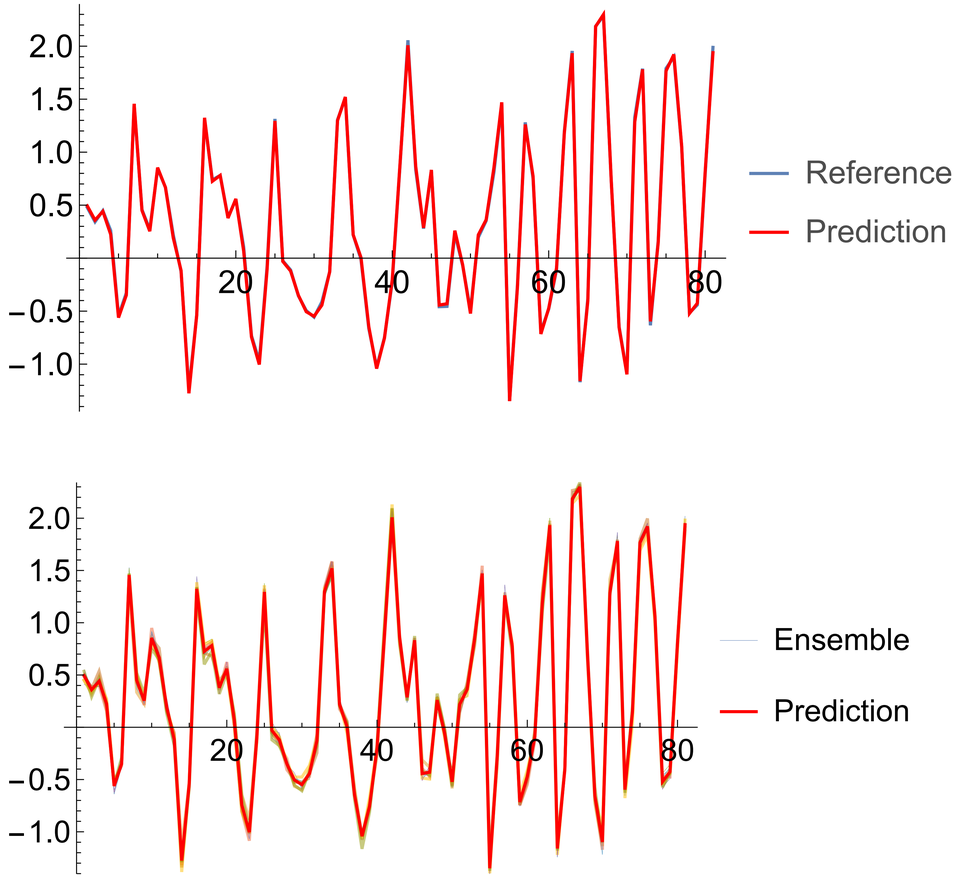




Statistical loss

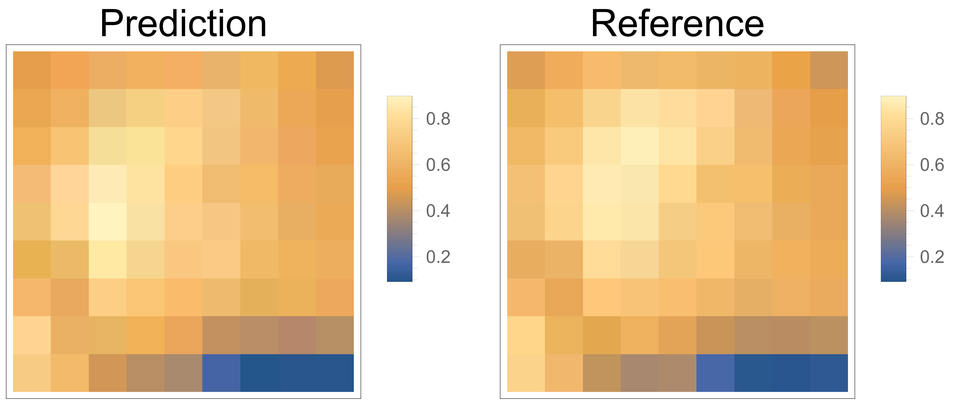
- Predictions:

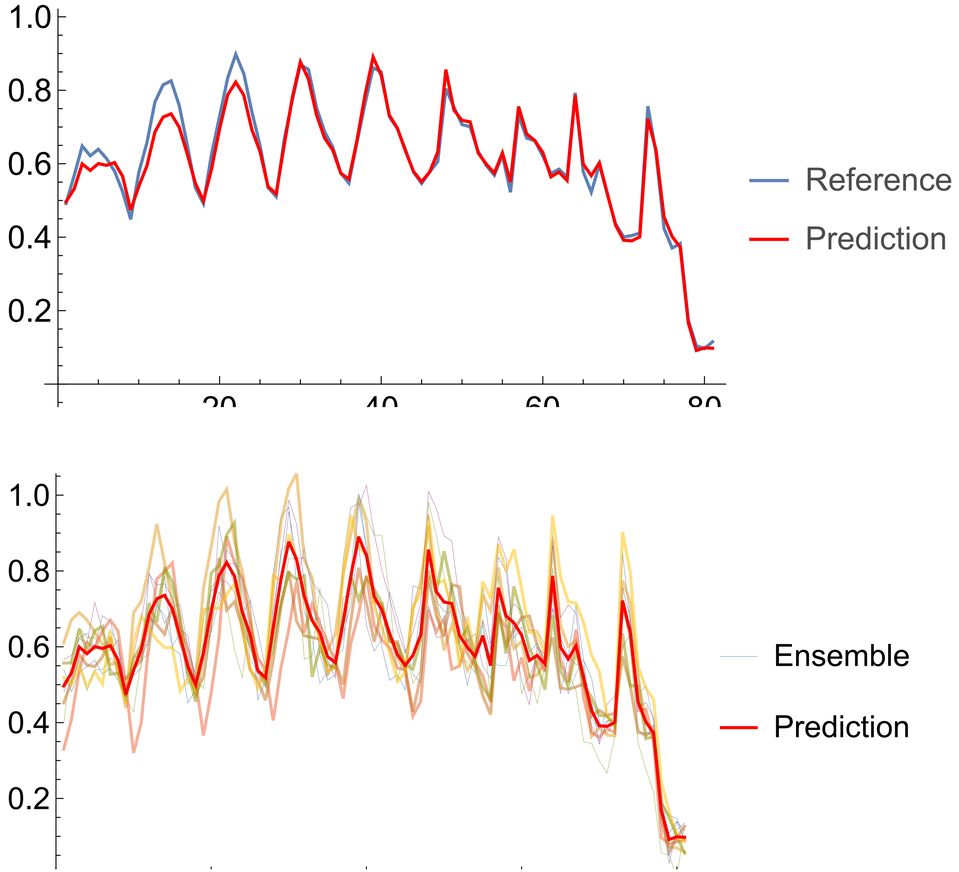




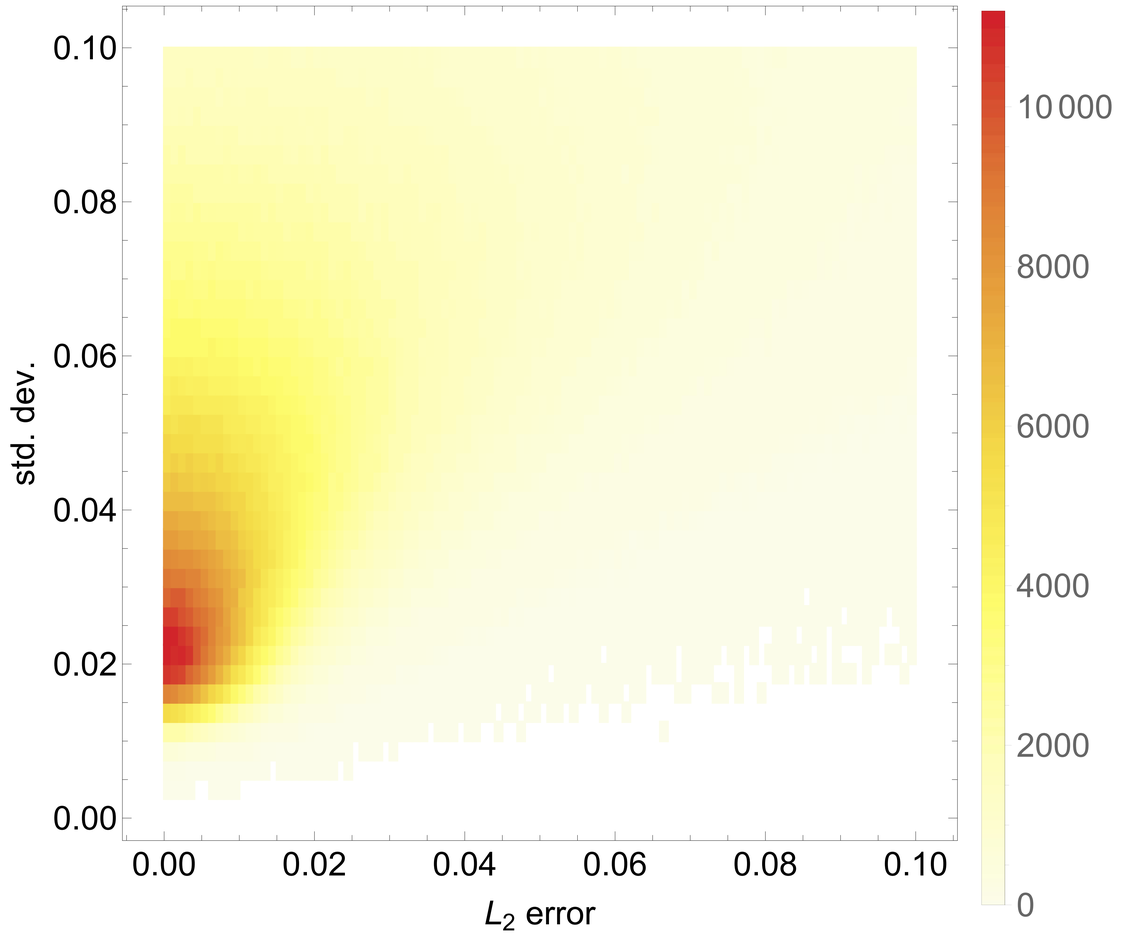
Statistical loss

- 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 $p_\theta\left(y|x,\alpha\right)$?

AtmoRep: in-context learning

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

AtmoRep: in-context learning

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

$$p_\theta\left(y|x,\alpha,\tau\right)$$

positional encoding
for output

AtmoRep: in-context learning

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

$$p_\theta\left(y|x,\alpha,\tau\right)$$

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

AtmoRep: in-context learning

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

$$p_\theta\left(y|x,\alpha,\tau\right)$$

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

- Does this make the network a scientific model?

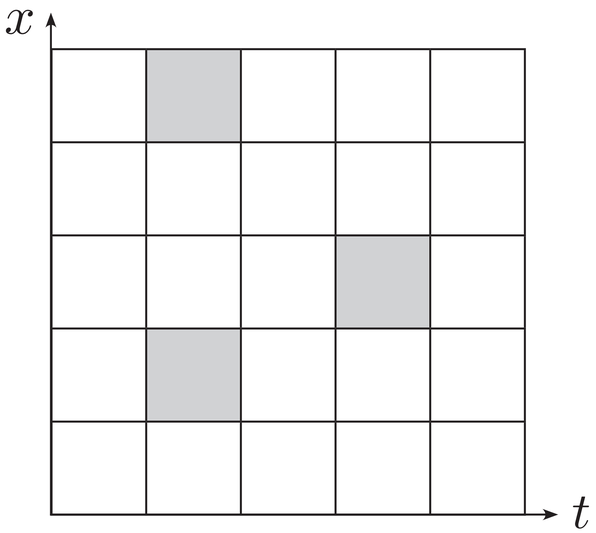
“I confess even to this day that I still don’t understand quantum mechanics, and I’m not even sure I really know how to use it all that well. And a lot of this has to do with the fact that I still don’t understand it.”

John Clauser, 2002

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

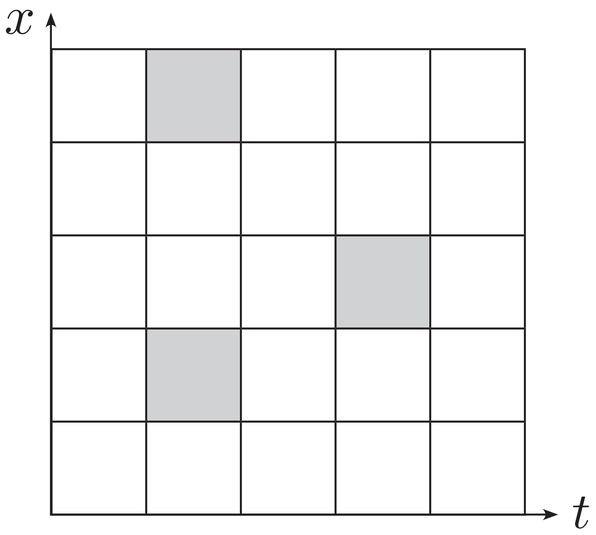
Zero shot forecasting

BERT

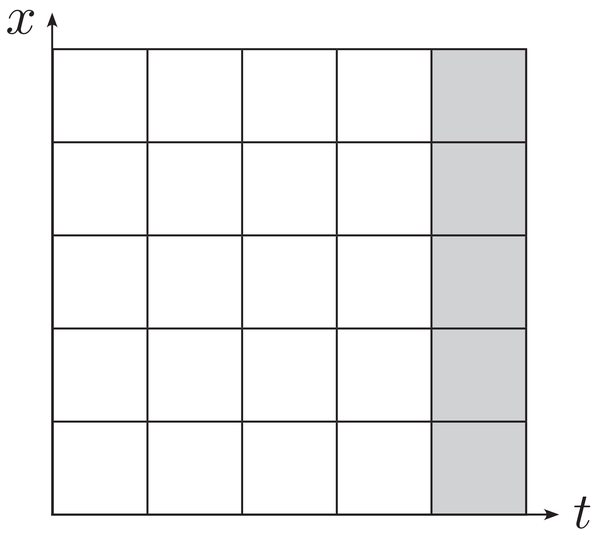


Zero shot forecasting

BERT

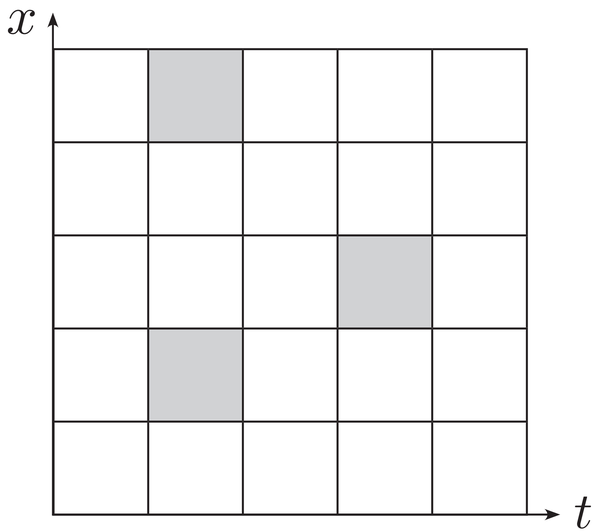


BERT-Forecast

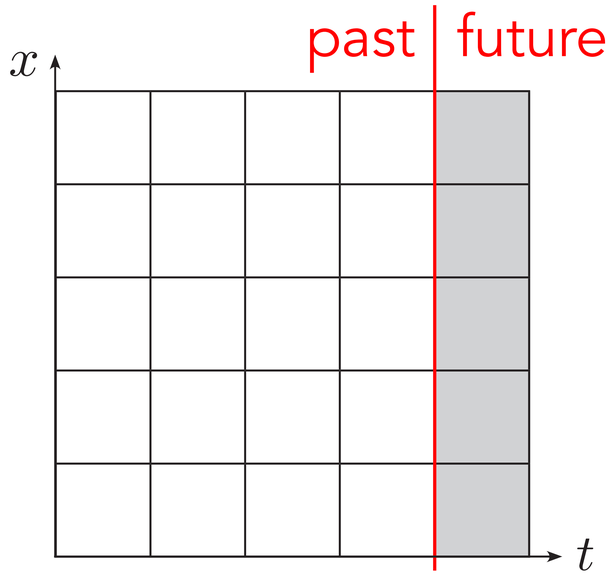


Zero shot forecasting

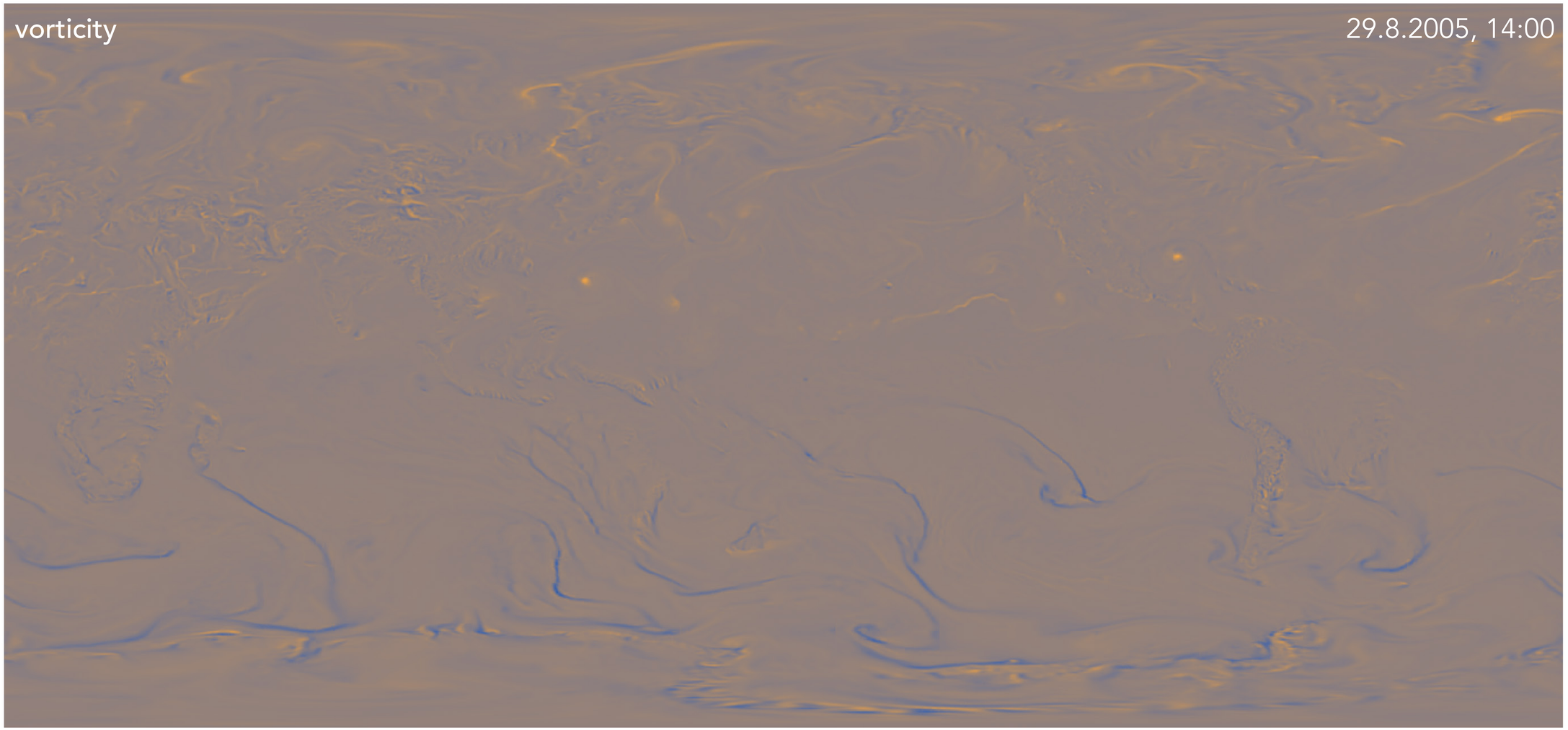
BERT



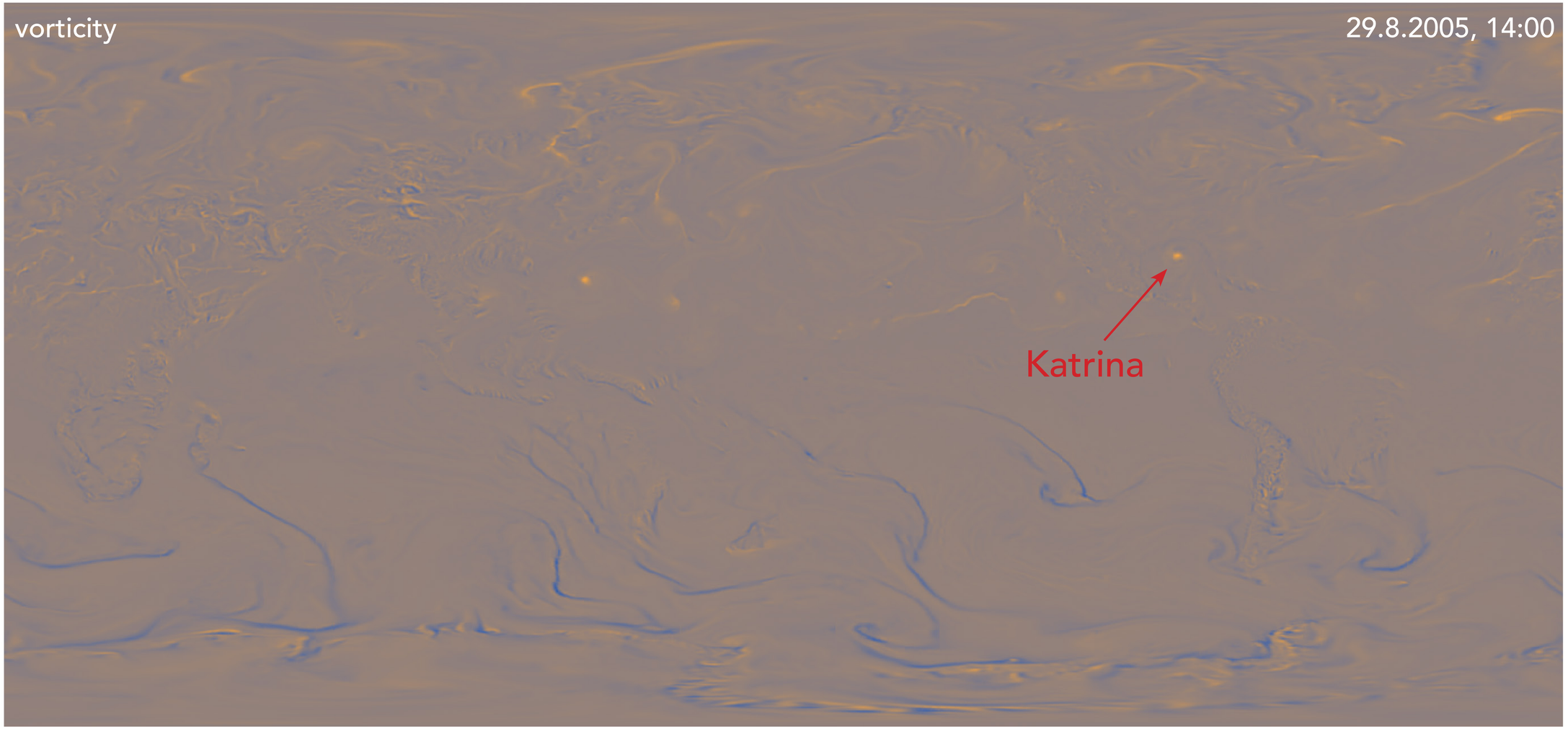
BERT-Forecast



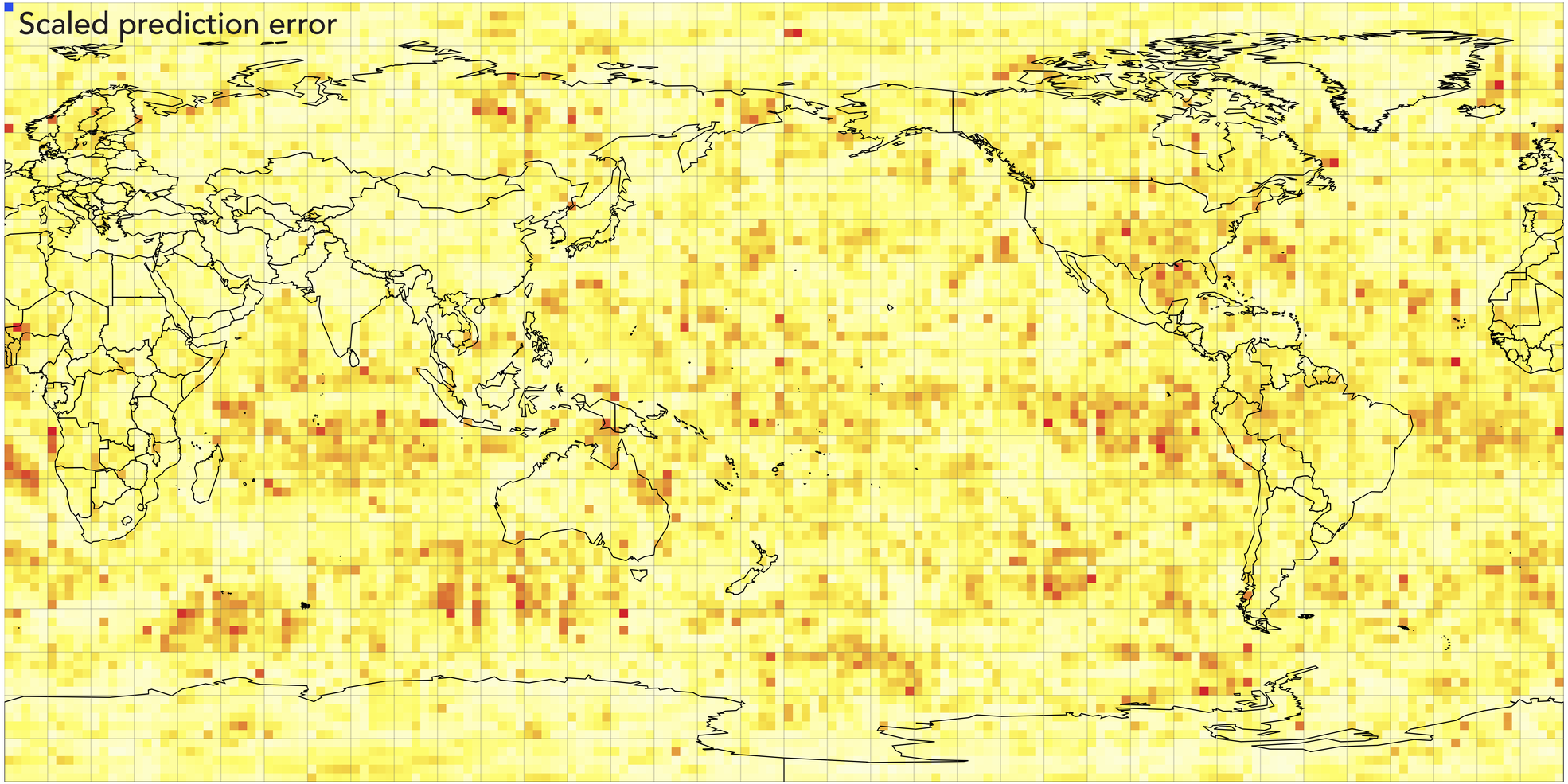
Zero shot forecasting



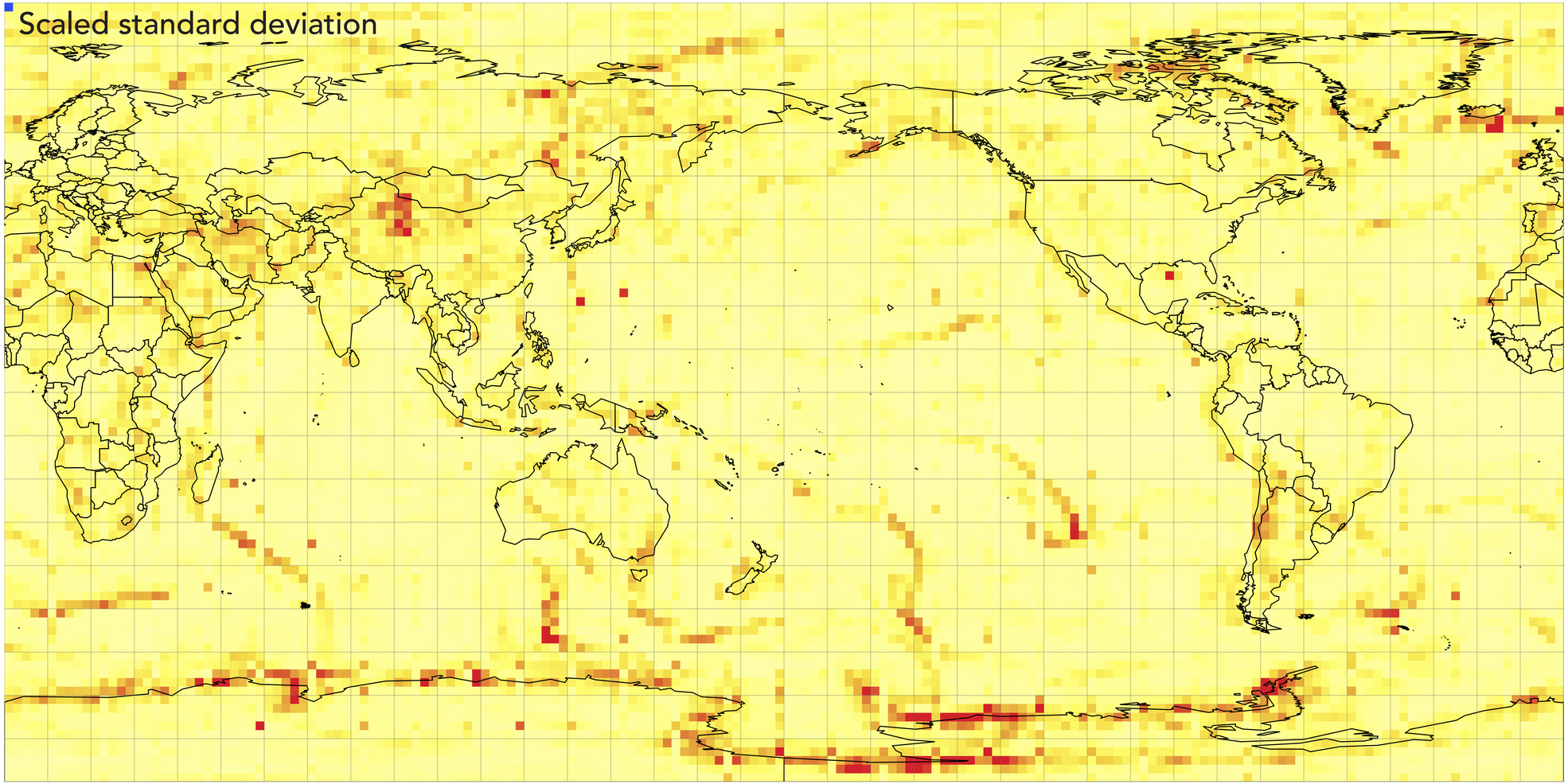
Zero shot forecasting



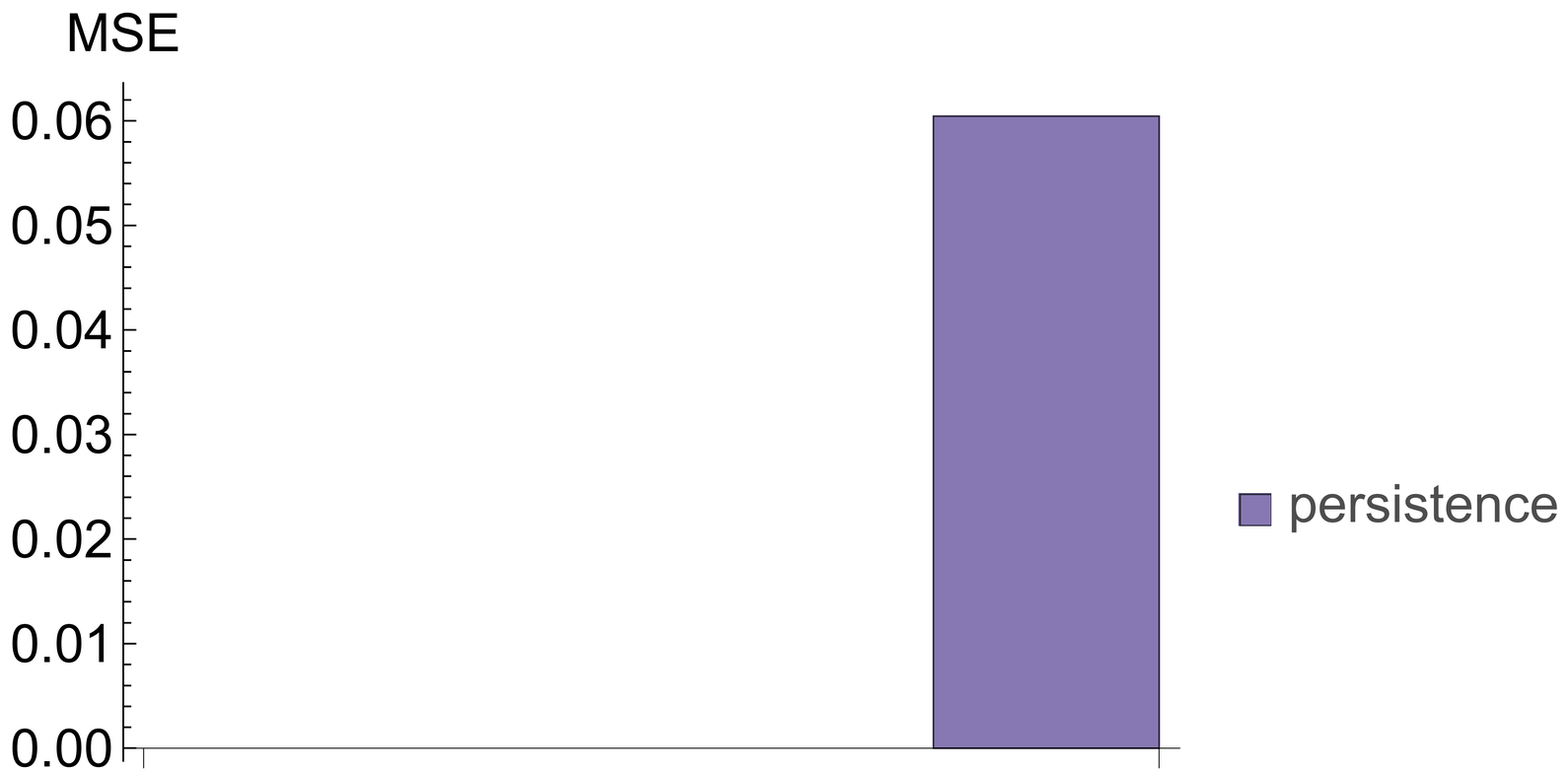
Zero shot forecasting



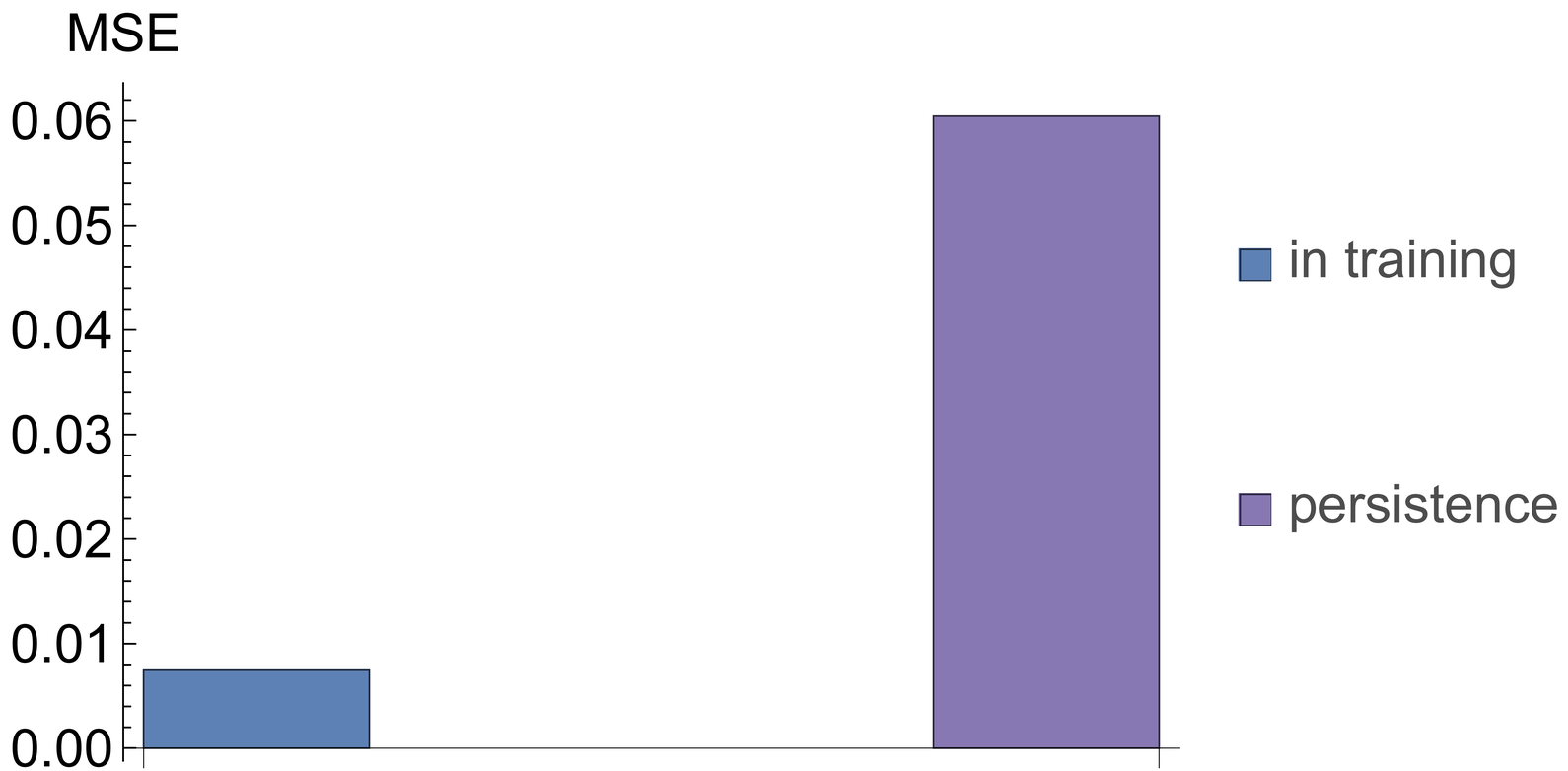
Zero shot forecasting



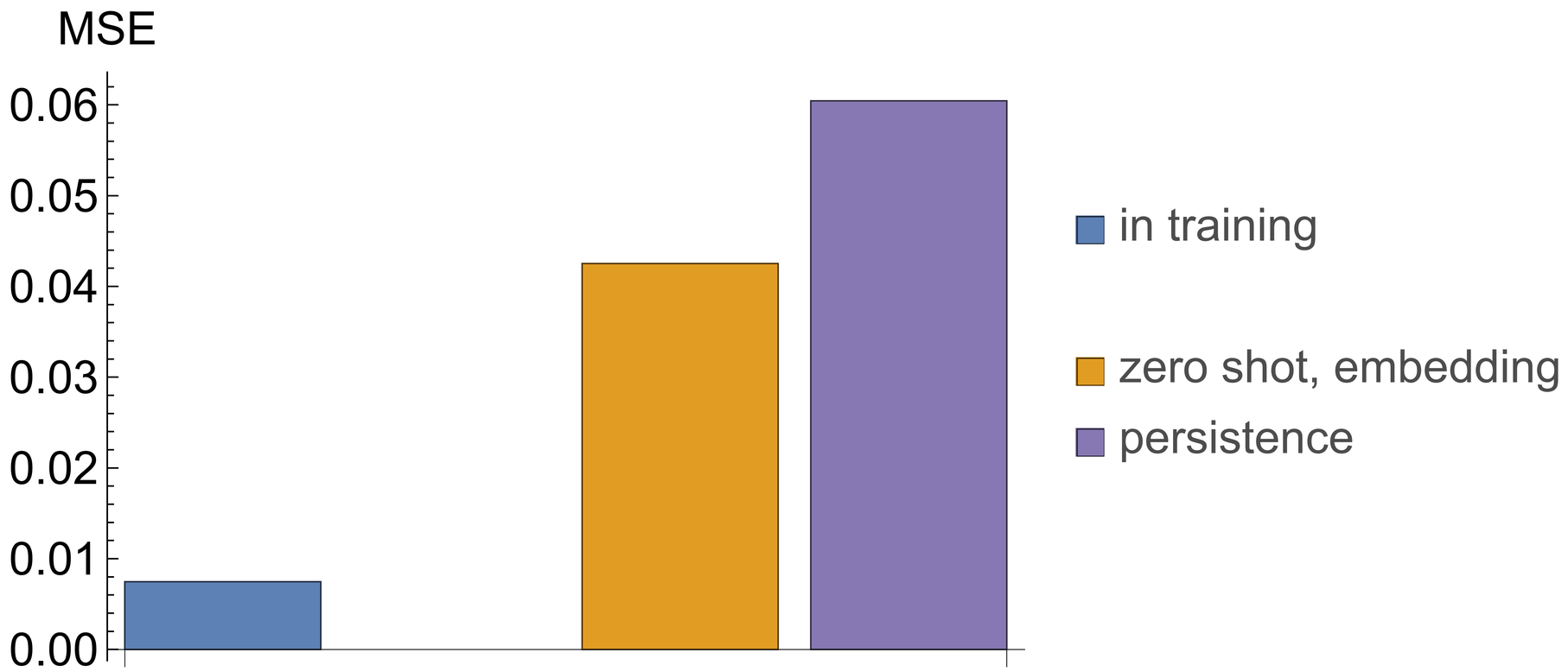
Zero shot forecasting



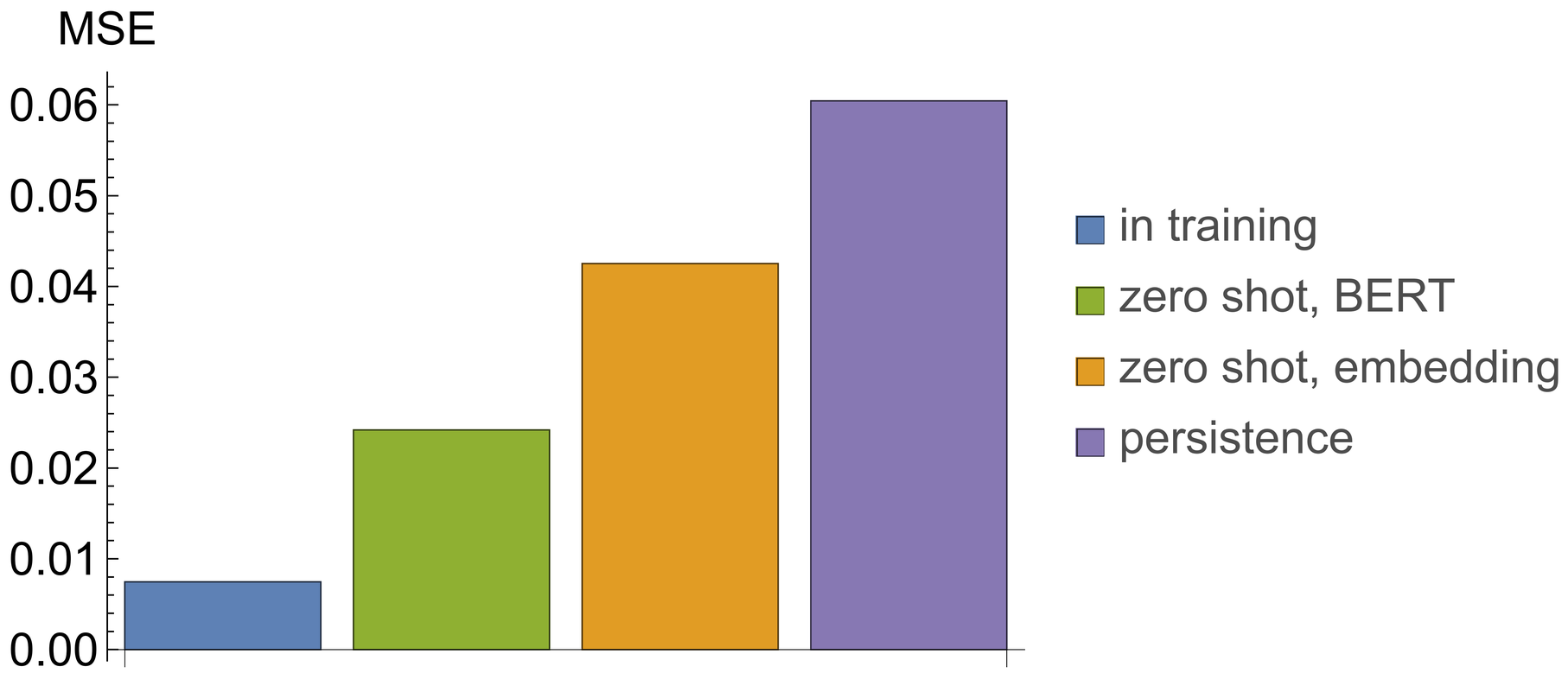
Zero shot forecasting



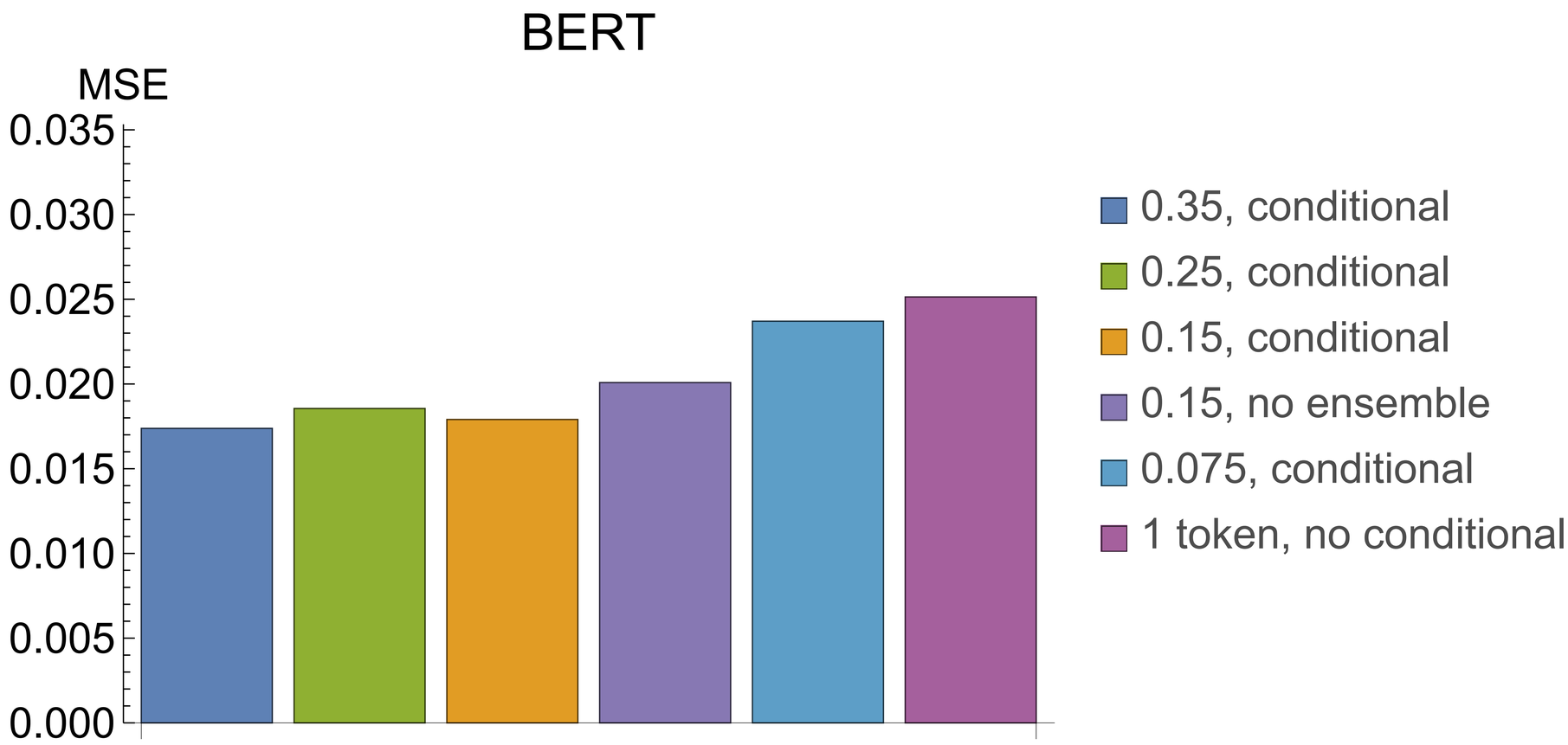
Zero shot forecasting



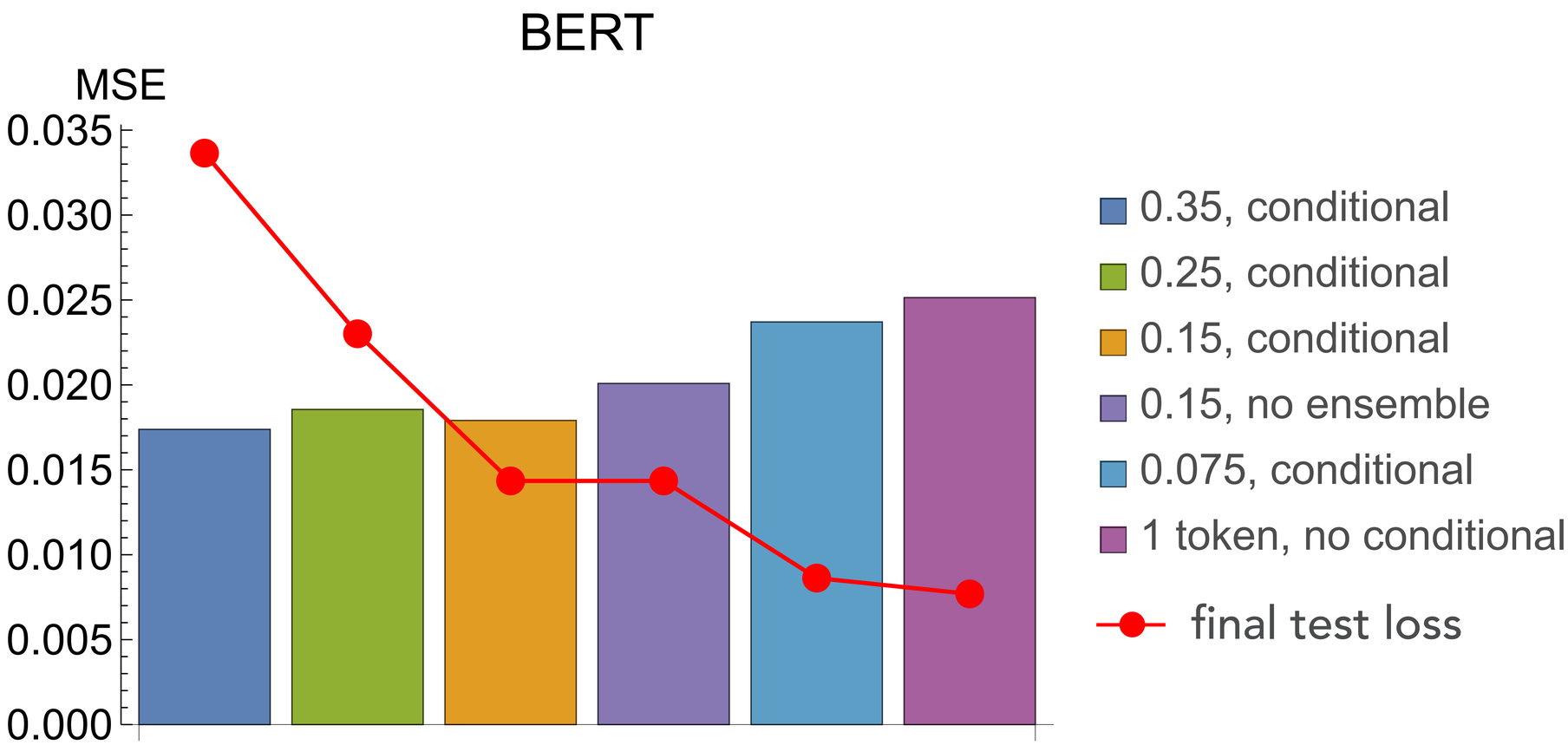
Zero shot forecasting



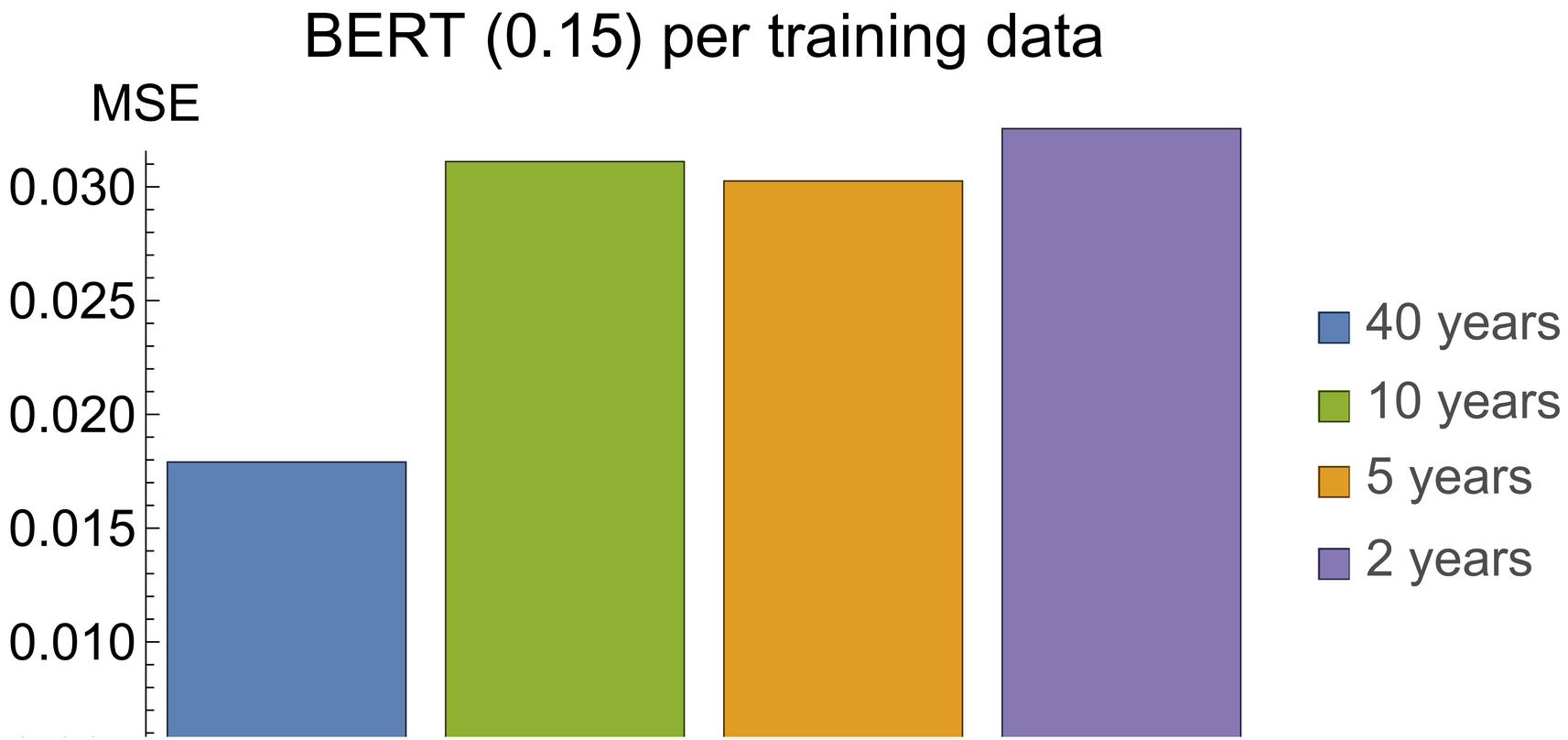
Zero shot forecasting



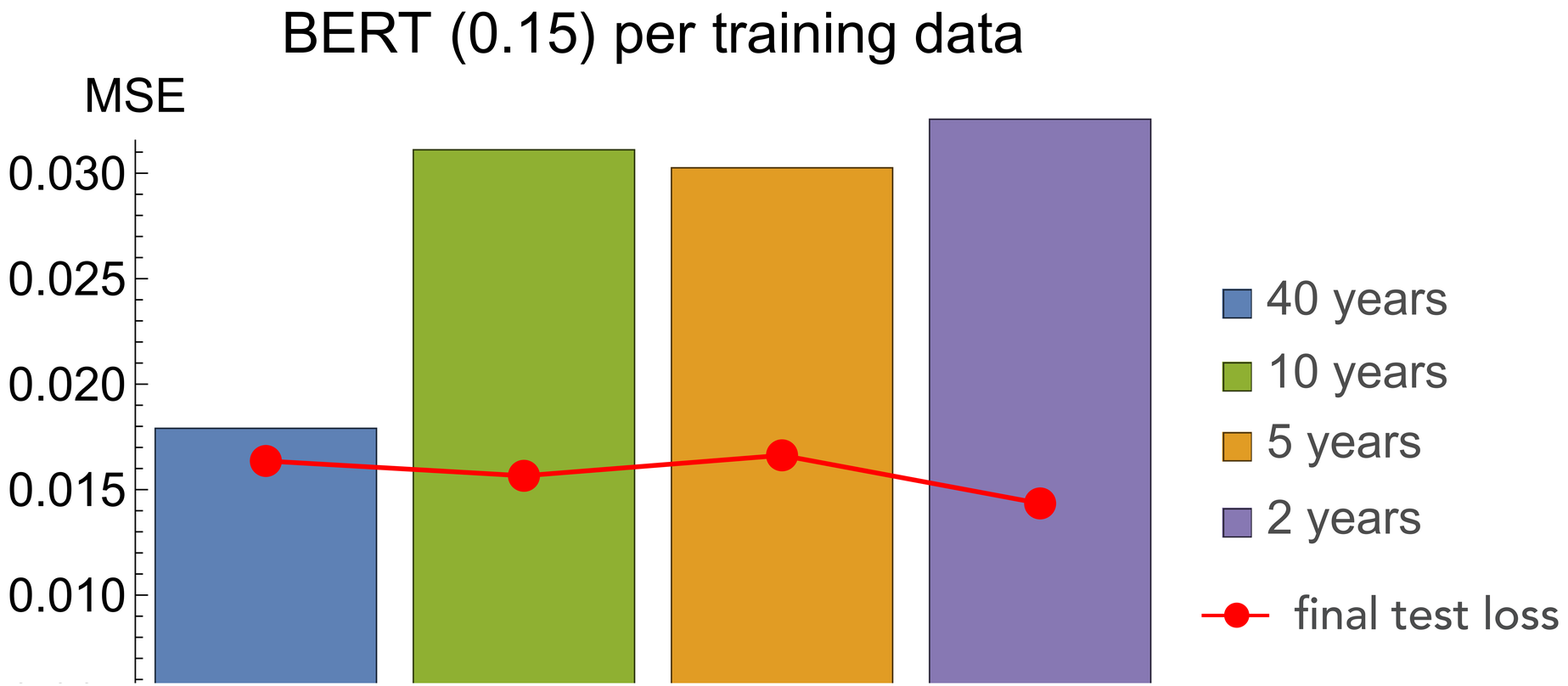
Zero shot forecasting



Zero shot forecasting

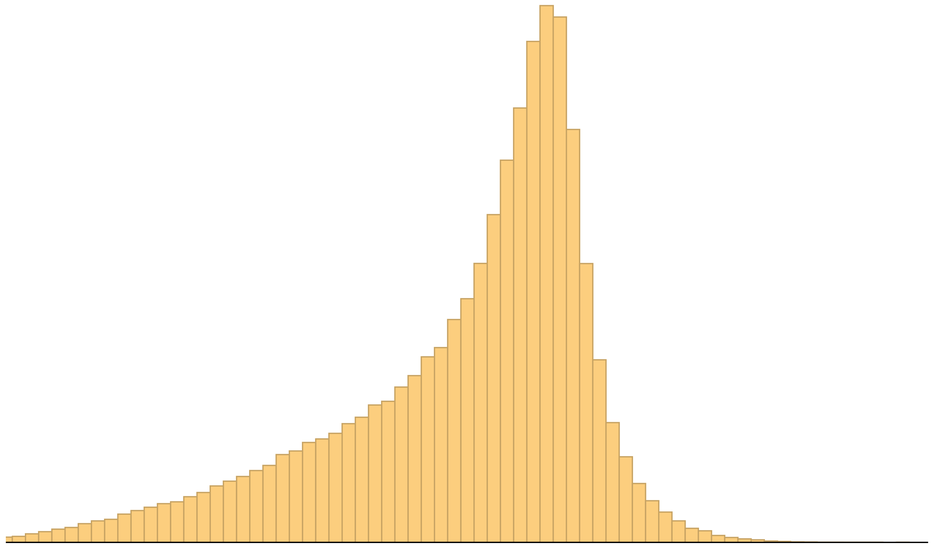


Zero shot forecasting



Zero shot forecasting

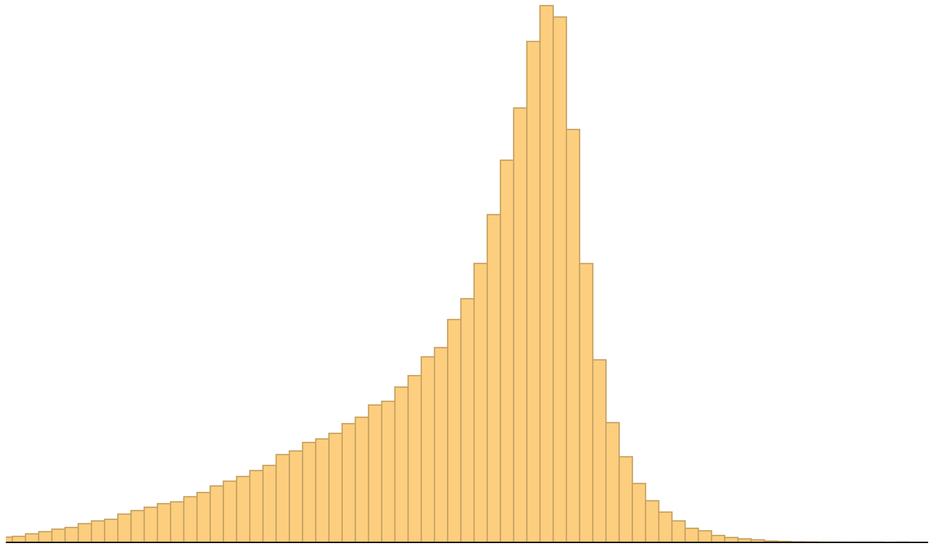
Cape town: historgram of vorticity



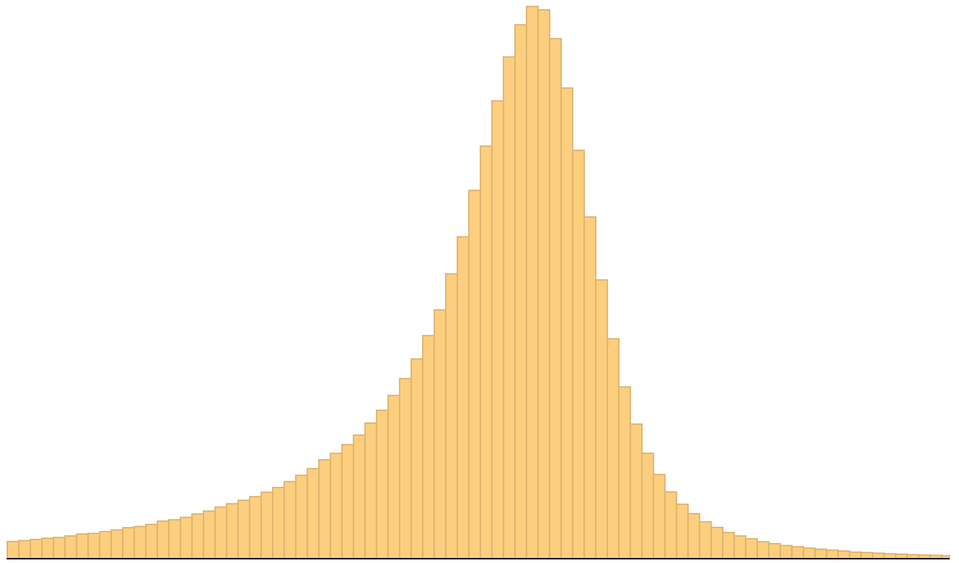
ERA5

Zero shot forecasting

Cape town: historgram of vorticity



ERA5



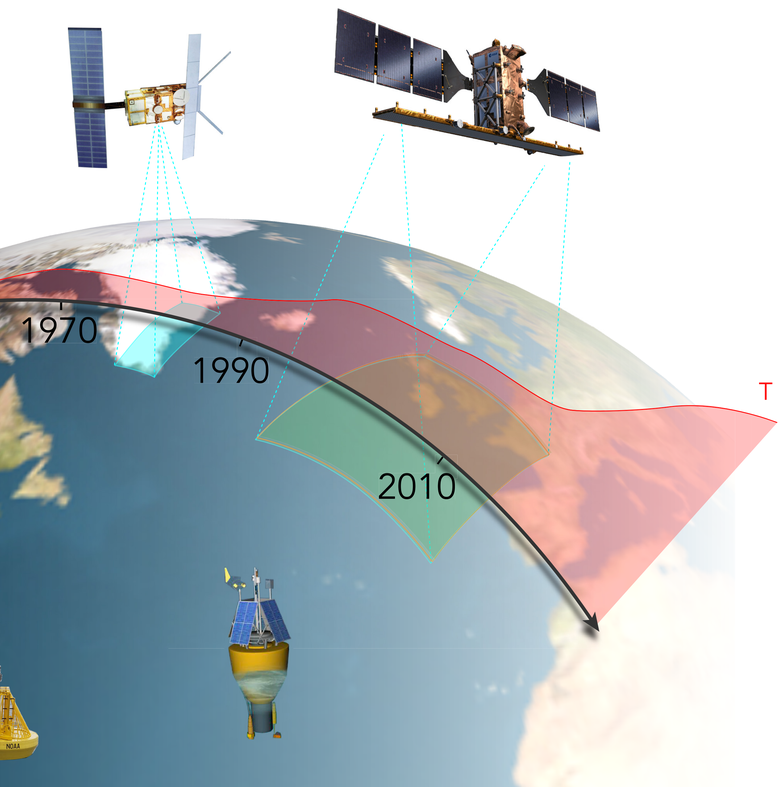
predictions

AtmoRep: longer term objectives

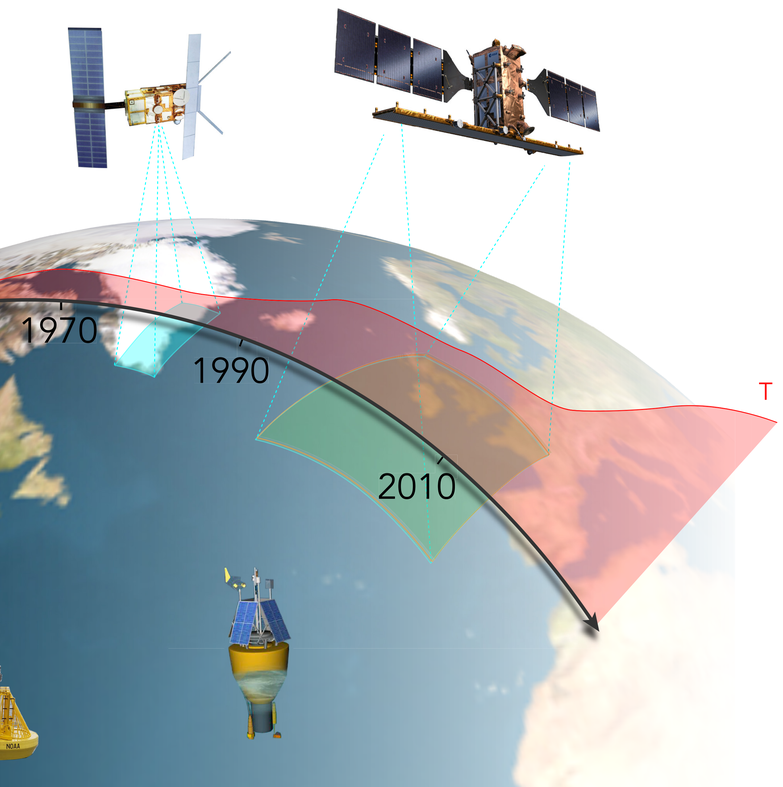
- Weather forecasting
- Climate projections
- Coupled Earth system
- Scientific model
- Training/fine-tuning on direct observational data

Current / next steps

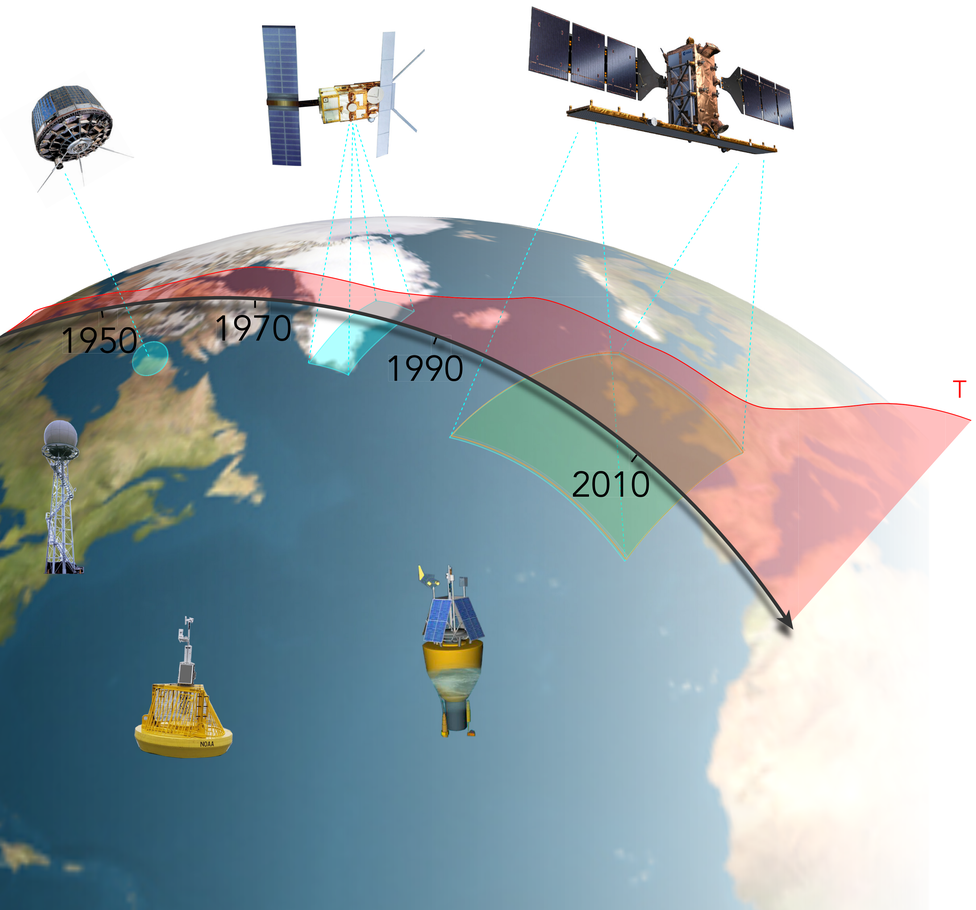
- Complete representation learning model
  - Scale data and network size
  - Different training tasks and protocols



Current / next steps

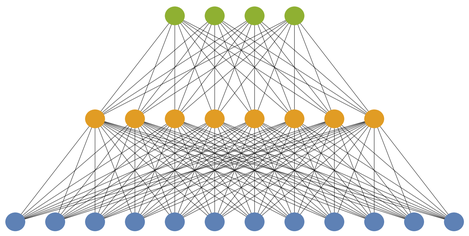


- Complete representation learning model
- Downstream applications
  - Weather forecasting
  - Downscaling
  - Model correction
  - ...



AtmoRep

Large scale representation learning of atmospheric dynamics

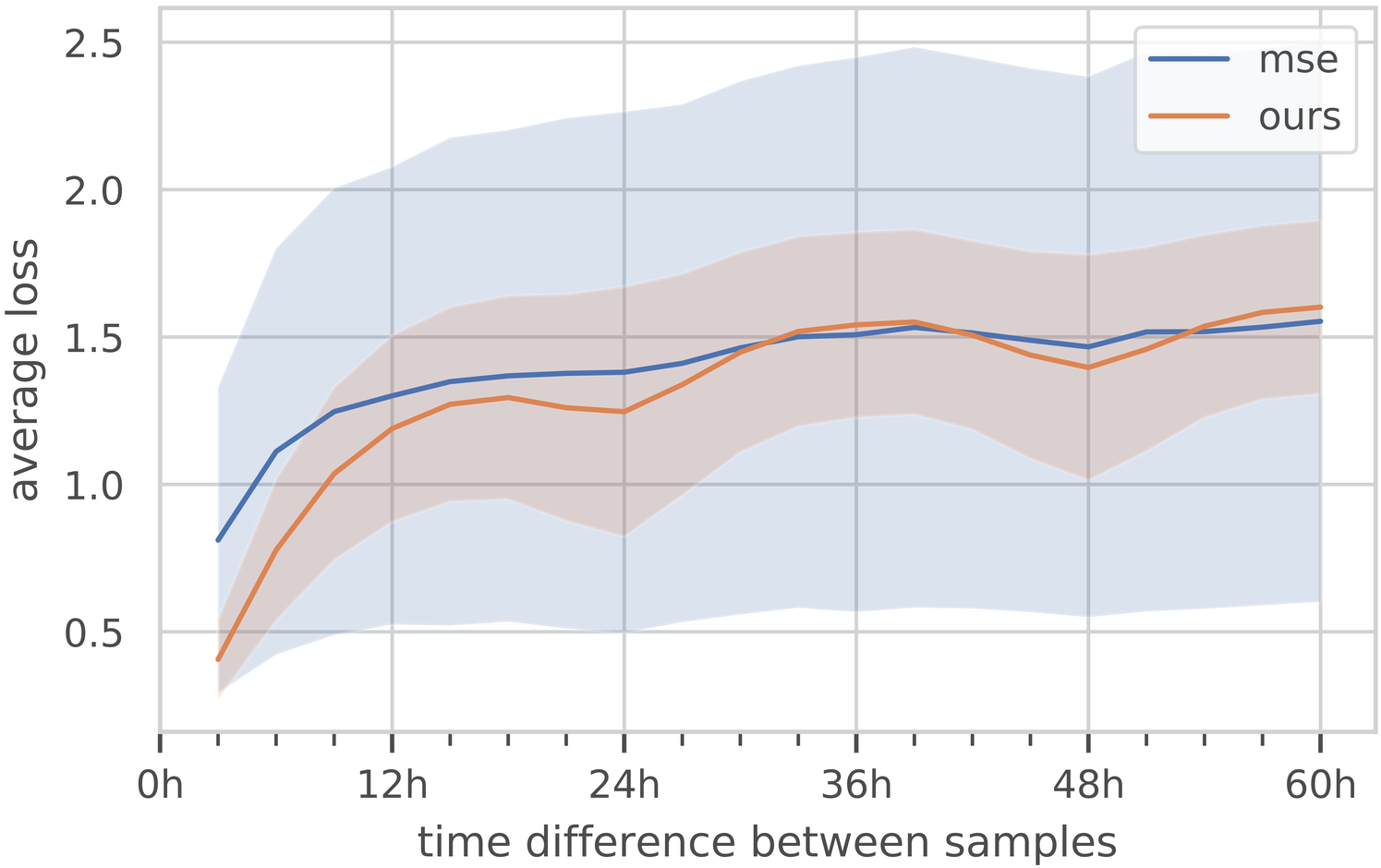


address climate change

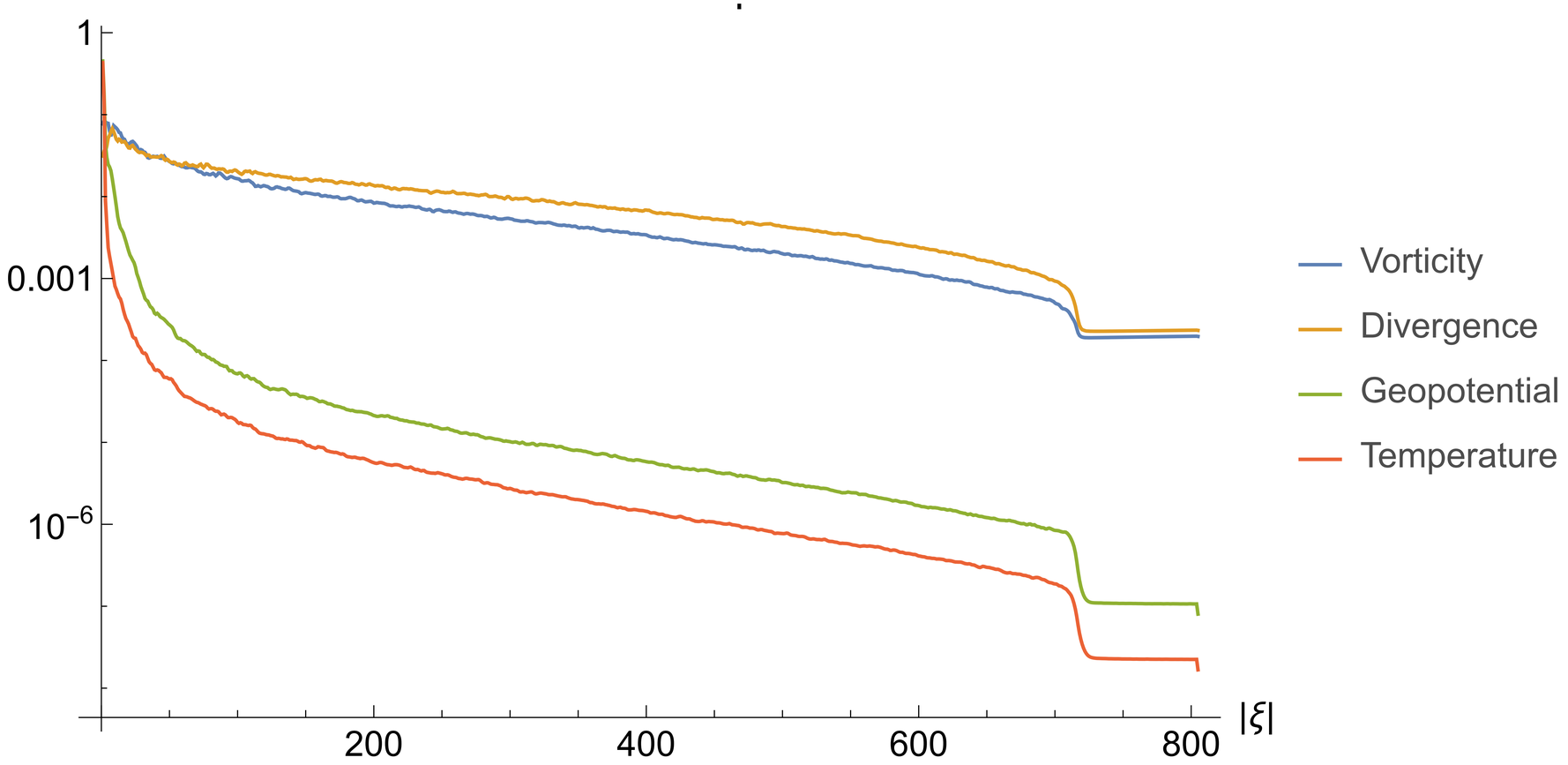
large scale machine learning

scientific insight

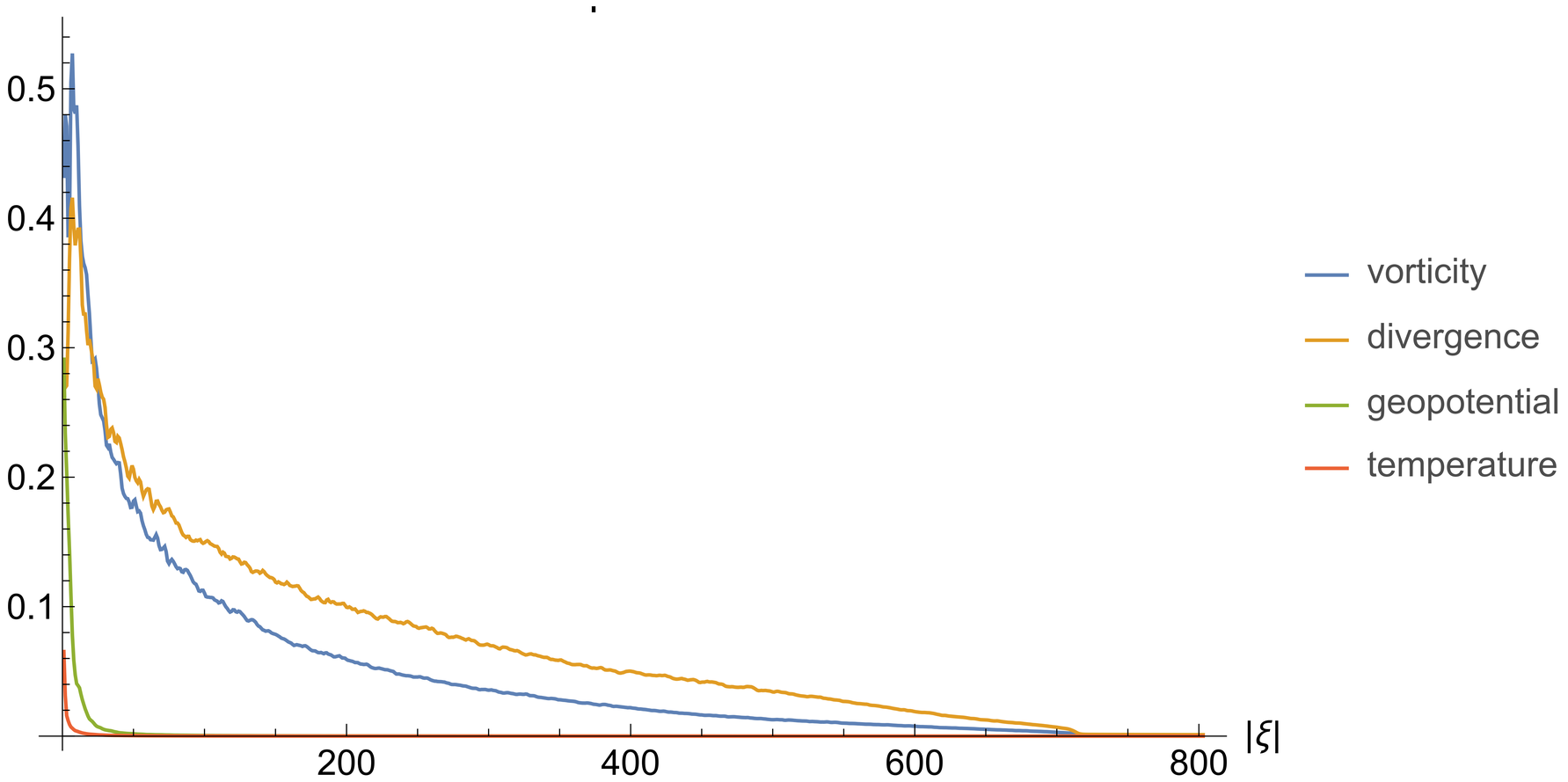
AtmoDist: evaluation



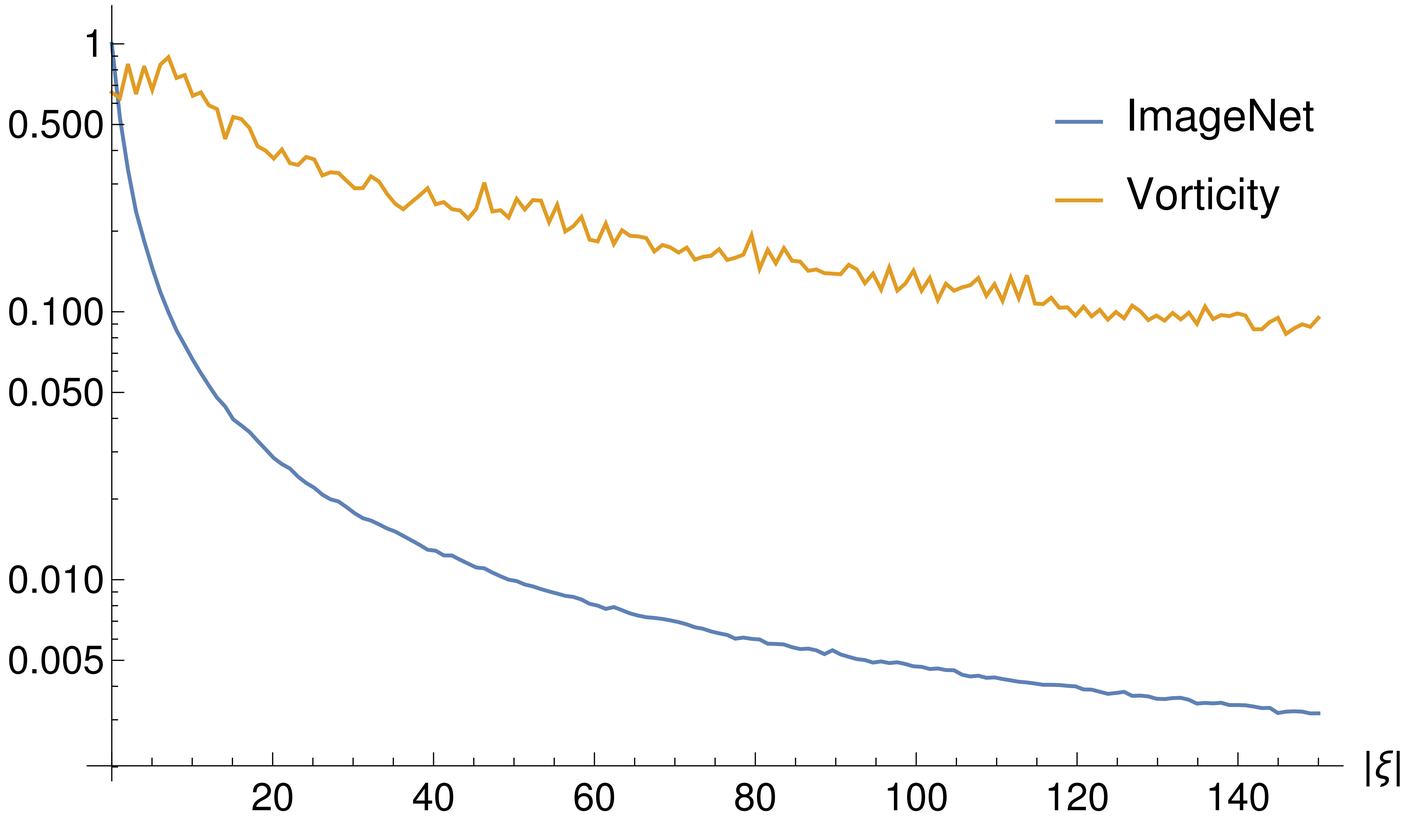
AtmoRep data



AtmoRep data



ERA5 versus ImageNet



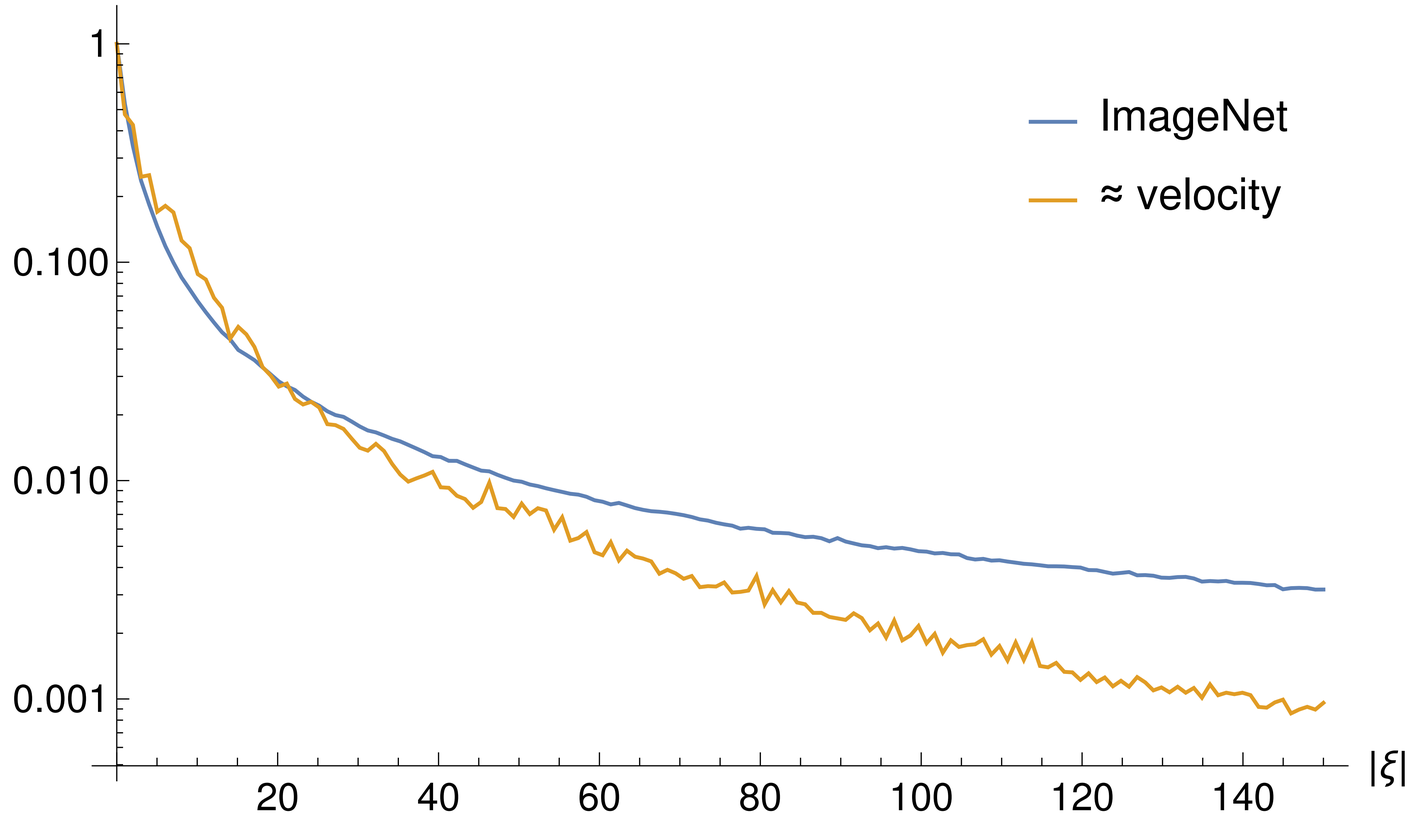
ERA5 versus ImageNet

stream function
velocity potential

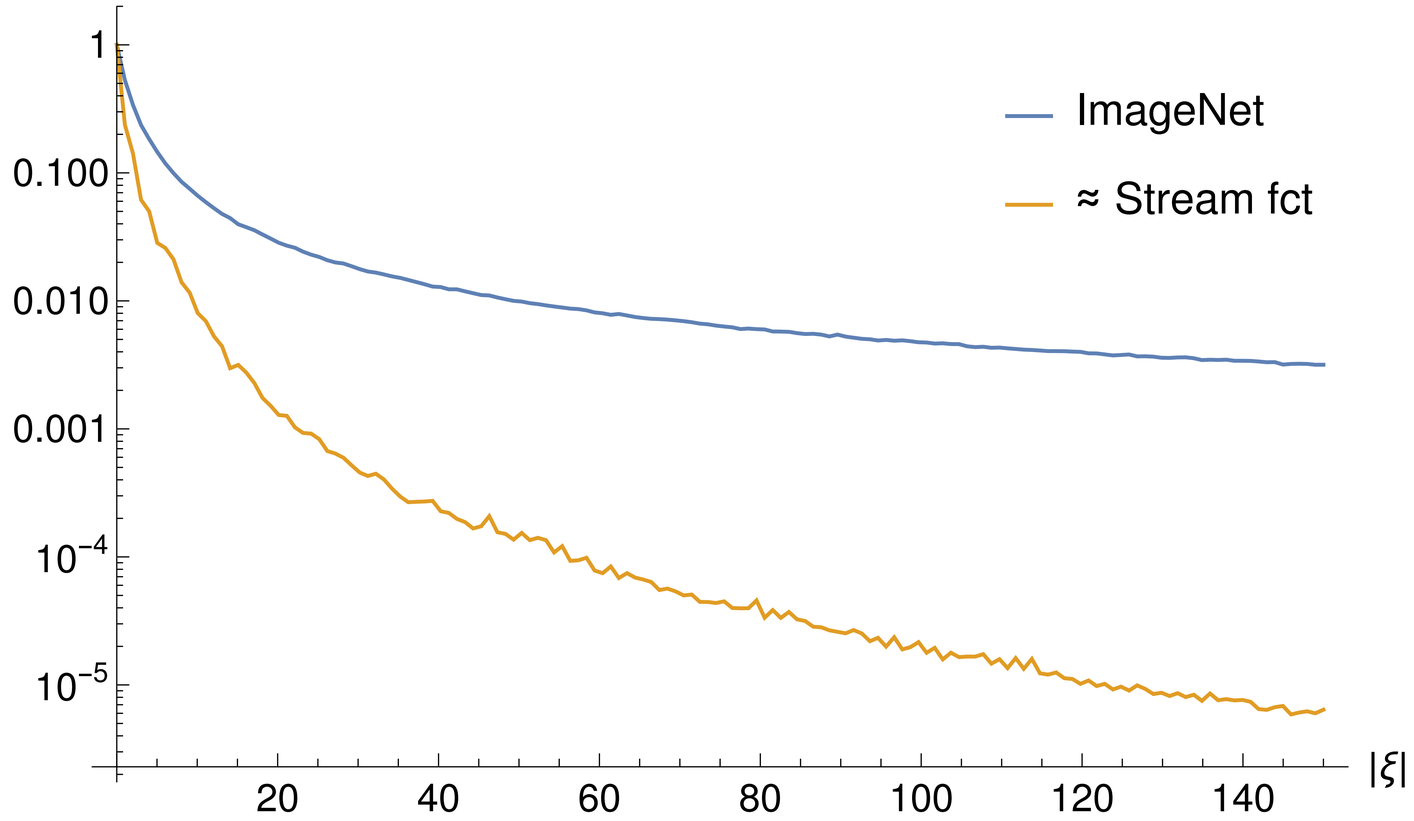
velocity
vector field

vorticity
divergence

ERA5 versus ImageNet



ERA5 versus ImageNet



ERA5 versus ImageNet

stream function
velocity potential

$\Leftrightarrow$

velocity
vector field

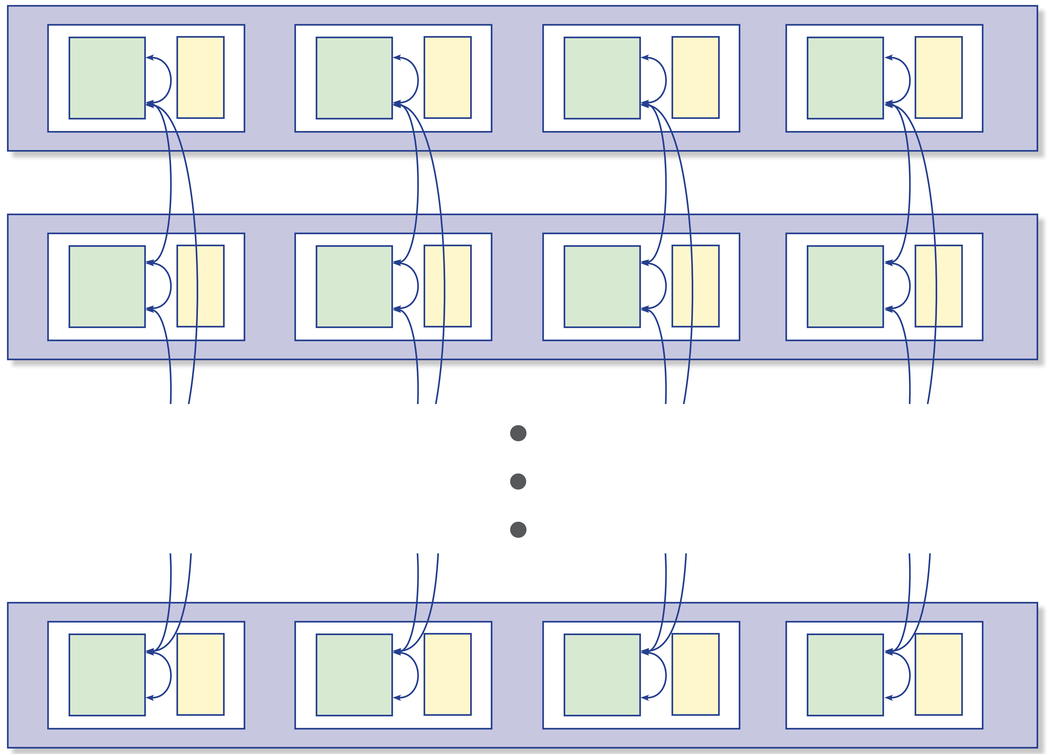
$\Leftrightarrow$

vorticity
divergence

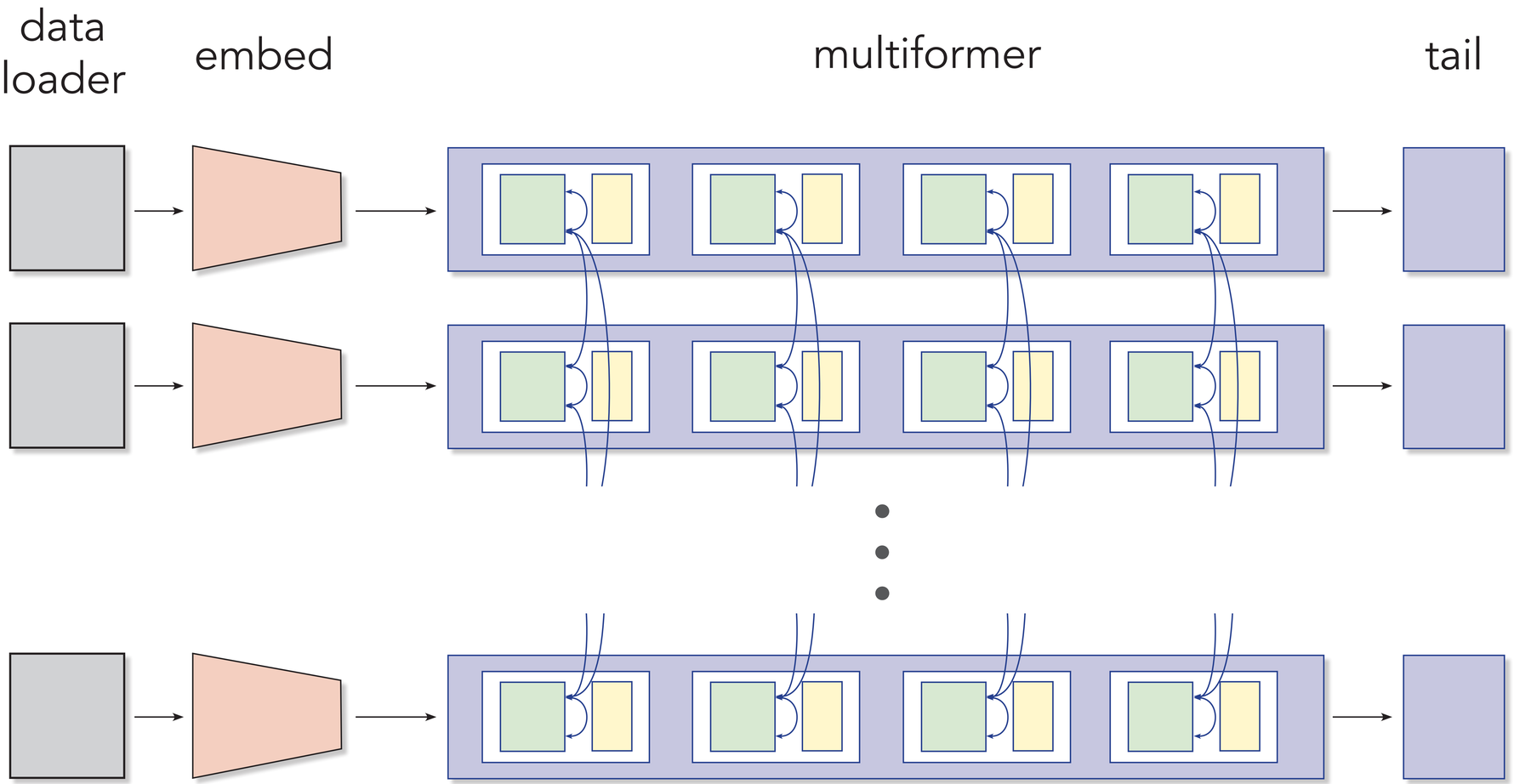
Work with velocity with norm that emphasizes small scale detail

Embedding of tokens

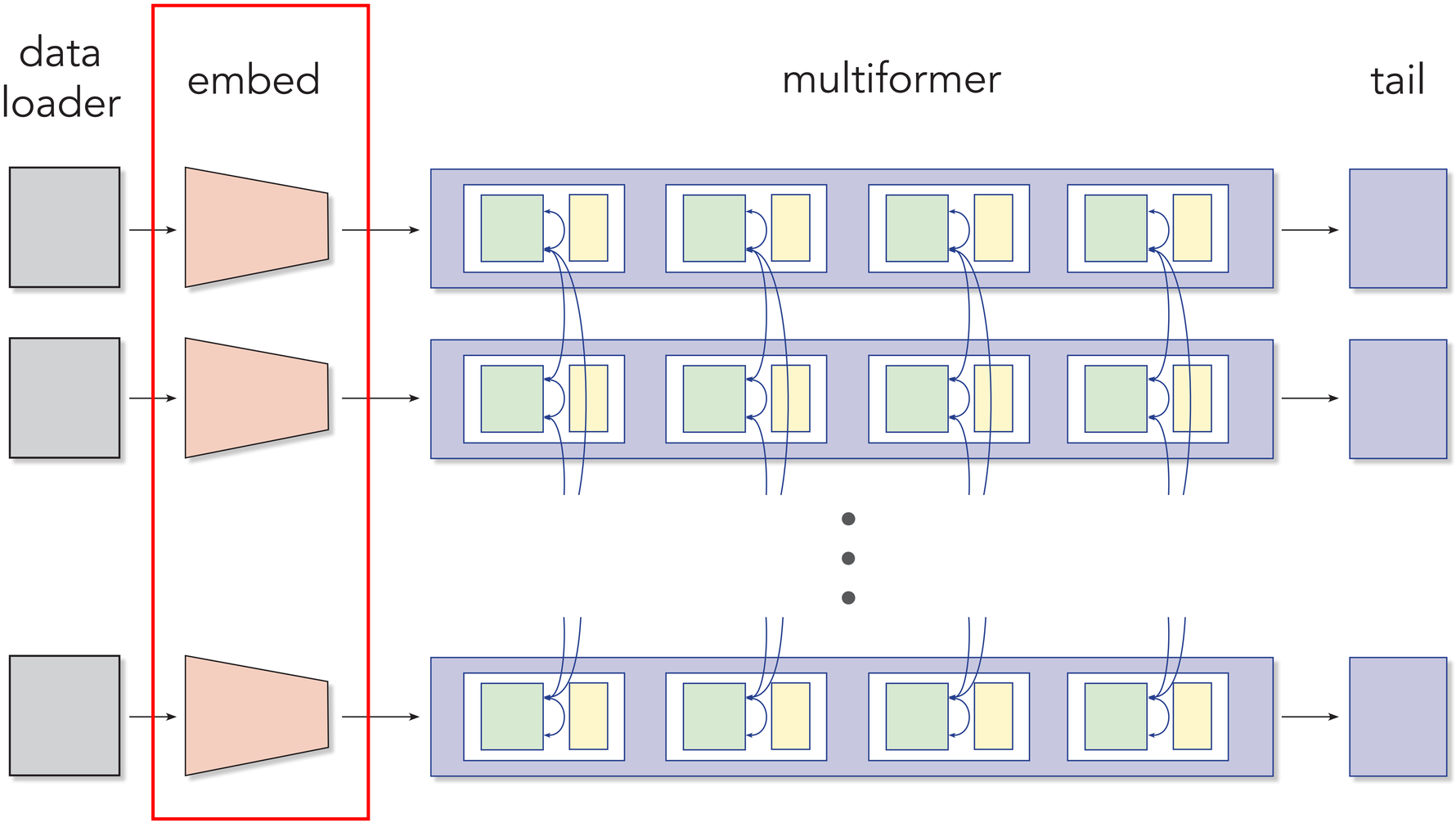
multiformer



Embedding of tokens



Embedding of tokens



Embedding network

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

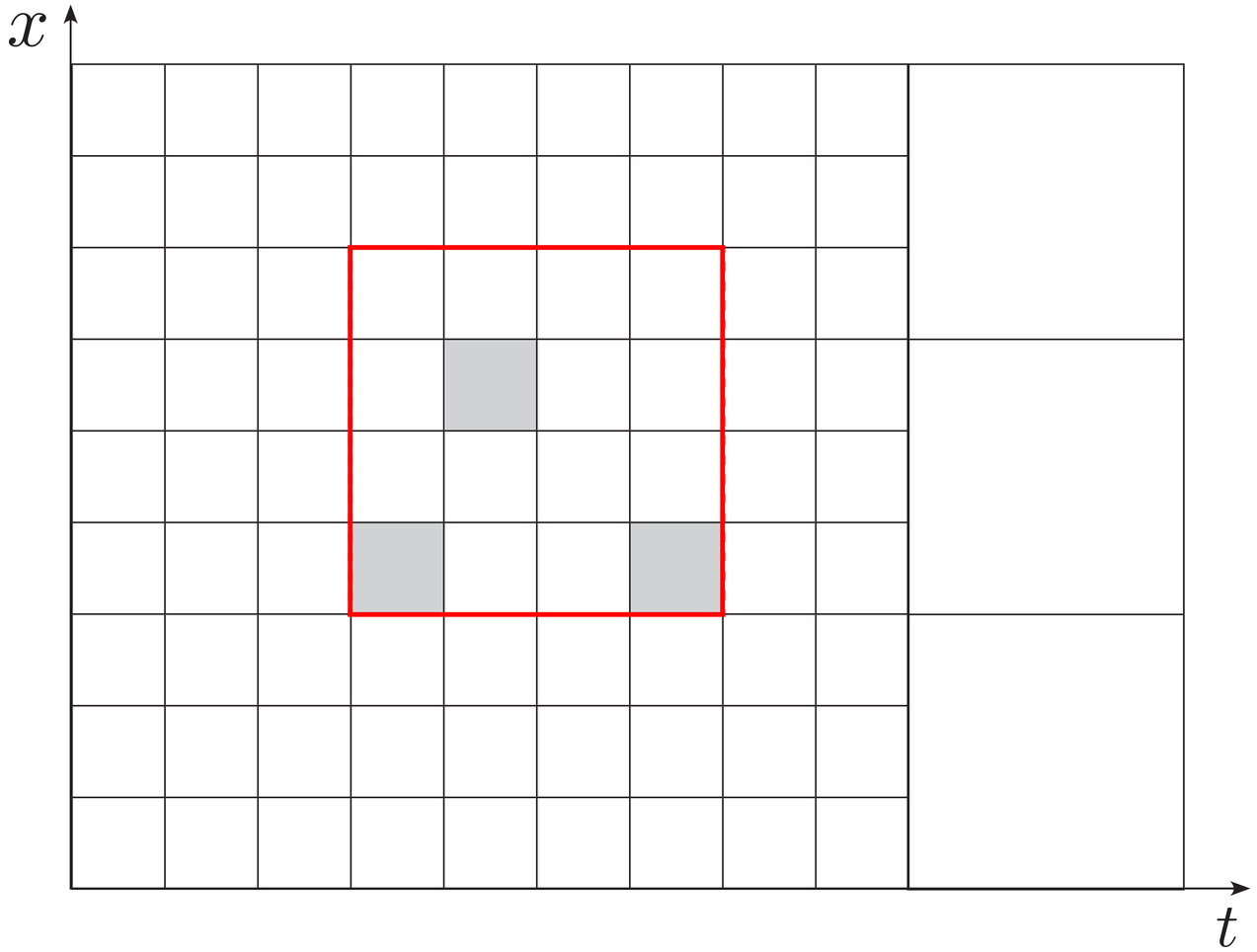
Embedding network

- 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

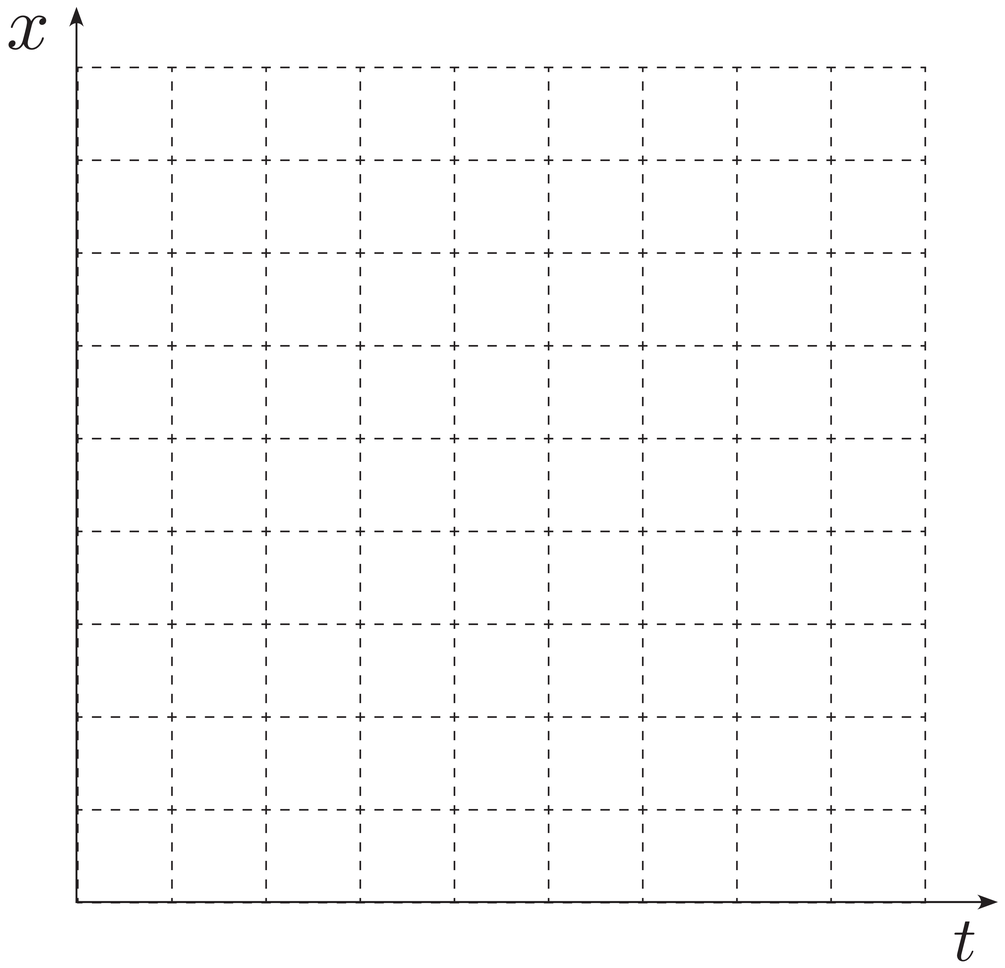
- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

Embedding of tokens

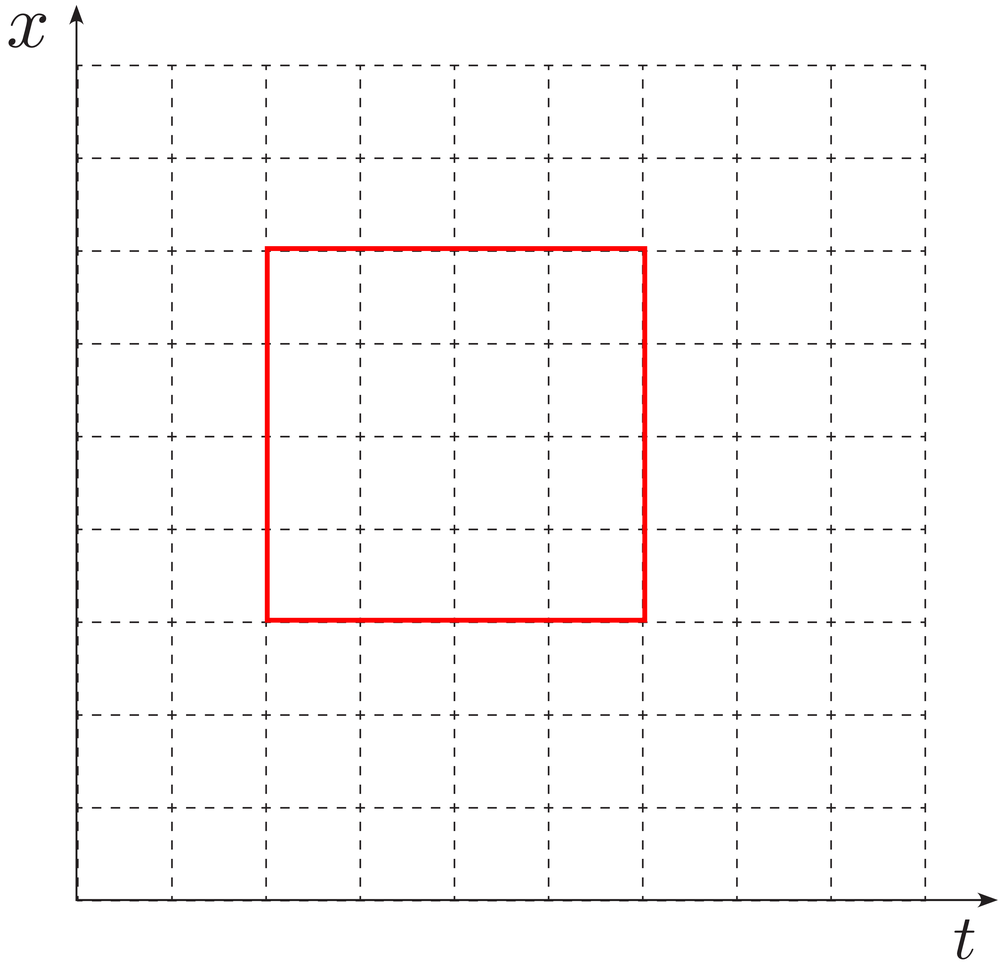
- Multiformer models longer range effects and field interactions in a rich latent space
  - › Embedding network provides rich encoding of input field
  - › Embedding network allows for multi-resolution representation per field, i.e. different token sizes

⇒ Use transformer as embedding network

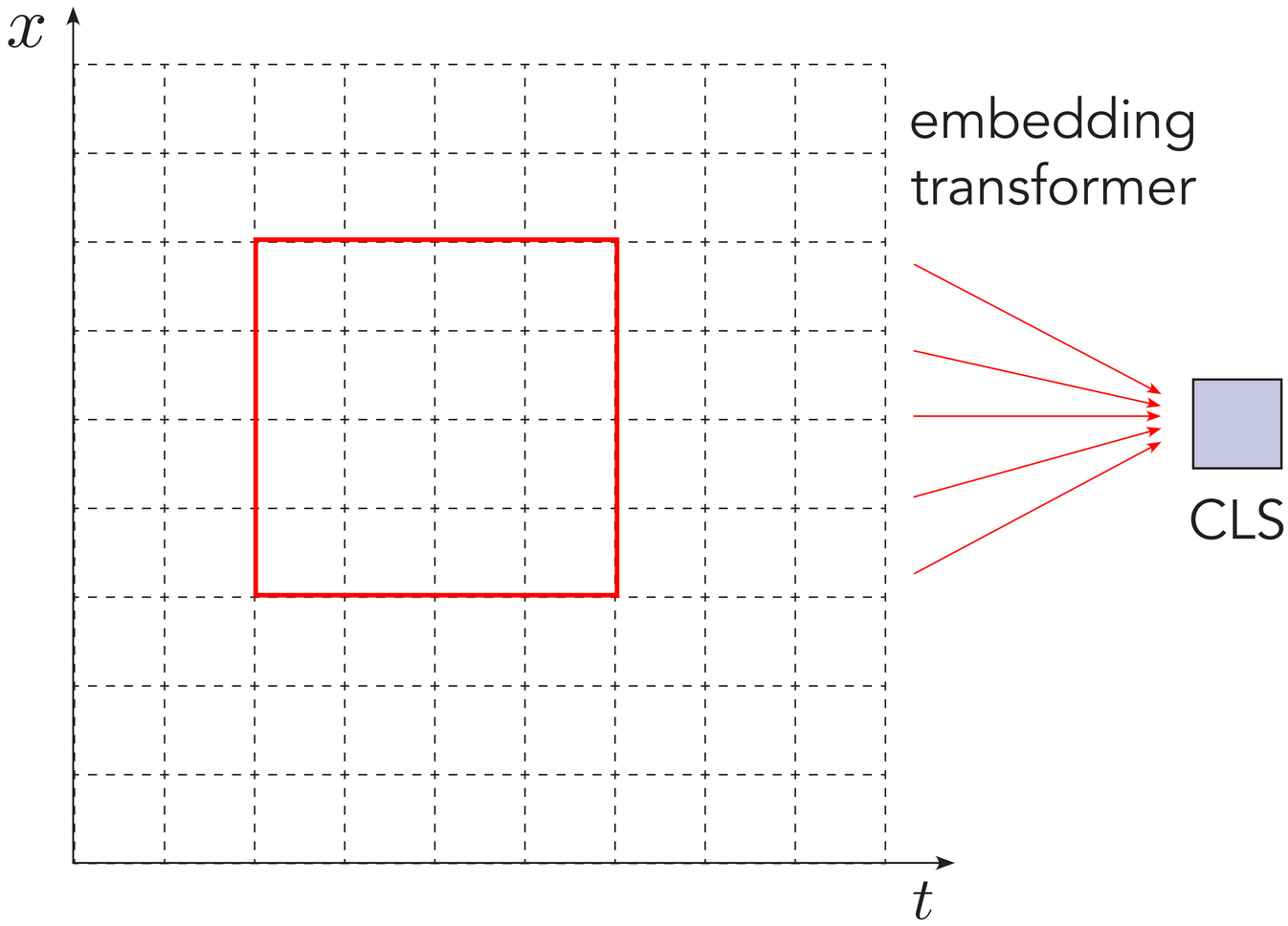
Embedding of tokens



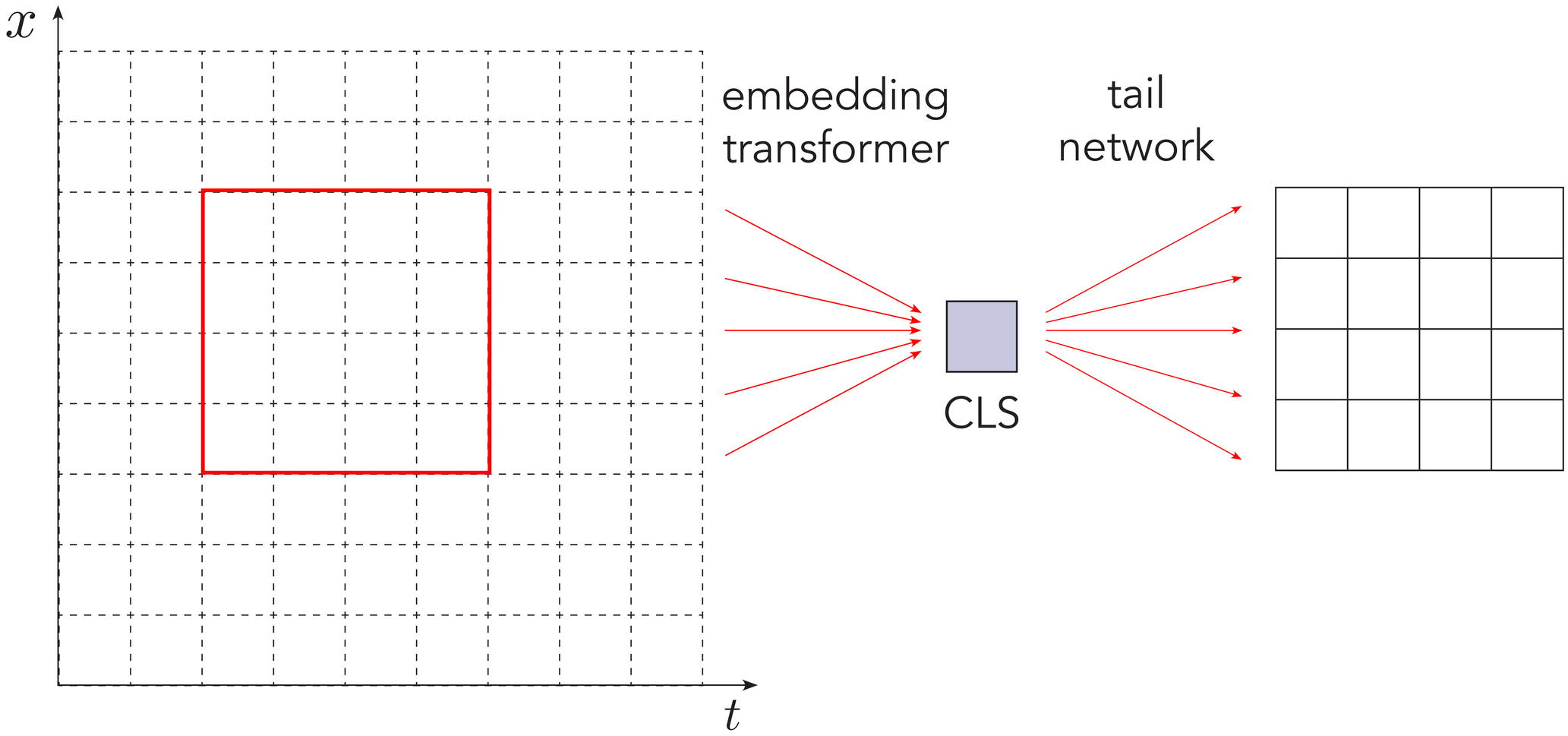
Embedding of tokens



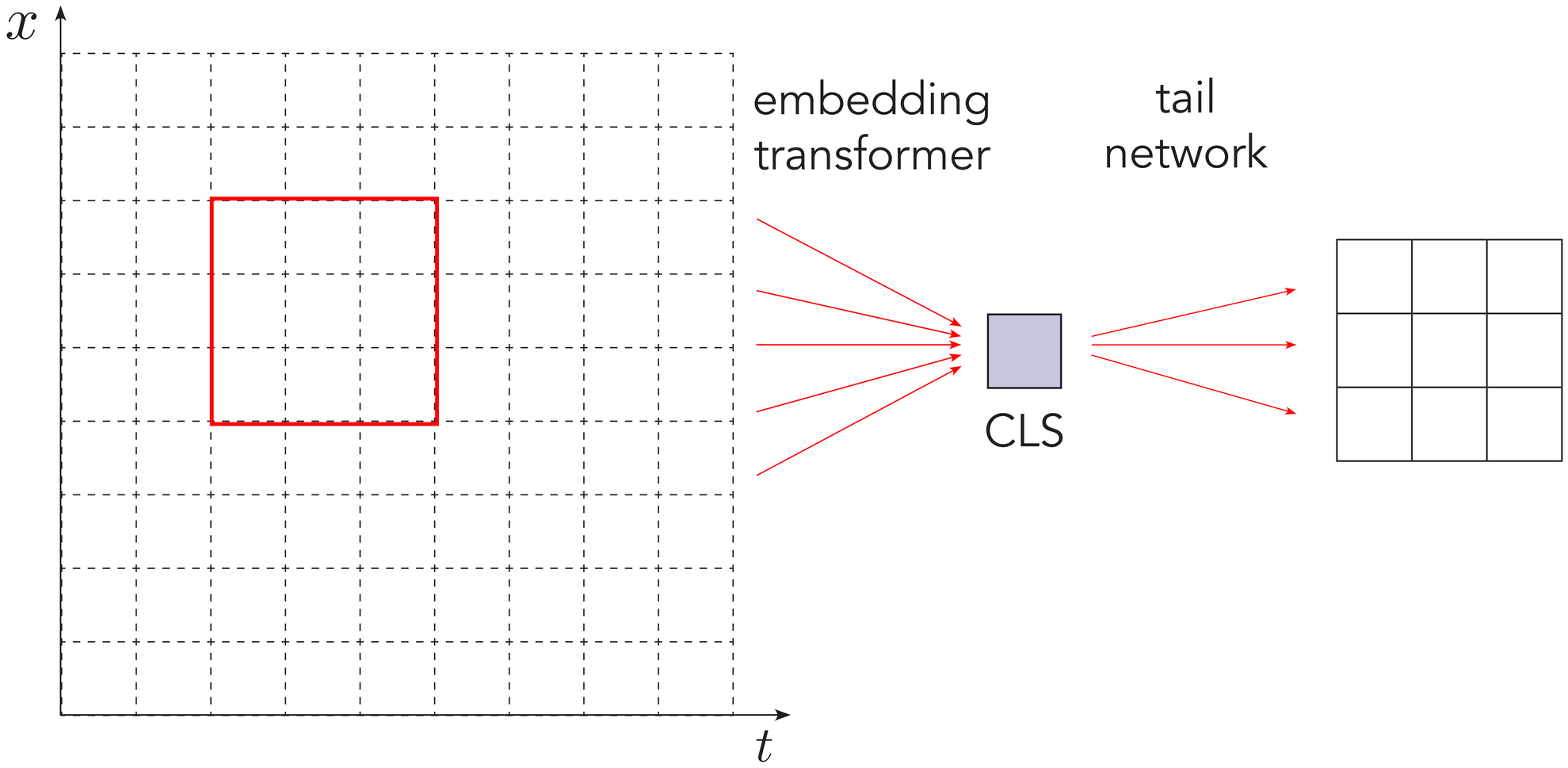
Embedding of tokens



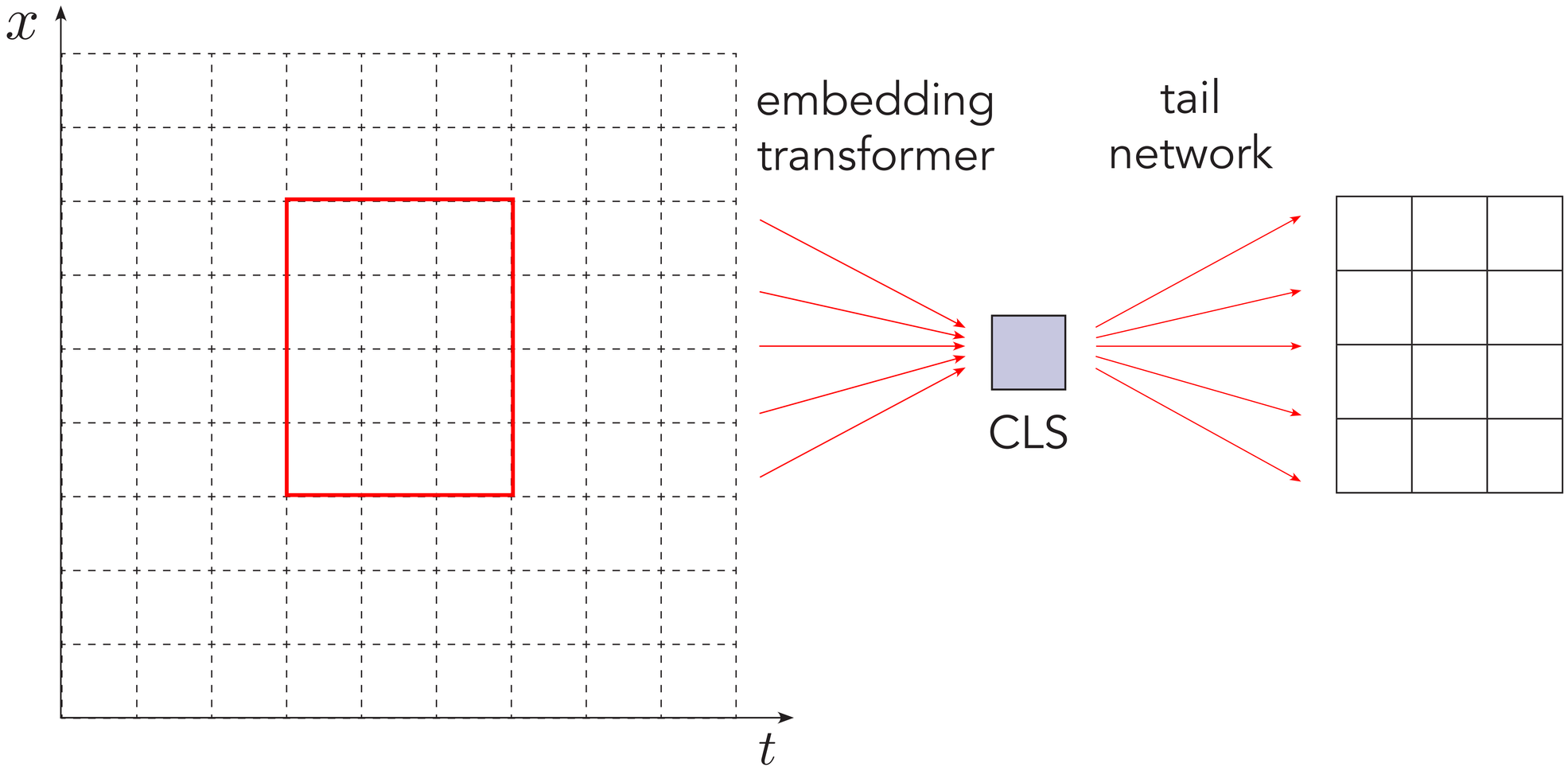
Embedding of tokens



Embedding of tokens



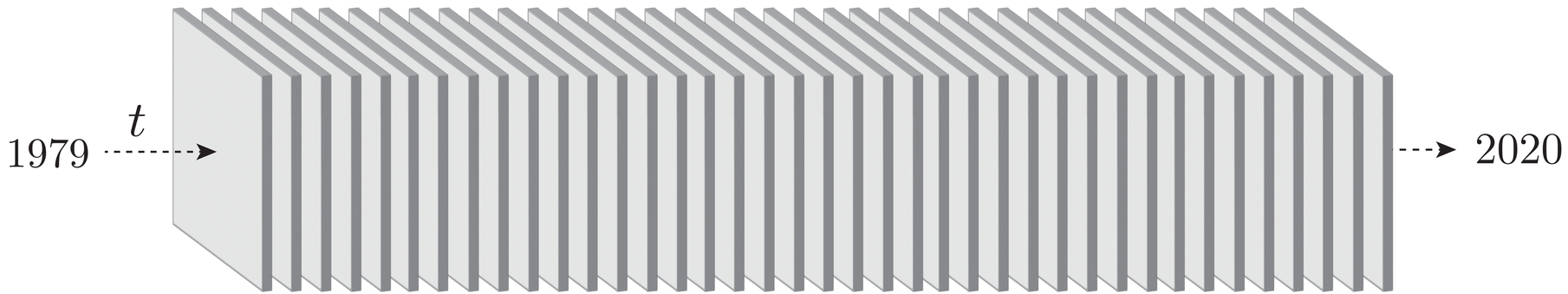
Embedding of tokens



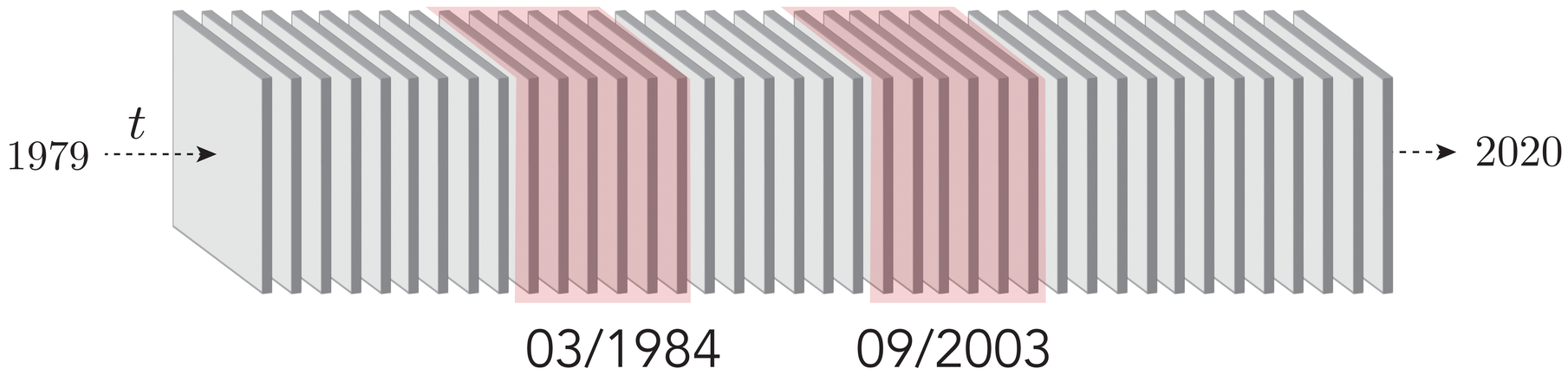
Training

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

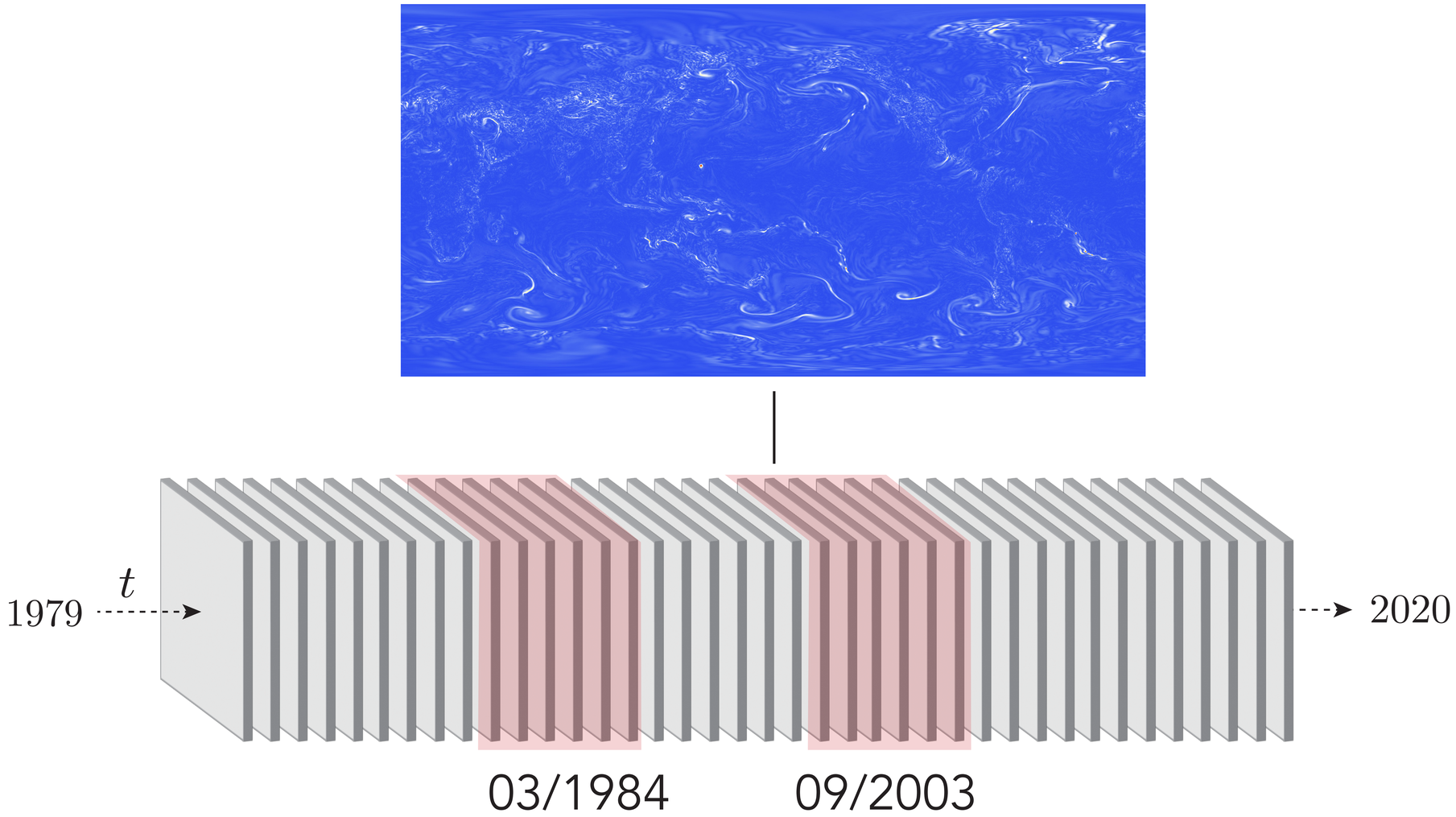
Training



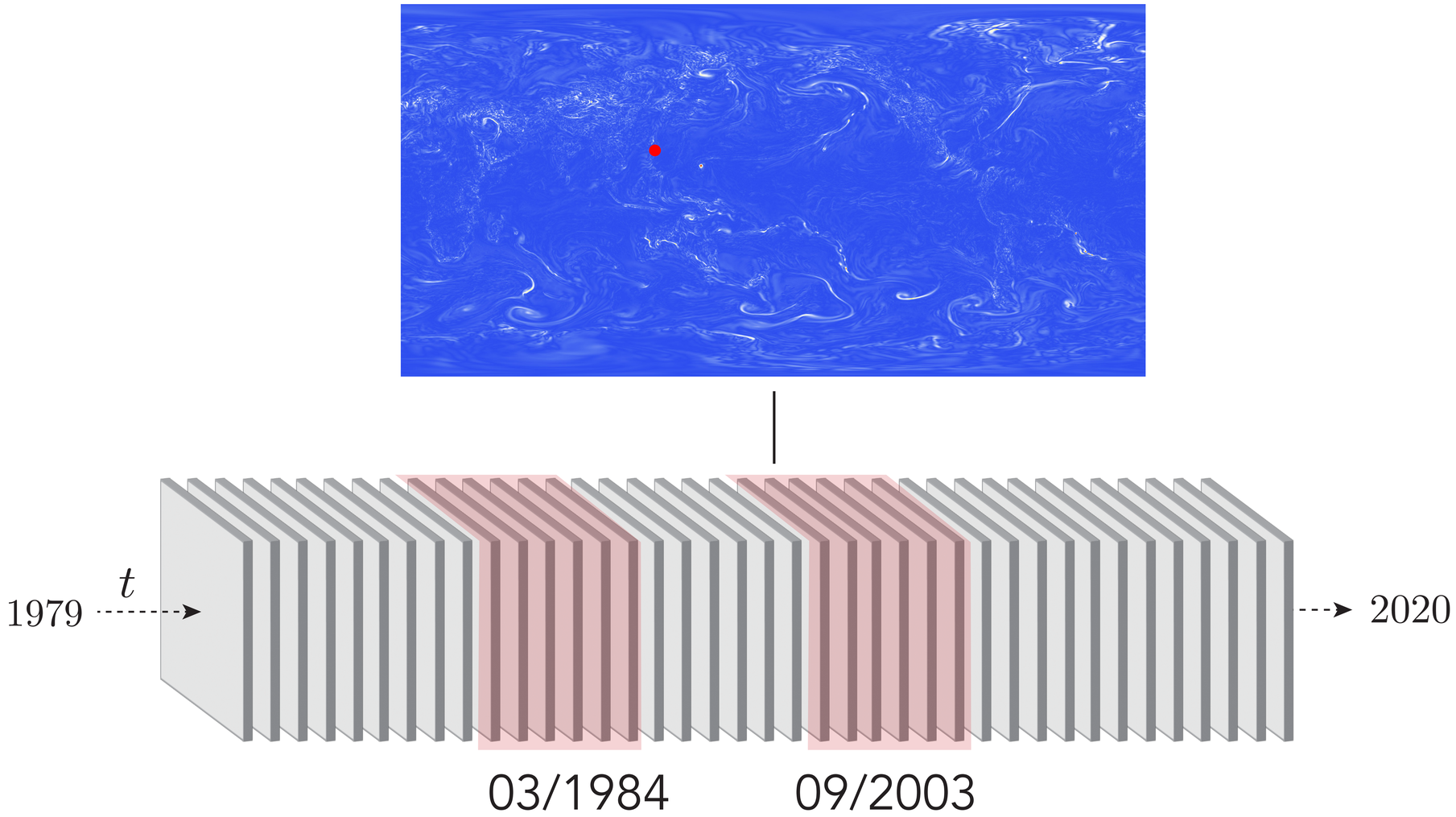
Training



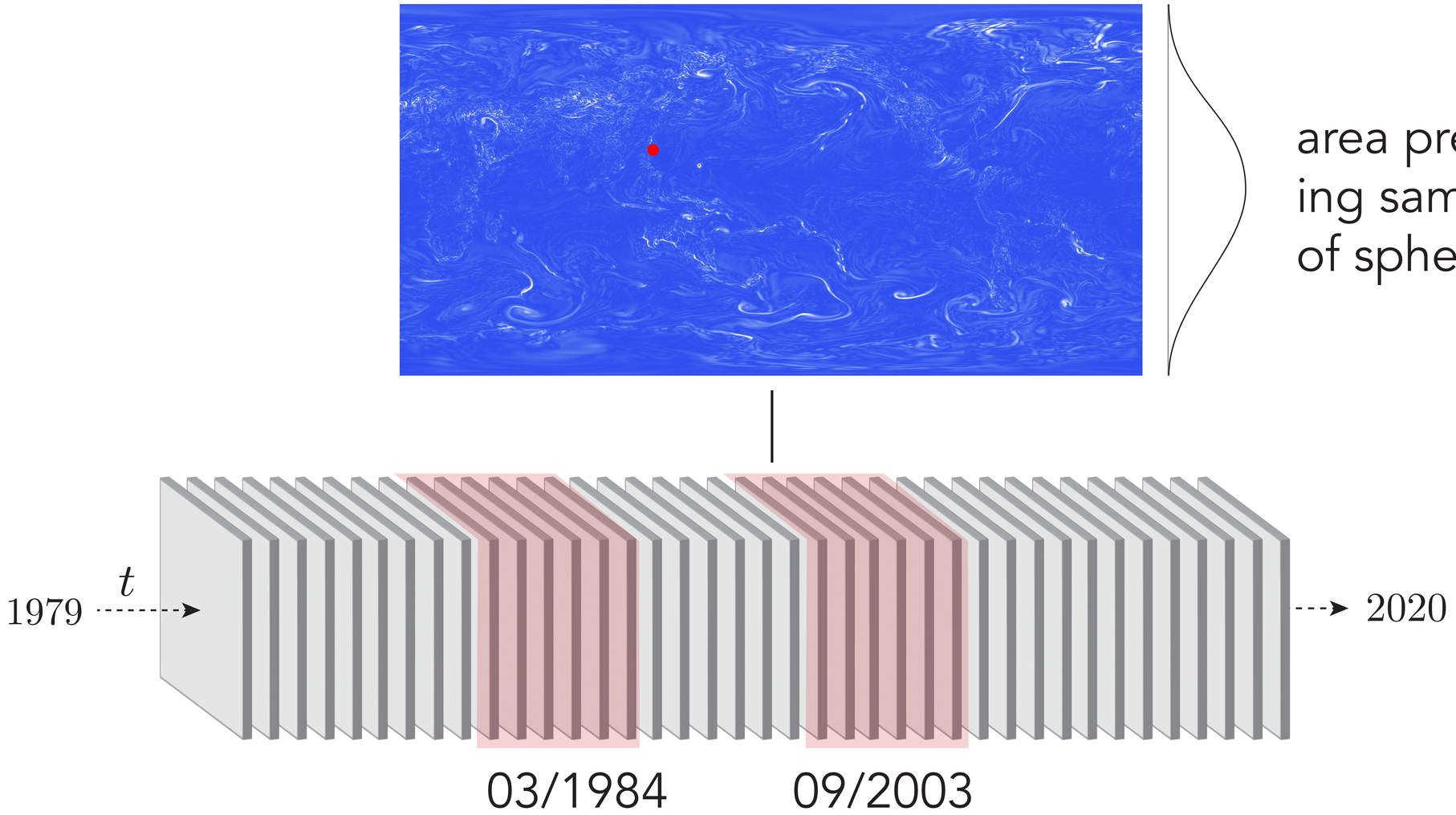
Training



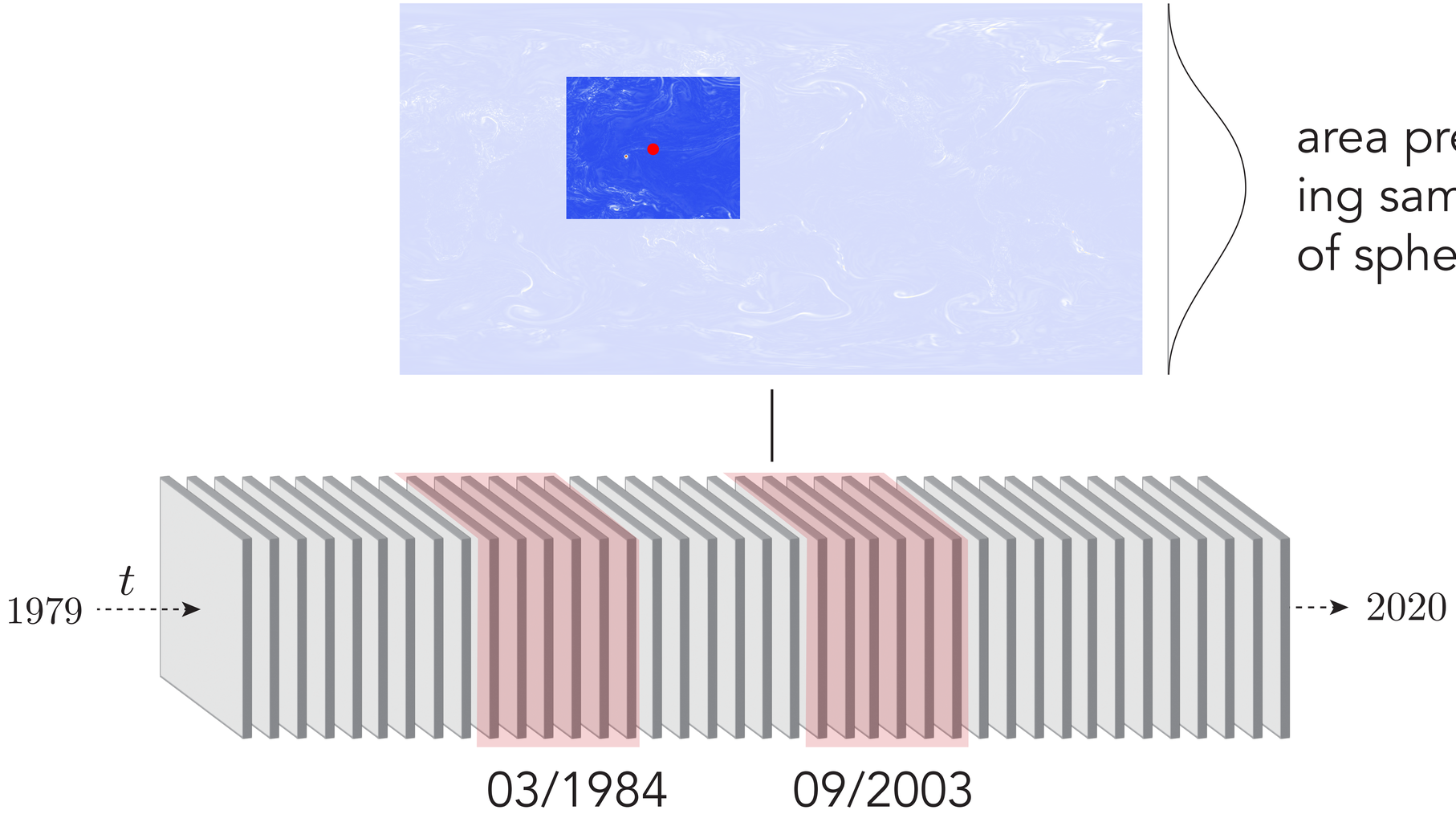
Training



Training

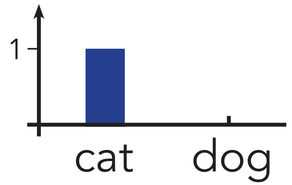


Training

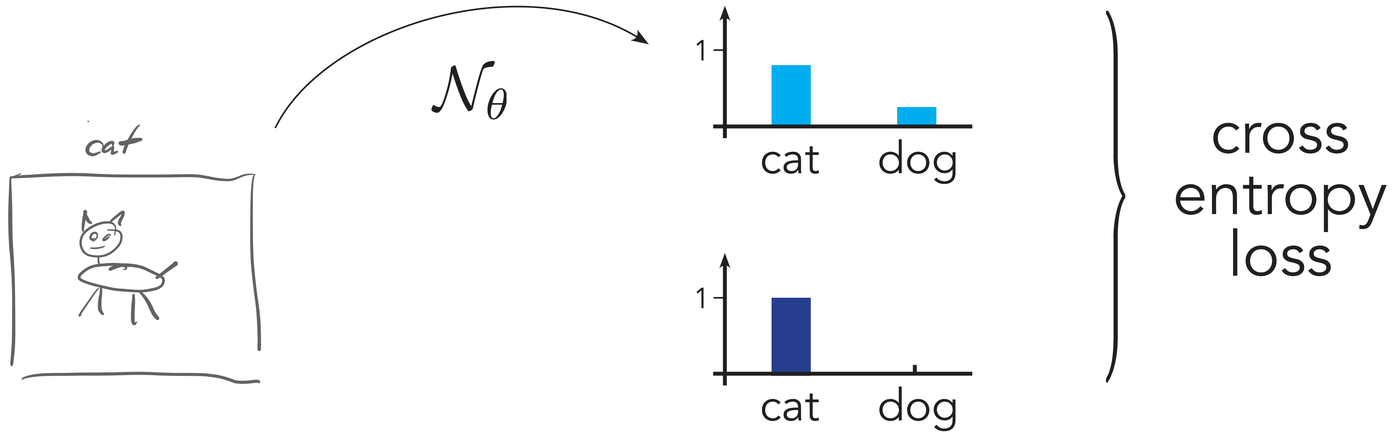


Statistical loss

Statistical loss

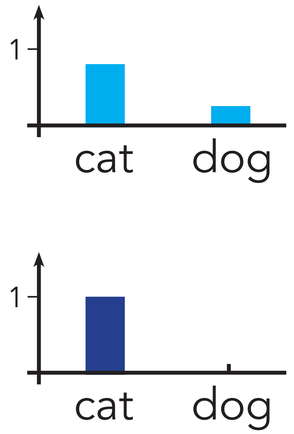


Statistical loss

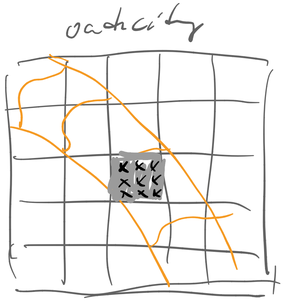


Statistical loss

$\mathcal{N}_\theta$

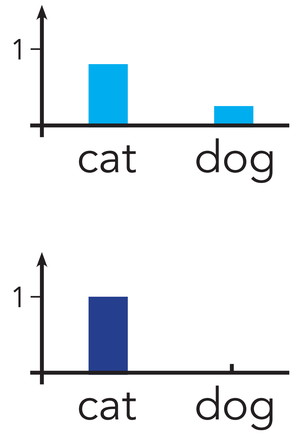


cross entropy loss

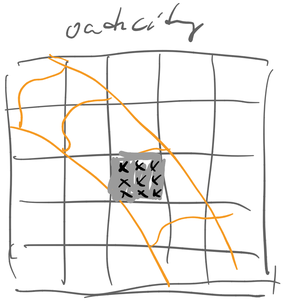


Statistical loss

$\mathcal{N}_\theta$



cross entropy 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:

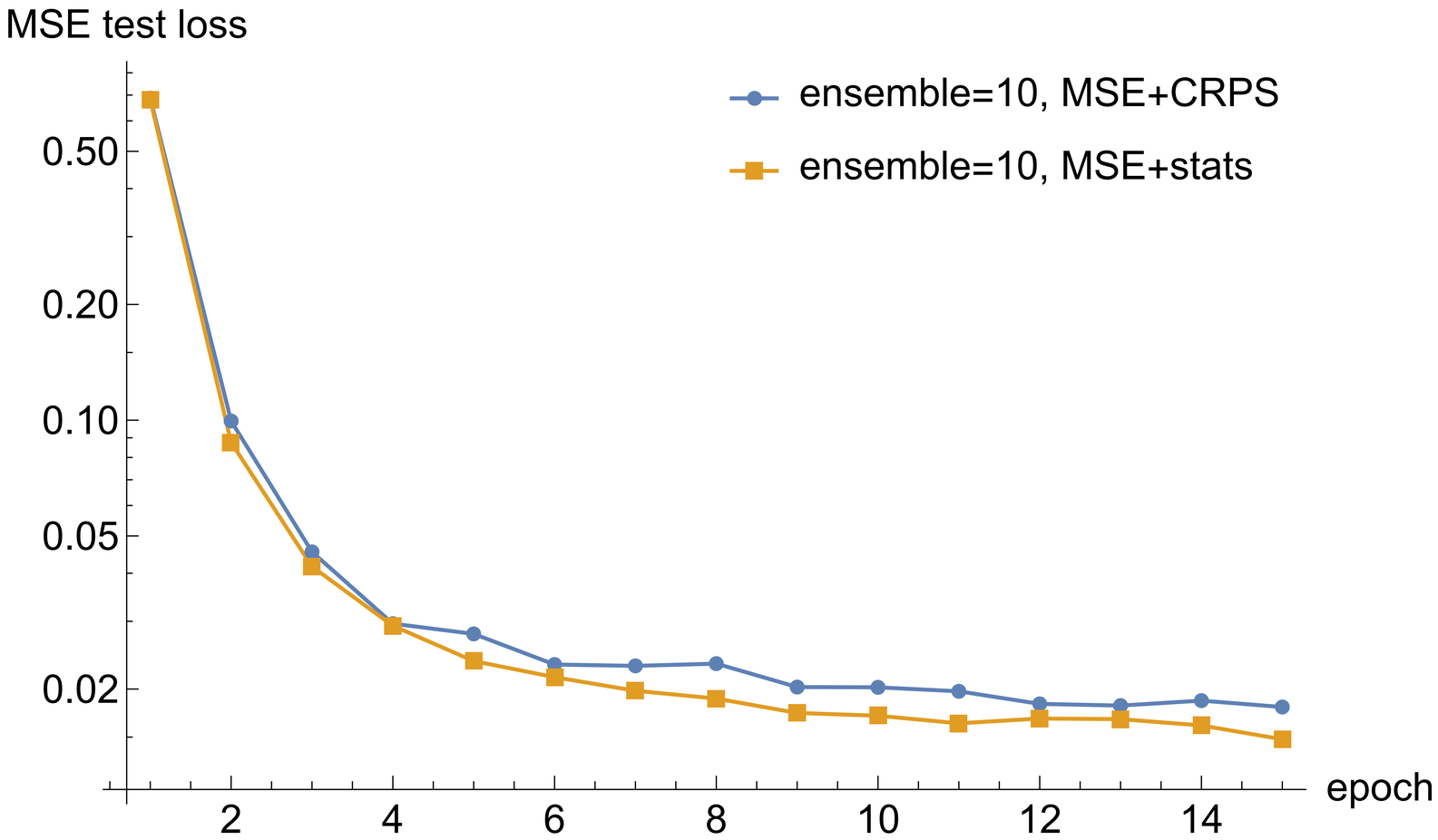
$$\mathcal{L}_{\text{stats}} = \left| 1 - \int_{\mathrm{R}} \delta_y(x)\, G_{\tilde{\mu},\tilde{\sigma}}(x)\, \mathrm{d}x \right|^2$$

- CRPS:[1]

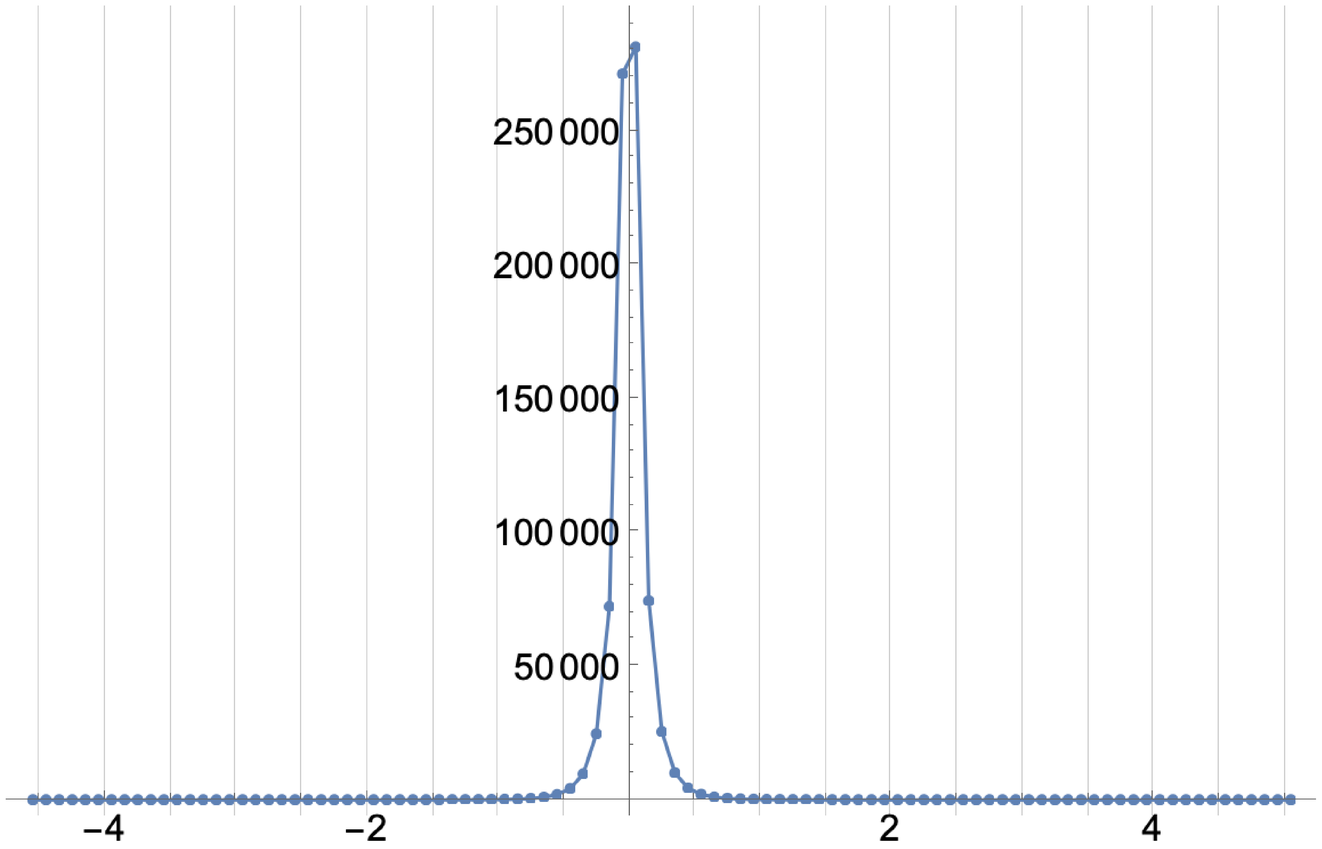
$$\mathcal{L}_{\text{CRPS}} = \int_{\mathrm{R}} \left| H_y(x)\, \mathrm{erf}_{\tilde{\mu},\tilde{\sigma}}(x) \right|^2 \mathrm{d}x$$

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

Statistical loss

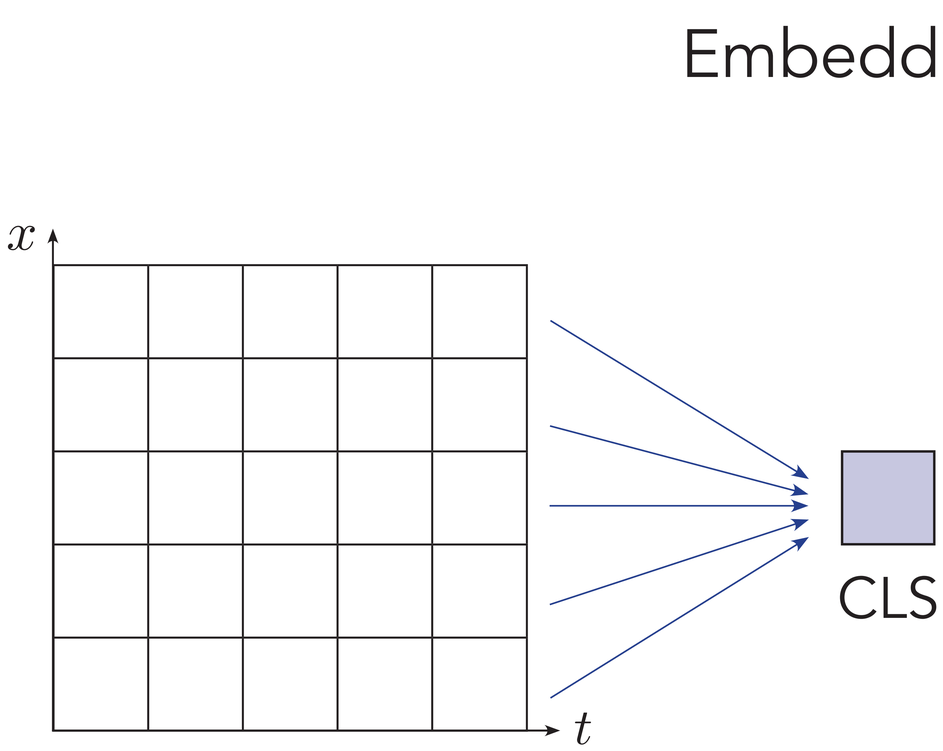


Statistical loss

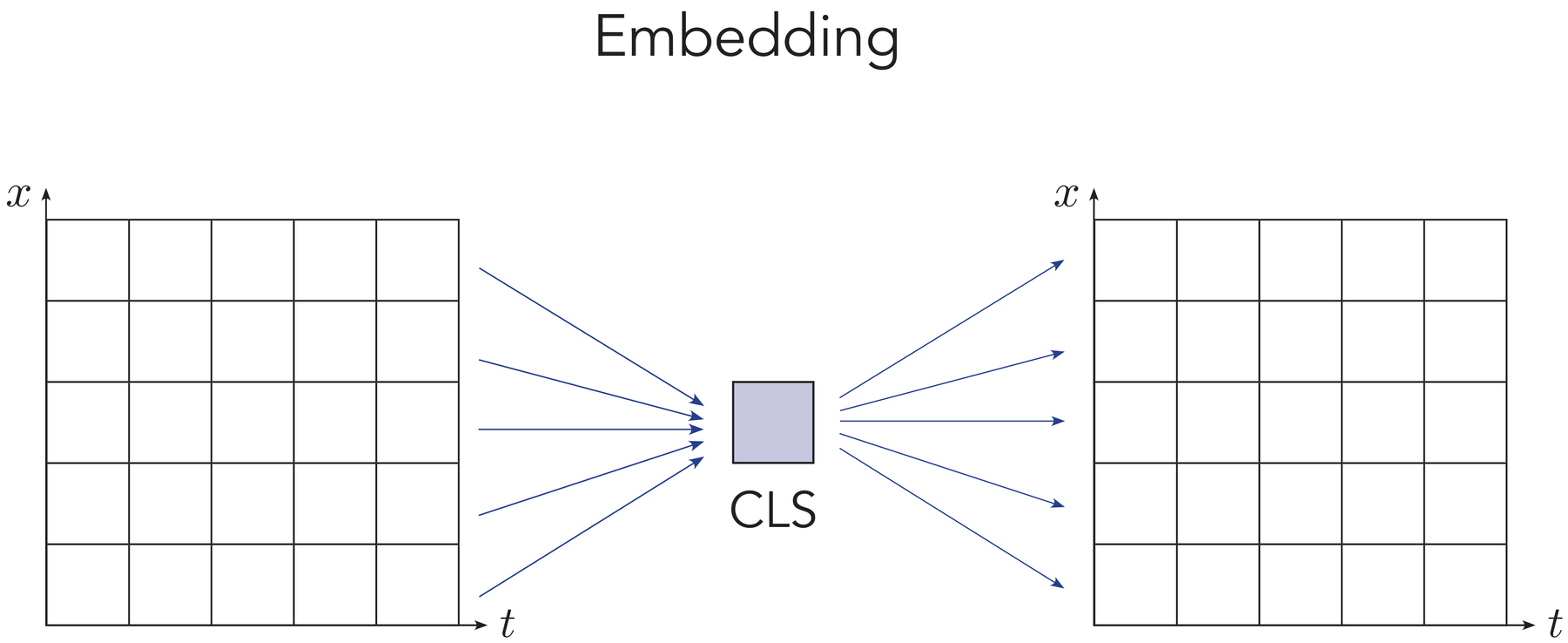


Histogram of ensemble errors

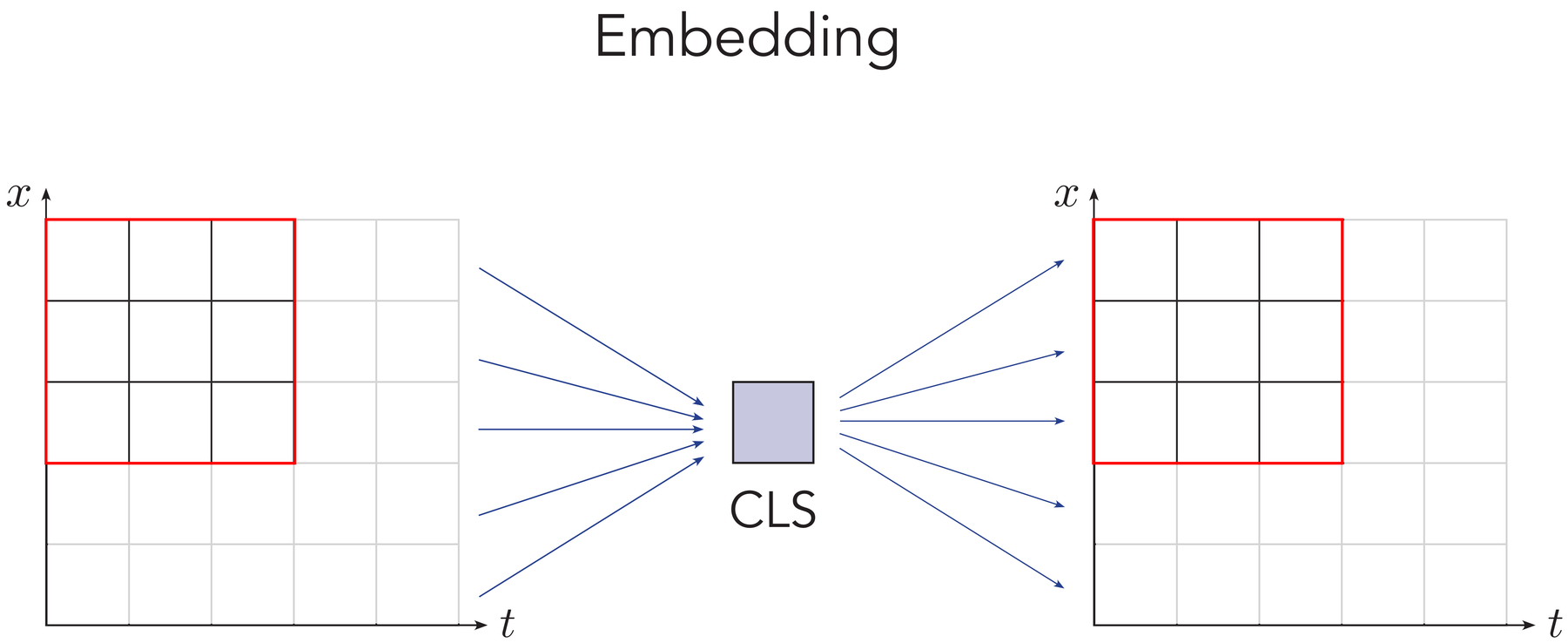
Zero shot evaluation



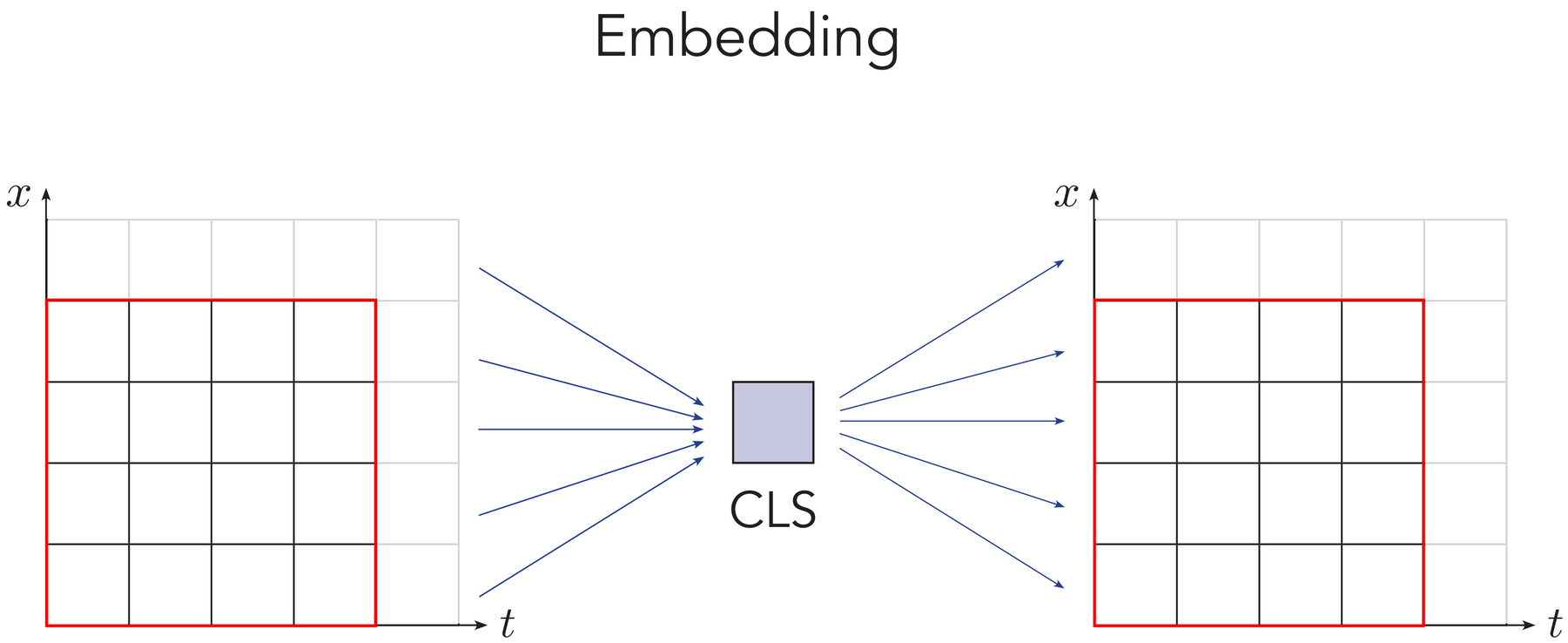
Zero shot evaluation



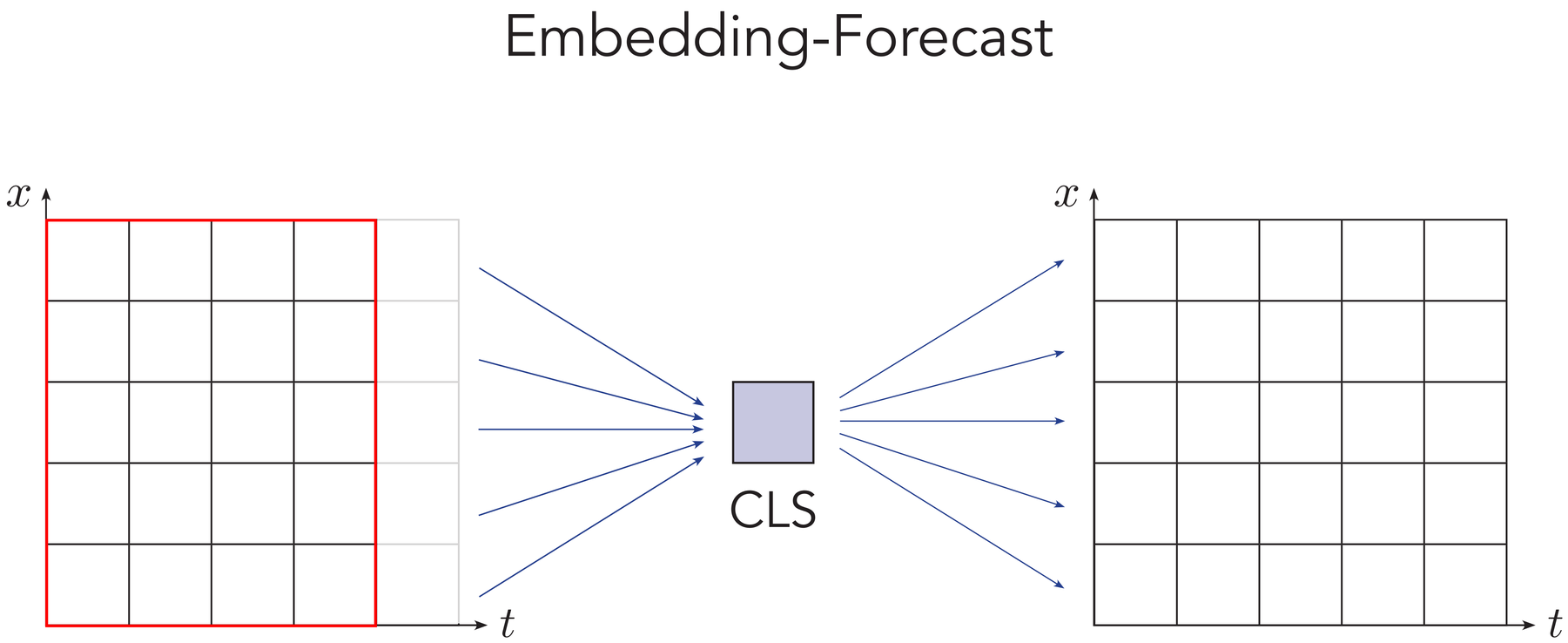
Zero shot evaluation



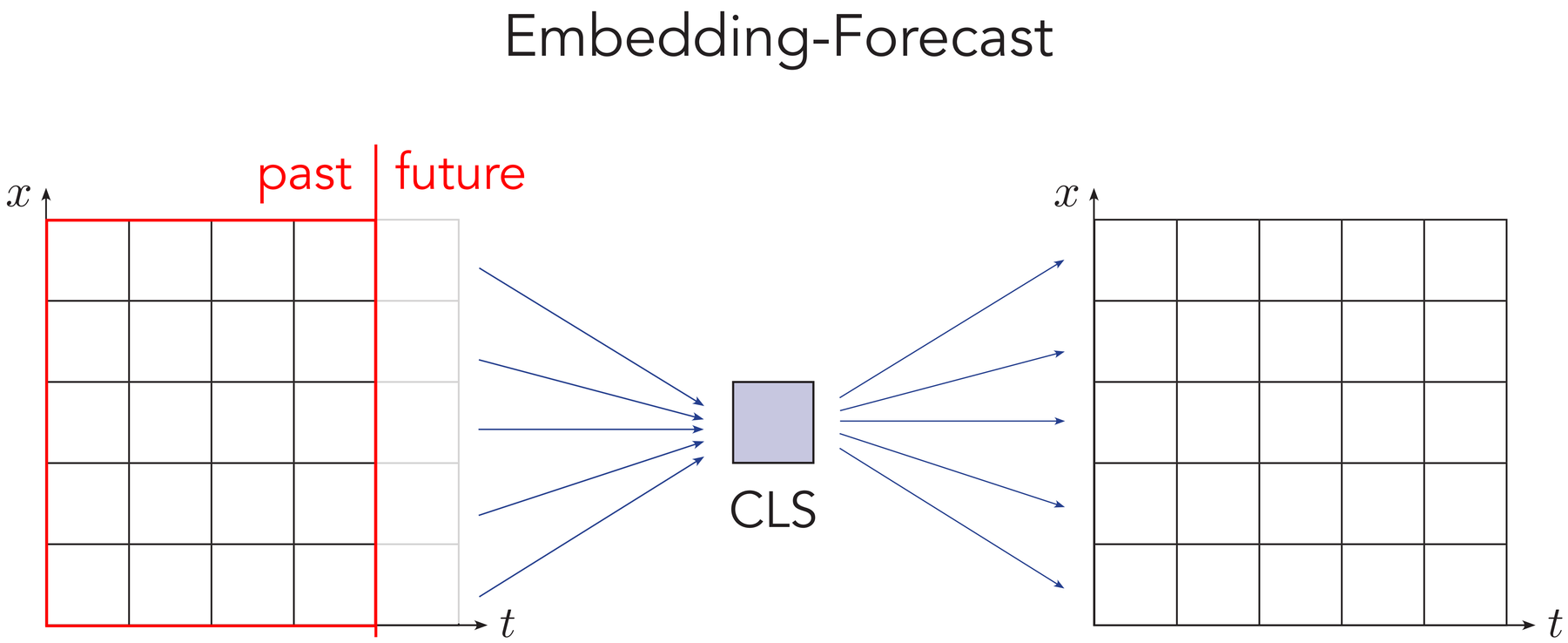
Zero shot evaluation



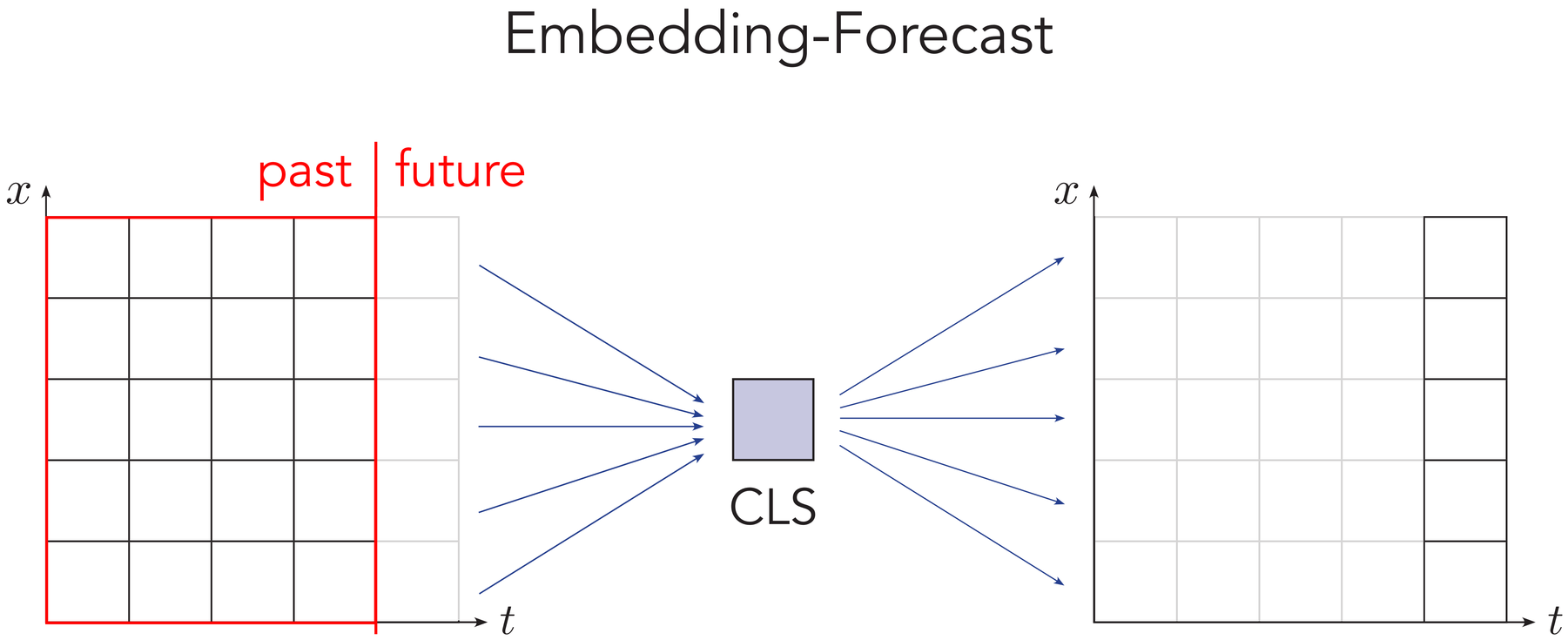
Zero shot evaluation



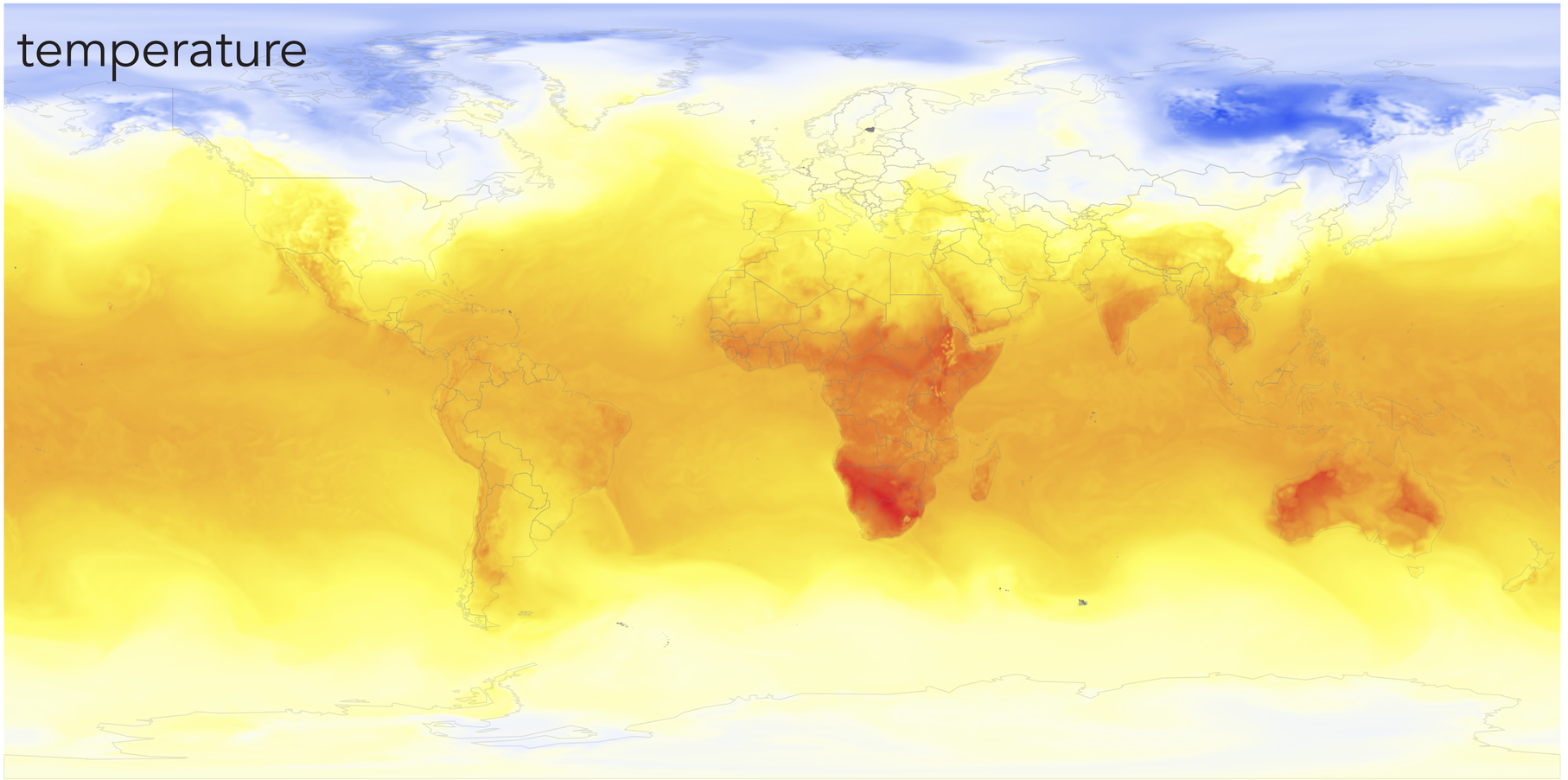
Zero shot evaluation



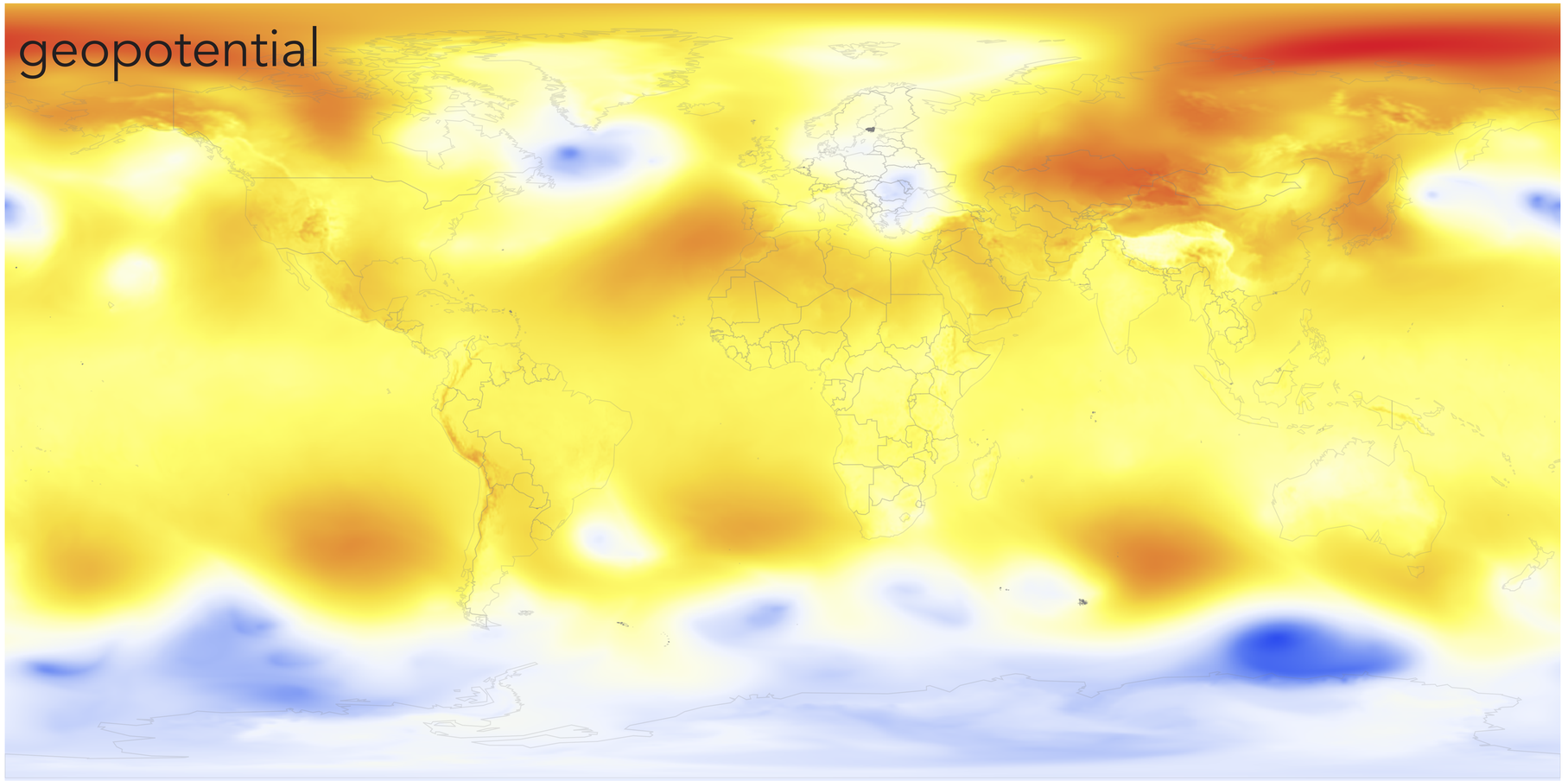
Zero shot evaluation



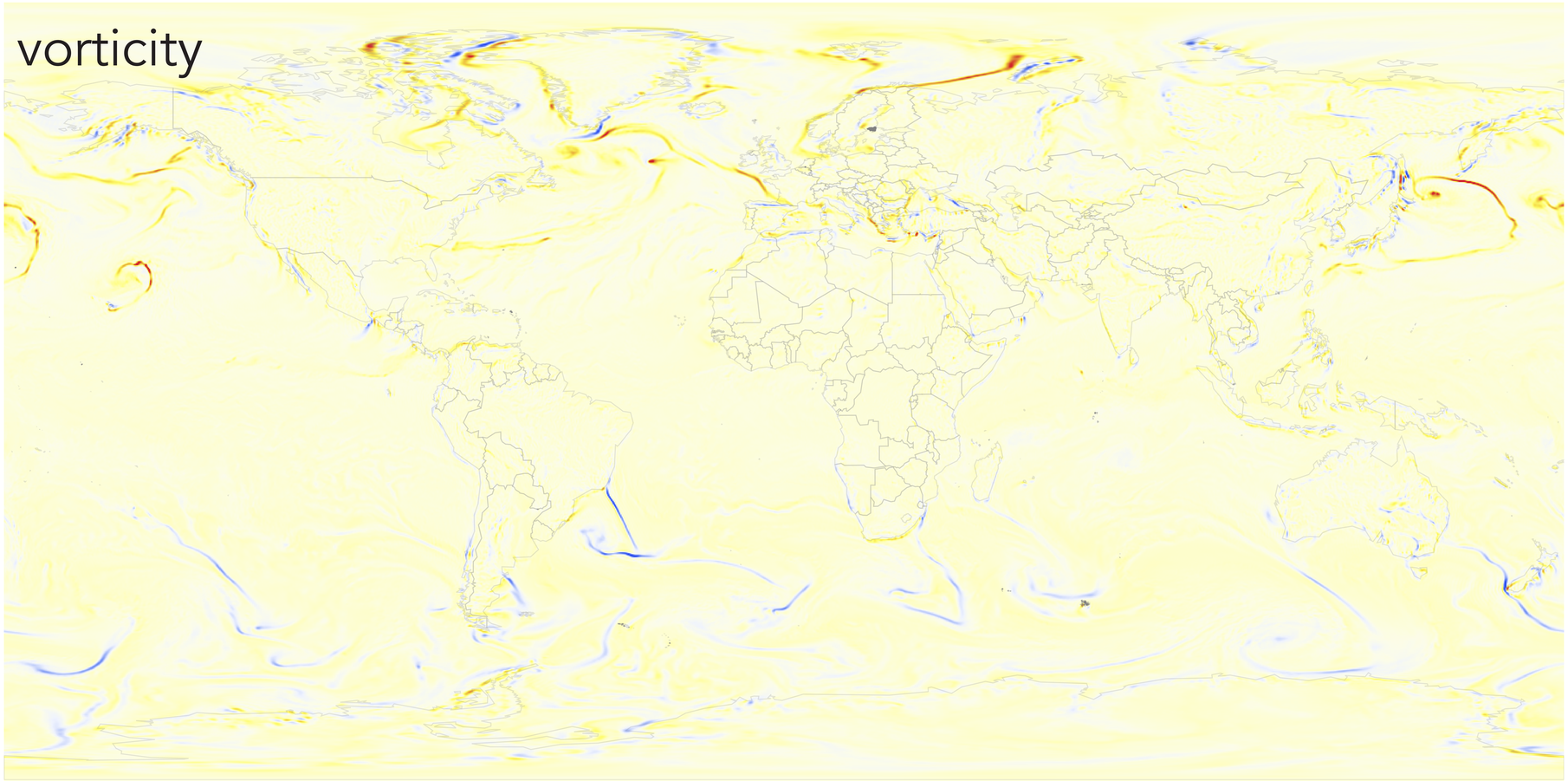
AtmoRep data



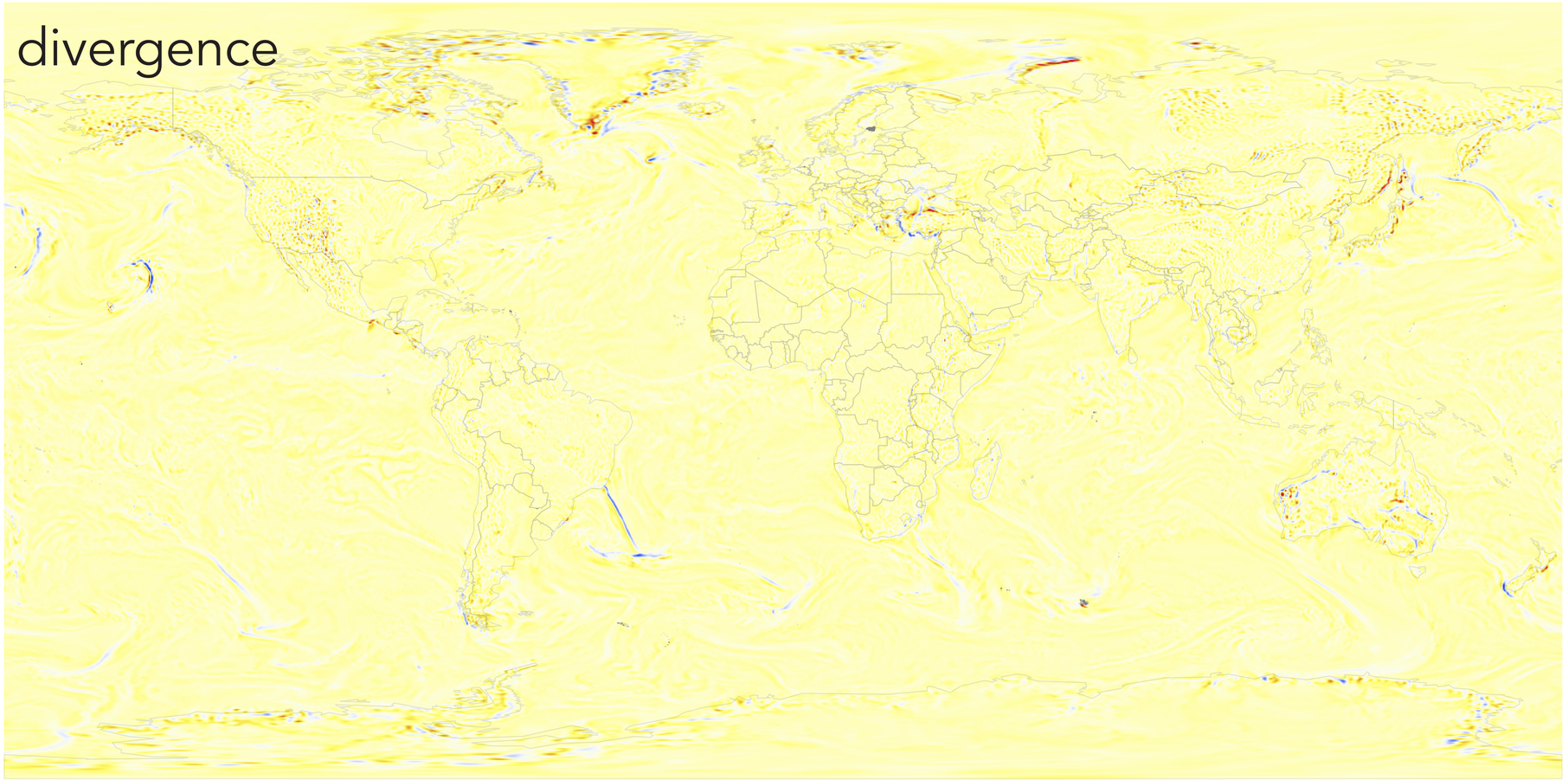
AtmoRep data



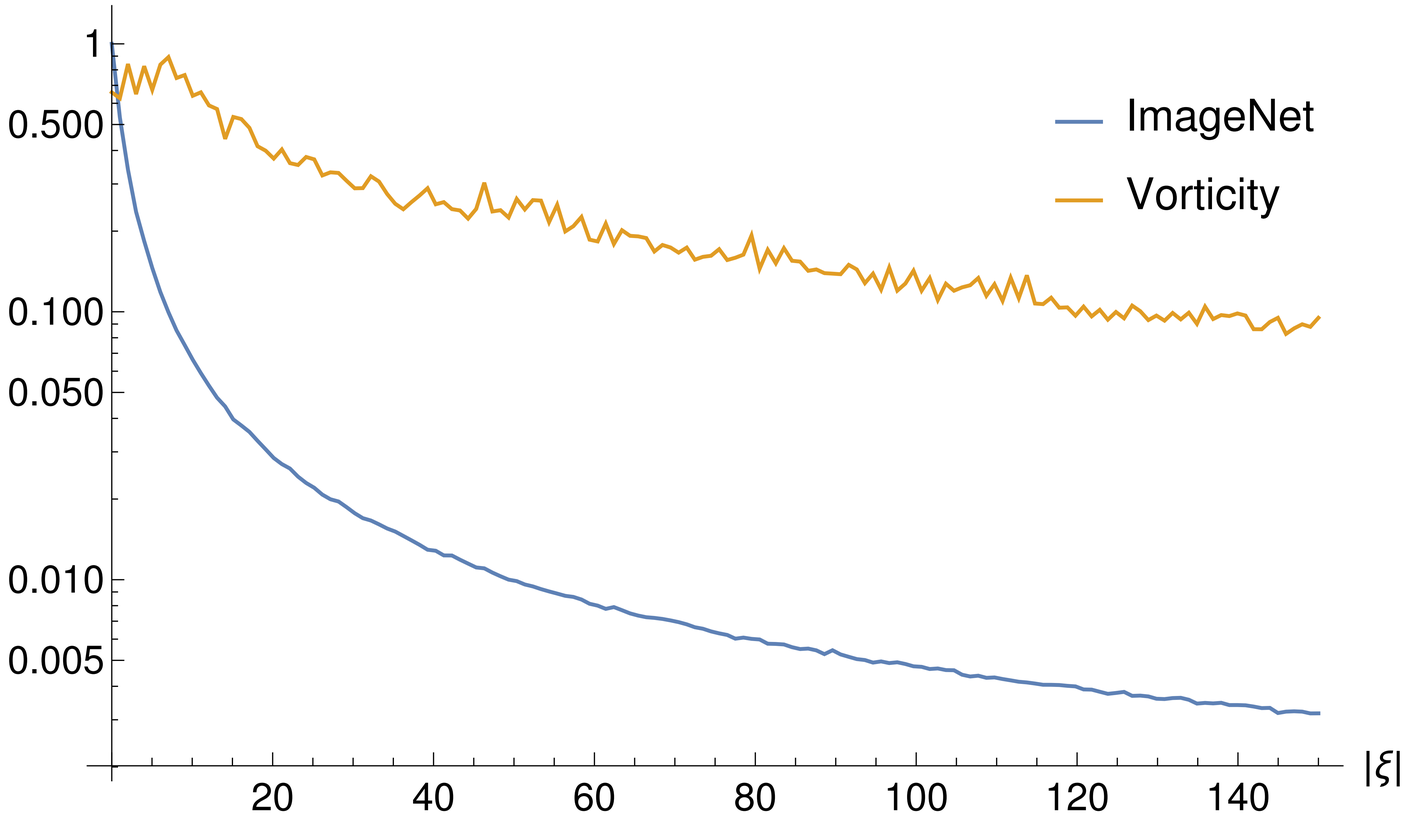
AtmoRep data



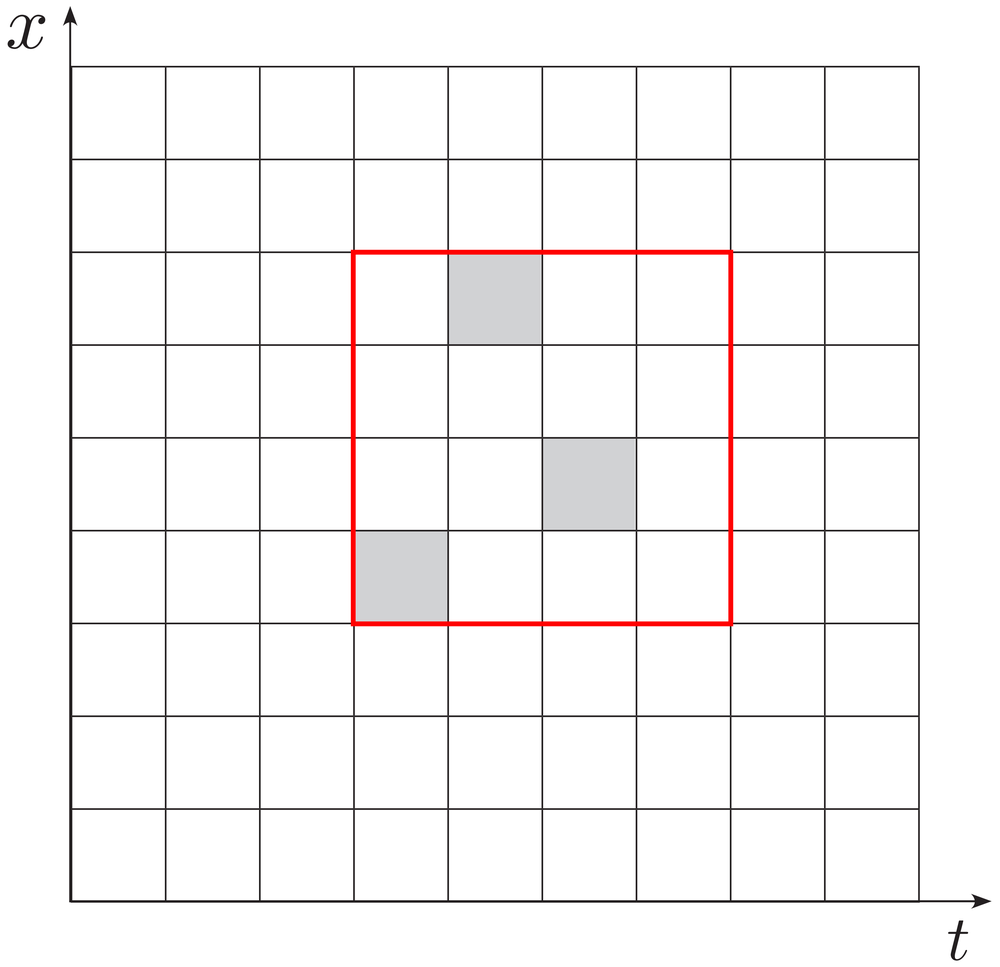
AtmoRep data



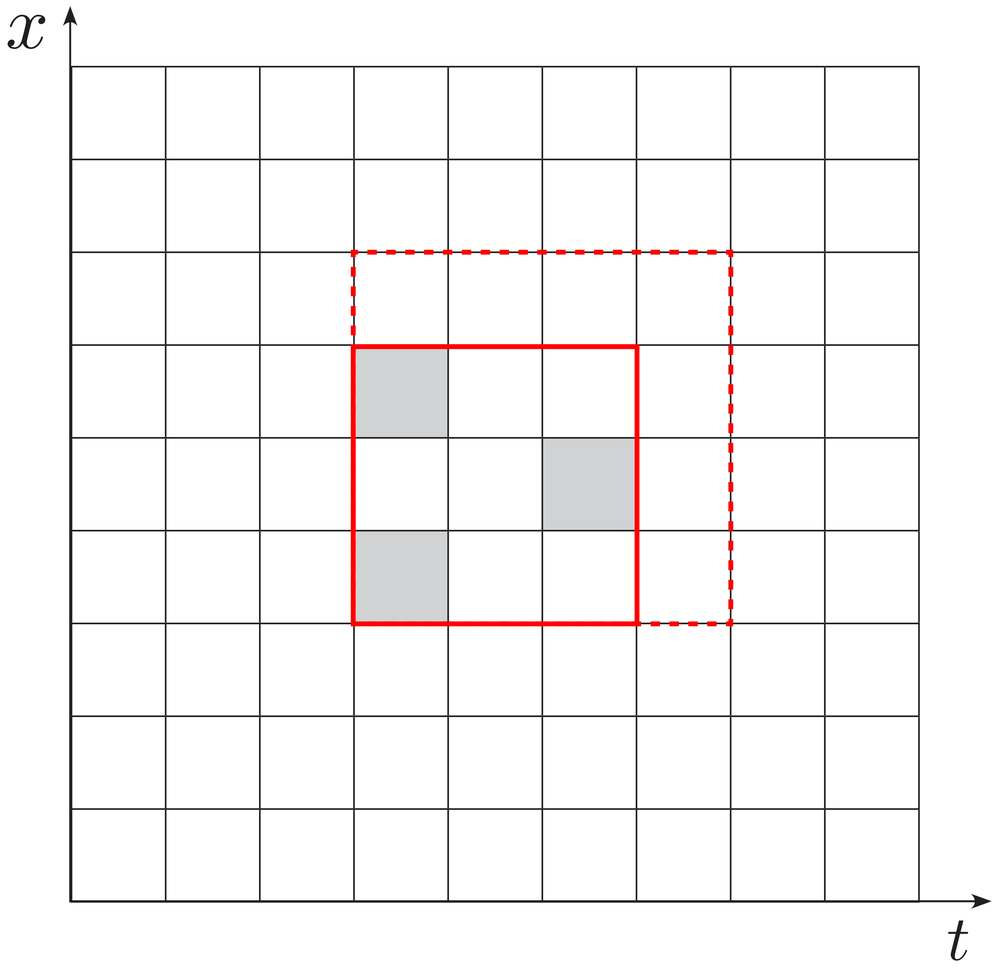
AtmoRep data



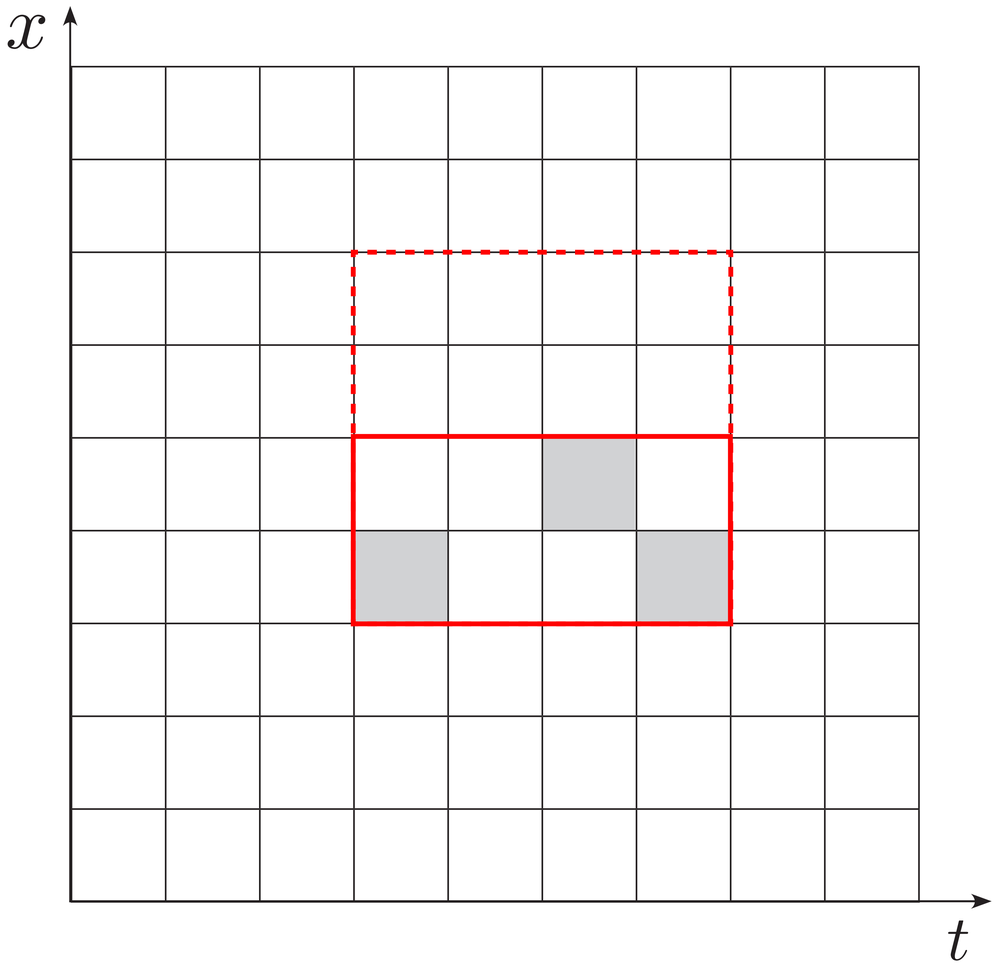
Spatio-temporal BERT



Spatio-temporal BERT

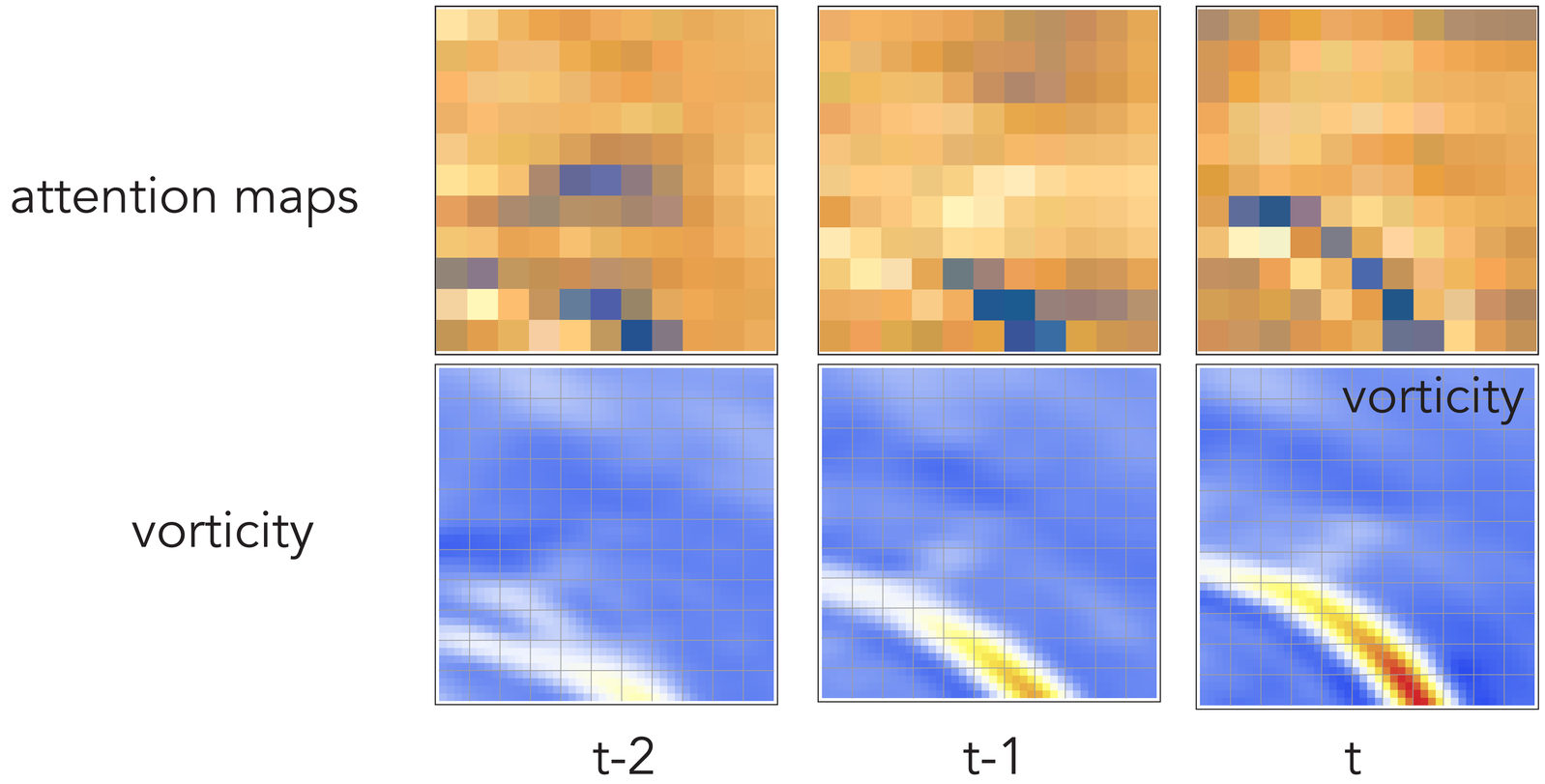


Spatio-temporal BERT

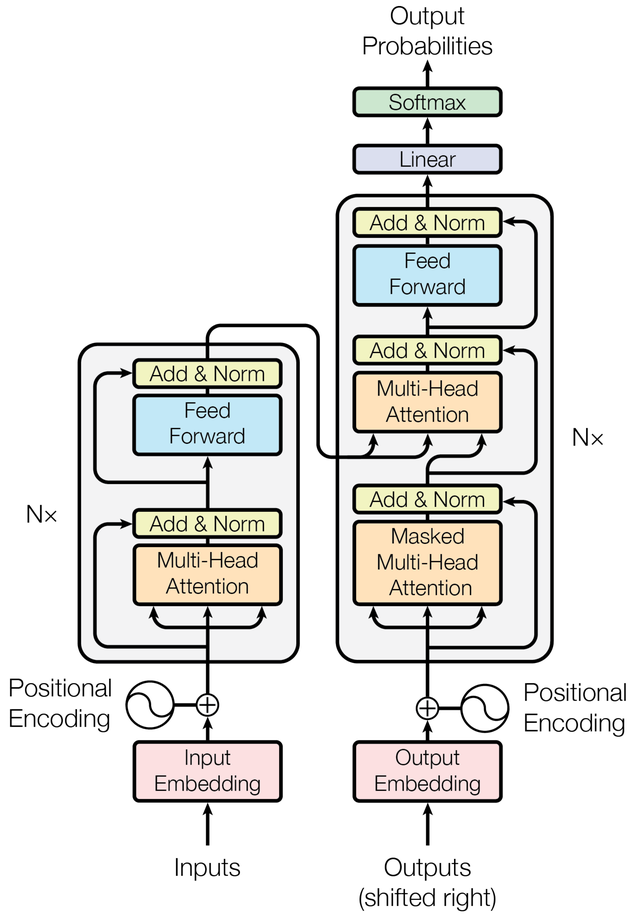


Statistical loss

- Attention maps:

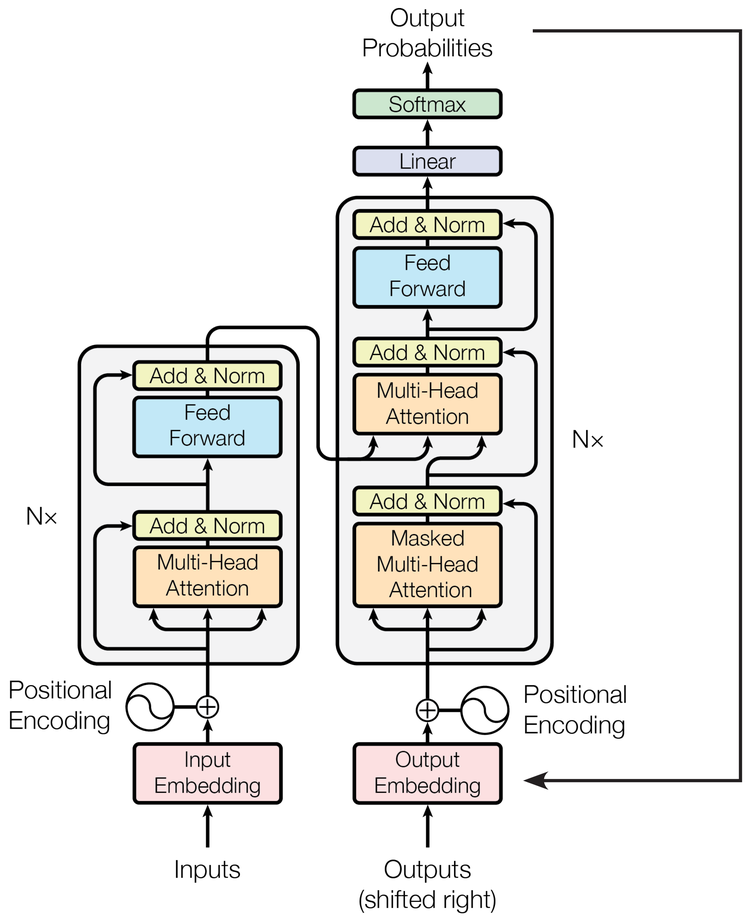


Forecasting / projections



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.

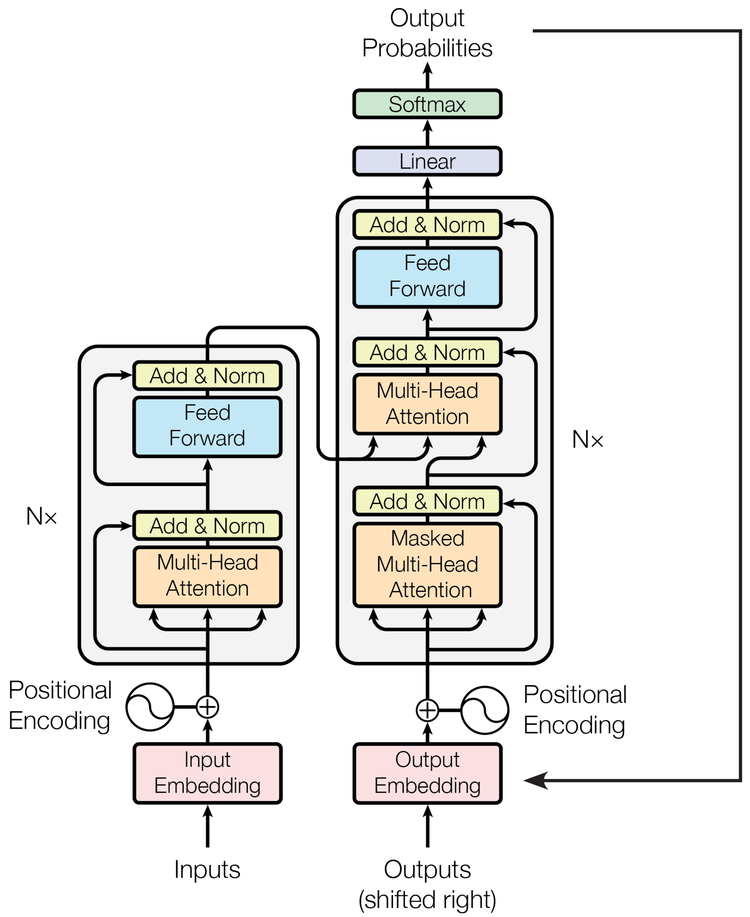
Forecasting / projections



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



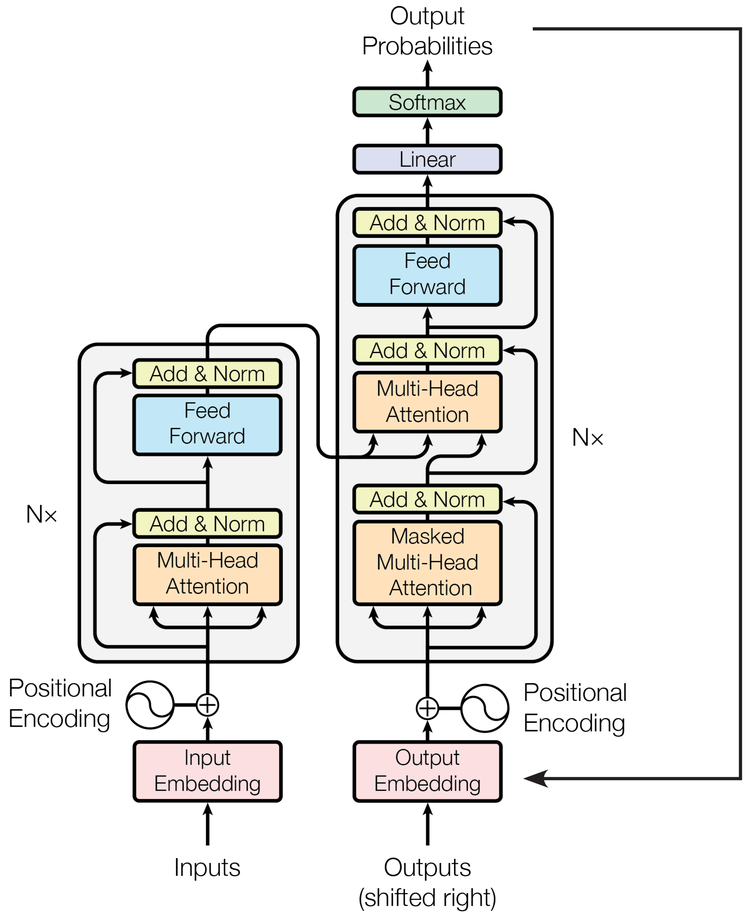
autoregressive, generative modeling

akin to time stepping loop (roll out) for 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

coarse scale/simple classical model



autoregressive, generative modeling

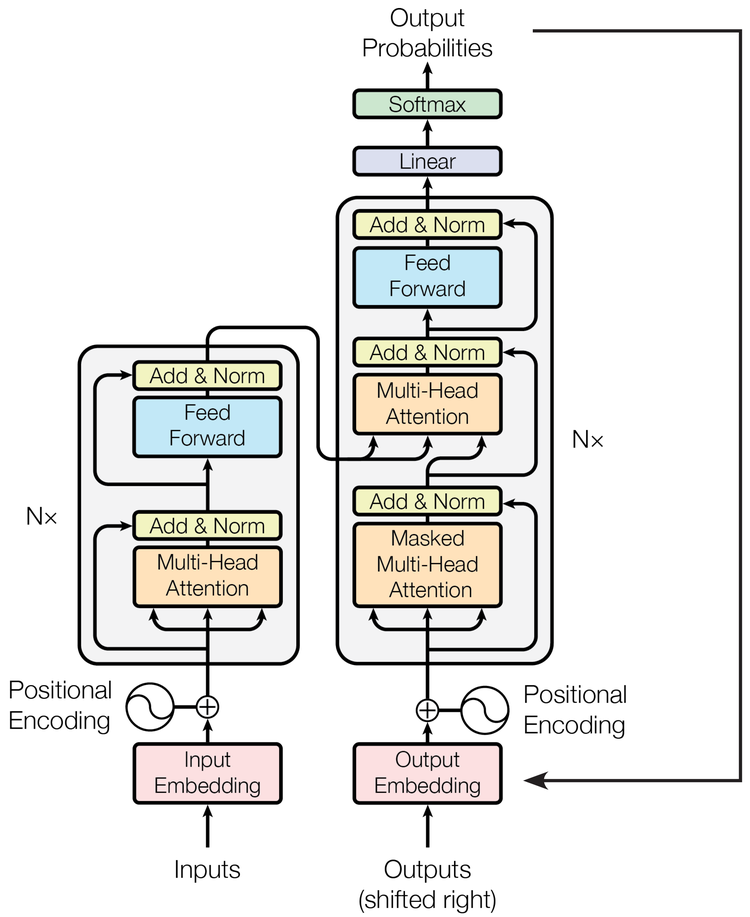
akin to time stepping loop (roll out) for forecasting/projections

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

coarse scale/simple classical model

slow climate variables



autoregressive, generative modeling

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

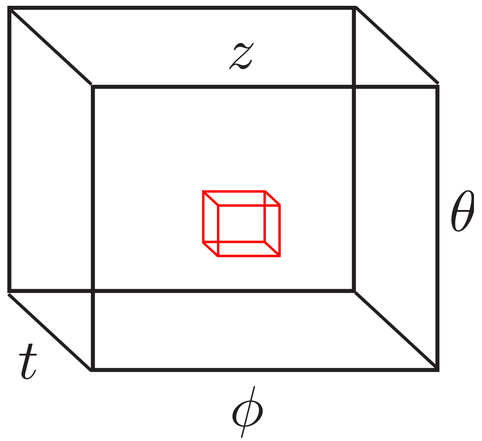
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

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 / inter-action 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