Learning all the text in the internet is easy for your network?

Try learning the weather on earth

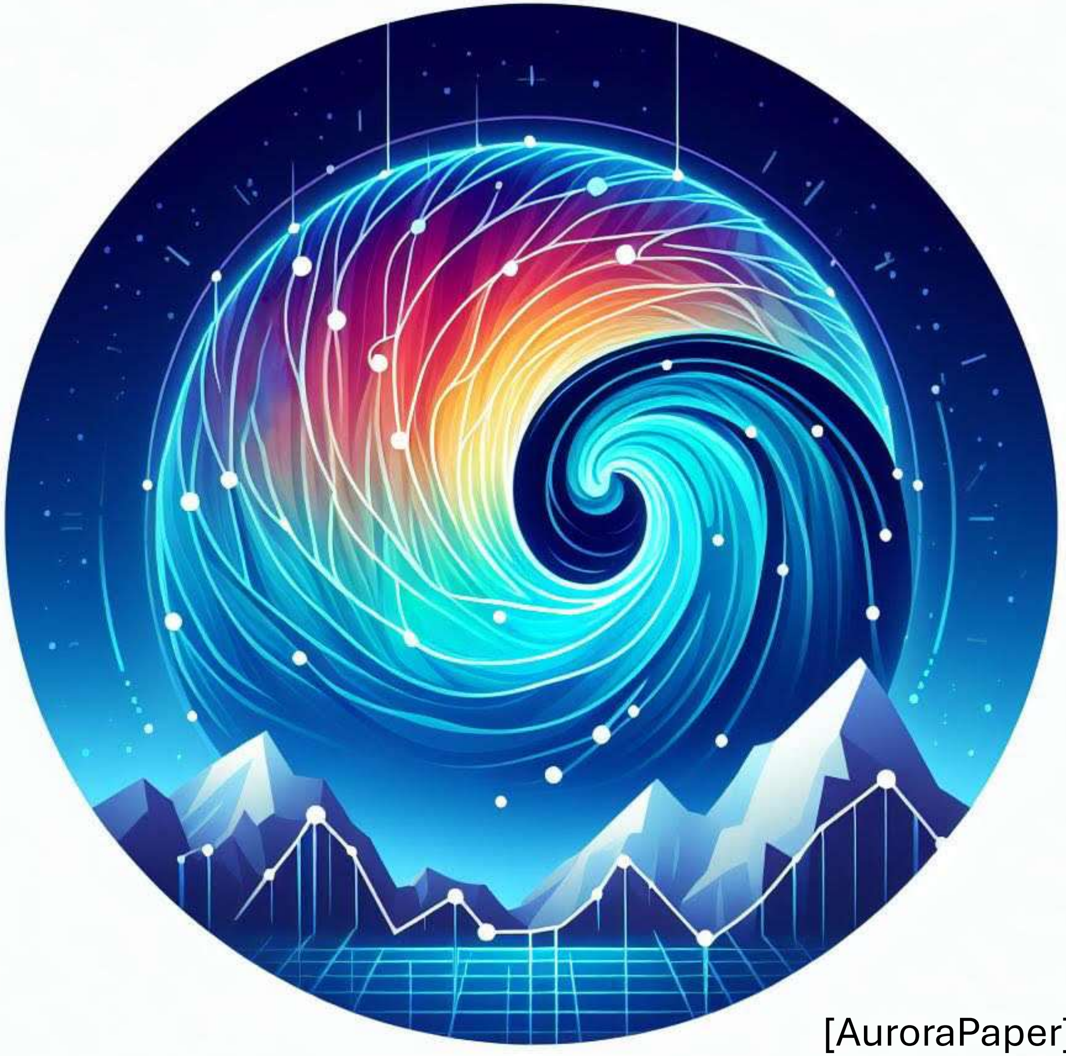
# Aurora: A Foundation Model of the Atmosphere

Frederic Findeis

Deep Learning for Natural Sciences

Lecturer: Karnik Ram

Date: November 12th, 2024



[AuroraPaper]

# Structure of the talk

- Introduction
- Data
- Architecture
- Training
- 2 Test Cases
- Conclusion

# Weather and Climate

[ButterflyImage]

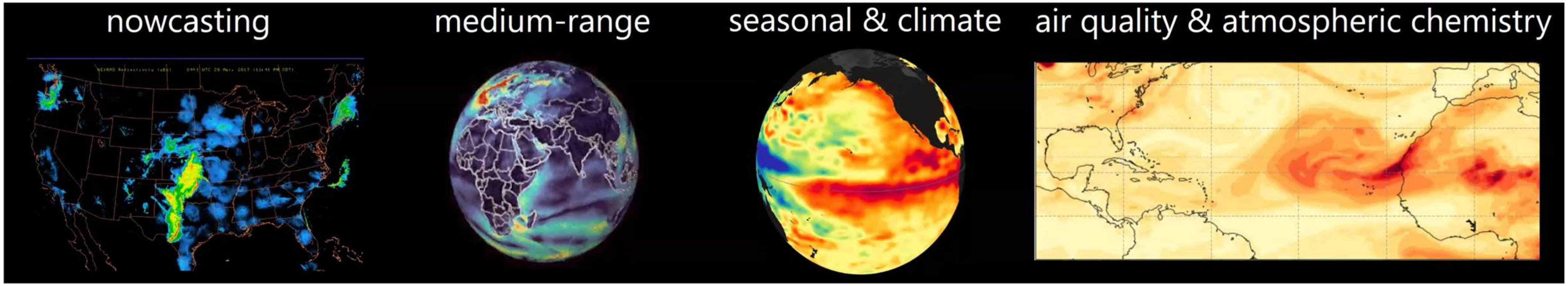
Weather:
- specific point of time
- the state of the atmosphere
- certain location

Climate:
- Average weather over time

➔ Complex physical system with lots of diverse data

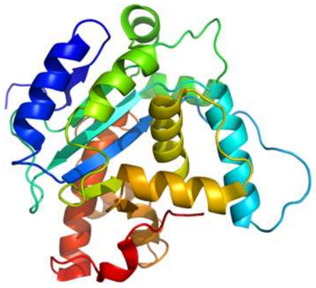
# Types of weather forecast

- Different tasks
- Different variables
- Different resolution 0.1 ° ( around 11 km)




[AuroraYoutubeVideo]

# Initial situation

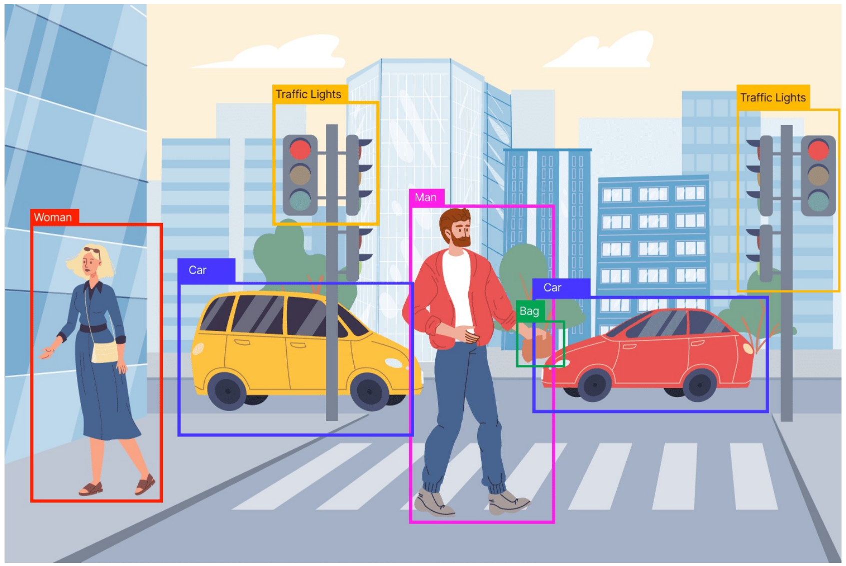


[Protein]

- Current state: Forecast by numerical weather prediction
  - Only parts of available data used
  - Costly
  - Long prediction times
  - Must be rerun, whenever new data gets in

➔Deep Learning:

great results with plenty training data




[ObjectDetection]

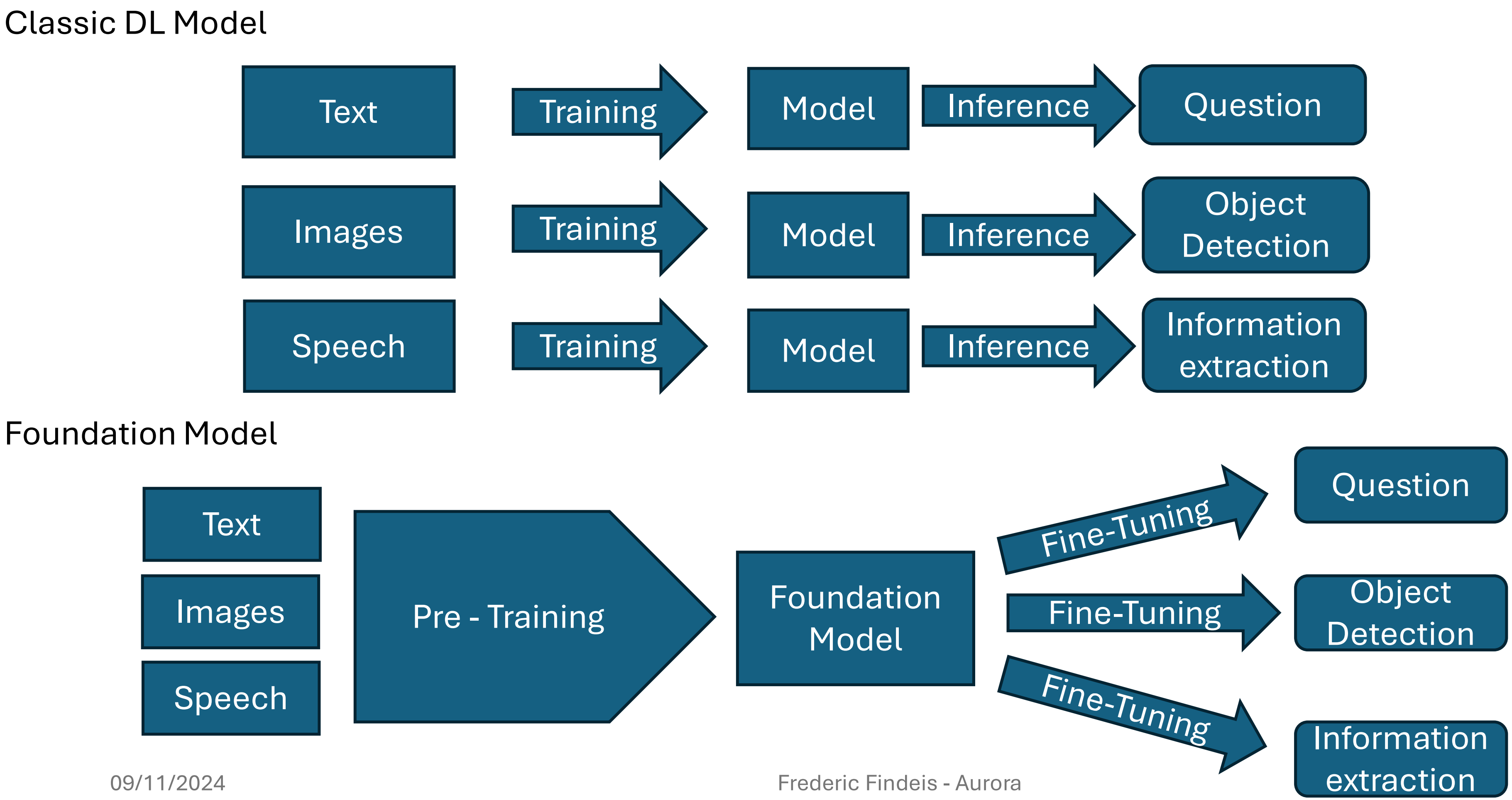
# Goal of Aurora Project

- Produce **operational forecasts** in short time
- outperform:
  - **state of the art simulation** tools
  - specialized **deep learning** models
  - In scenarios with limited data

**Why possible:**
Foundation Model, trained on diverse data

[AIFoundationModels]

# Repetition: Foundation Model



# Dataset Types

- **Forecasts:** prediction of weather
- **Analysis Data:** measurements
- **Reanalysis Data:** historic observations with fixed models

➔ reconstruction past conditions

- **Reforecasts:** reanalysis as initial condition (true forecast)

➔ calibrate current models

- **Climate Simulation:** model to predict response of climate system for different scenarios

# Most important datasets

- ECMWF = European Centre for Medium-Range Weather Forecasts
  - HRES High Resolution
  - ERA5 reanalysis
  - CAMS = Copernicus Atmosphere Monitoring Service
- NCEP = U. S. National Centers for Environmental Prediction
  - GFS ( Global Forecast System)

[ECMWFLogo]

[NWS]

# Pretraining –
# Mixture 6 weather and climate datasets

- Diverse data:
  - Resolution
  - Different times
  - Variables
  - Pressure levels
- Big data

Pretraining Datasets

| Name | Resolution | Timeframe | Surface Variables | Atmospheric Variables | Num levels | Size (TB) | Num frames |
|---|---|---|---|---|---|---|---|
| ERA5 | 0.25° × 0.25° | 1979-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 13 | 105.43 | 367,920 |
| HRES-0.25 | 0.25° × 0.25° | 2016-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 13 | 42.88 | 149,650 |
| IFS-ENS-0.25 | 0.25° × 0.25° | 2018-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 3 | 518.41 | 6,570,000 |
| GFS Forecast | 0.25° × 0.25° | 2015-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 13 | 130.39 | 560,640 |
| GFS Analysis | 0.25° × 0.25° | 2015-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 13 | 2.04 | 8,760 |
| GEFS Reforecast | 0.25° × 0.25° | 2000-2019 | 2T, MSL | U, V, T, Q, Z | 3 | 194.02 | 2,920,000 |
| CMCC-CM2-VHR4 | 0.25° × 0.25° | 1950-2014 | 2T, U10, V10, MSL | U, V, T, Q | 7 | 12.6 | 94,900 |
| ECMWF-IFS-HR | 0.45° × 0.45° | 1950-2014 | 2T, U10, V10, MSL | U, V, T, Q | 7 | 3.89 | 94,900 |
| MERRA-2 | 0.625° × 0.5° | 1980-2020 | 2T, U10, V10, MSL | U, V, T, Q | 13 | 5.85 | 125,560 |
| IFS-ENS-Mean | 0.25° × 0.25° | 2018-2020 | 2T, U10, V10, MSL | U, V, T, Q, Z | 3 | 10.37 | 131,400 |
| | | | | | Total | 1,219.91 | 11,023,730 |

[AuroraPaper]

BUT: ALSO INCOMPLETE DATA

# Split

**Pretraining**
Validate with IFS HRES at 0.25° from 2020

| Name | Timeframe |
|---|---|
| ERA5 | 1979-2020 |
| HRES-0.25 | 2016-2020 |
| IFS-ENS-0.25 | 2018-2020 |
| GFS Forecast | 2015-2020 |
| GFS Analysis | 2015-2020 |
| GEFS Reforecast | 2000-2019 |
| CMCC-CM2-VHR4 | 1950-2014 |
| ECMWF-IFS-HR | 1950-2014 |
| MERRA-2 | 1980-2020 |
| IFS-ENS-Mean | 2018-2020 |

**Test**
Years 2022 / 2023
Depending on dataset

| Name | Timeframe |
|---|---|
| HRES-0.25 | 2022 |
| HRES-0.1 | 2023 |
| CAMS Analysis | June 2022 – Nov 2022 |

**Finetune**

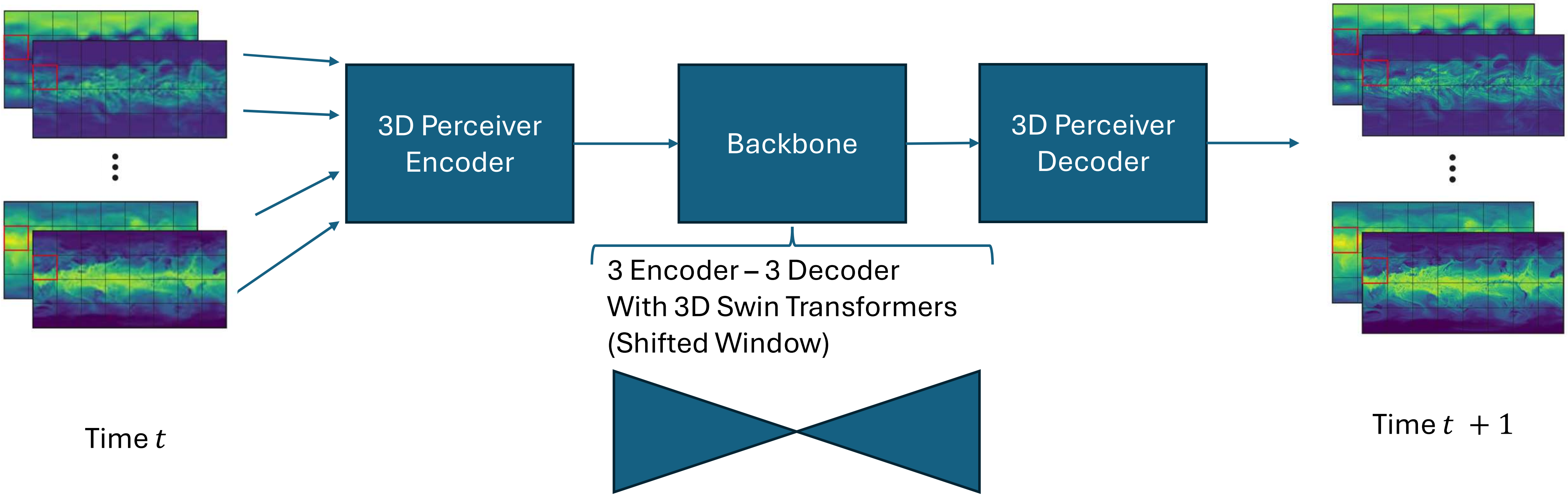
| Name | Timeframe |
|---|---|
| HRES-0.25 | 2016 - 2021 |
| HRES-0.1 | 2016 – 2022 |
| CAMSRA | 2003 – 2021 |
| CAMS Analysis | Oct 2017 – May 2022 |

# Problem statement

- Observed state of atmosphere $X^t$ at time t: $X^t = V \times H \times W$
- Discretization of time $t$
- $V$ number of variables
- $V$ can be split in $V_S$ (surface) and $V_A$ (atmosphere)
- $H, W$ number of latitude/longitude coordinates
- Goal:
  - $\Phi\colon (\mathrm{X}^{\mathrm{t}-1},\ \mathrm{X}^{\mathrm{t}}) \rightarrow \hat{X}^{t+1}$
  - Recursively for future

# Architecture

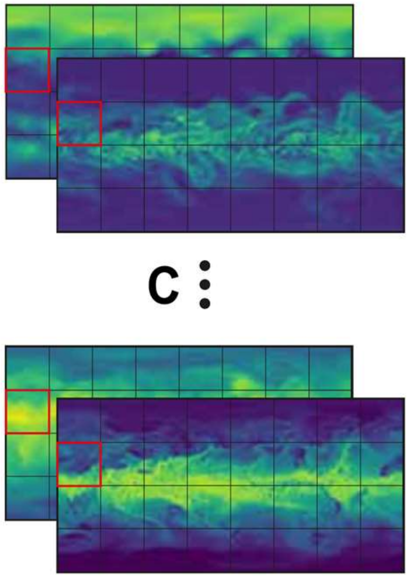
- Image processing – not graph



# 3D Perceiver Encoder (1)

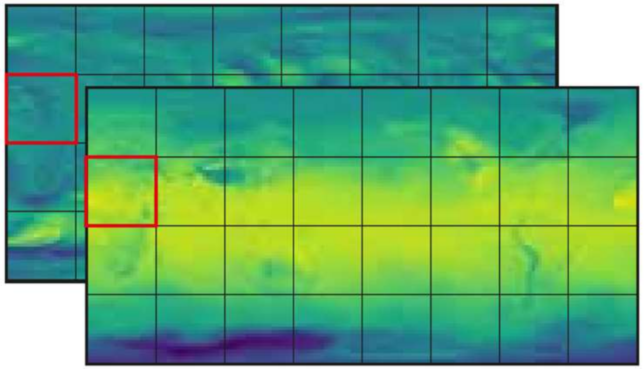
- All variables as $H \times W$ images at time $t$ and $t-1$

Input Atmospheric Variables



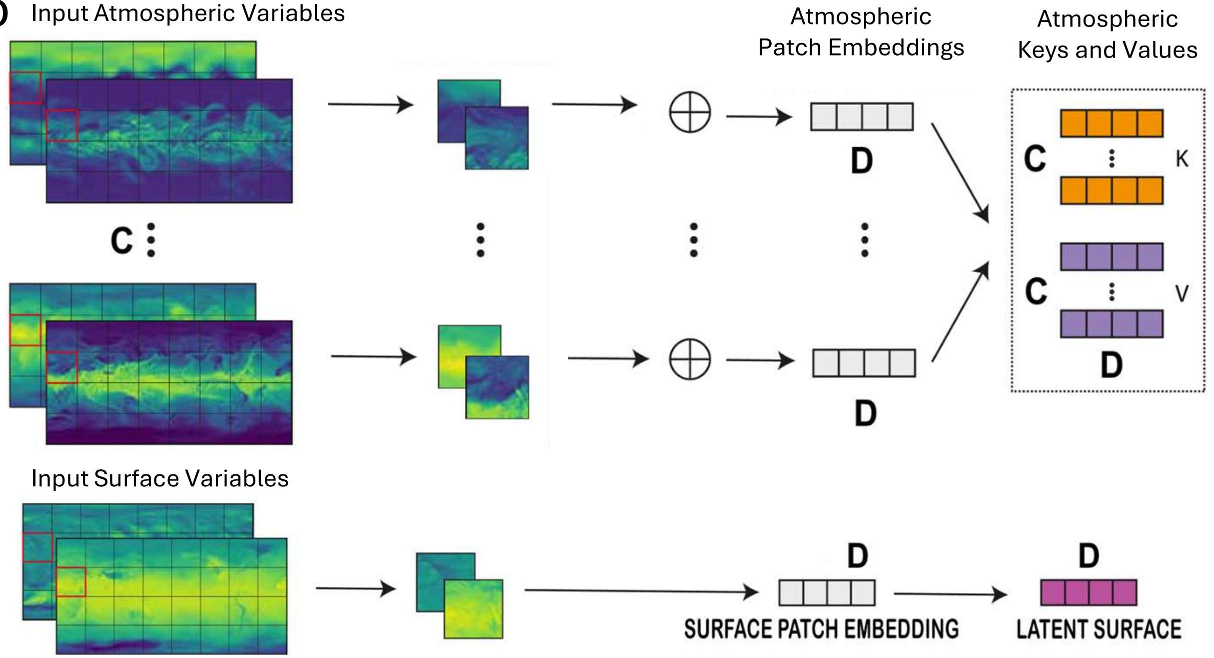

- Static variabel (from ERA5):
  - (1) geopotential at the surface
  - (2) land-sea mask
  - (3) soil-type mask

Input Surface Variables



# 3D Perceiver Encoder (2)

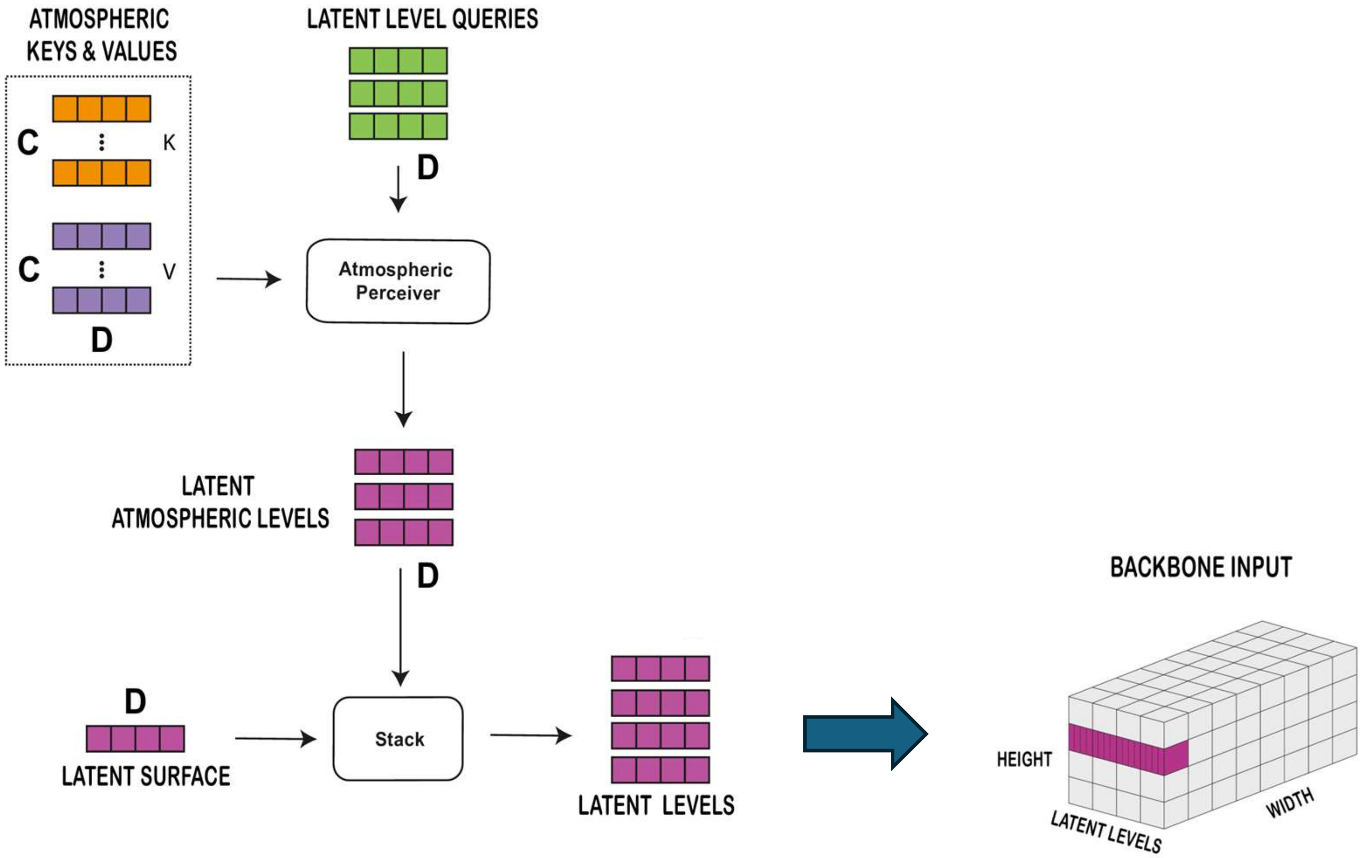
- Level Embeddings:
  - Split $H \times W$ images in $P \times P$ patches
  - Map patches at each level into vectors $\mathbb{R}^D$ via linear layer
  - $V_S \times T \times P \times P \ \rightarrow 1 \times D$
  - Tag vector with level encoding
  - Stack levels




(simplified version)

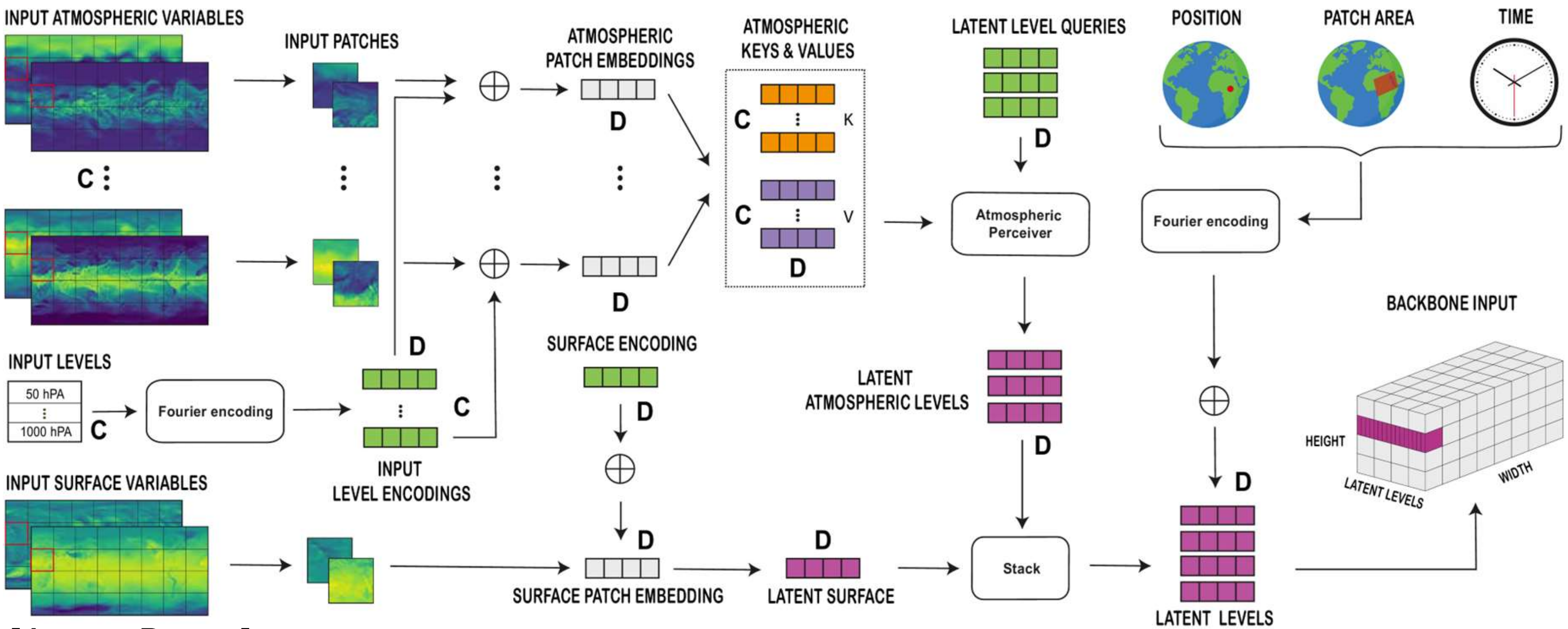
# 3D Perceiver Encoder (3)

- Level aggregation



# 3D Perceiver Encoder (4)

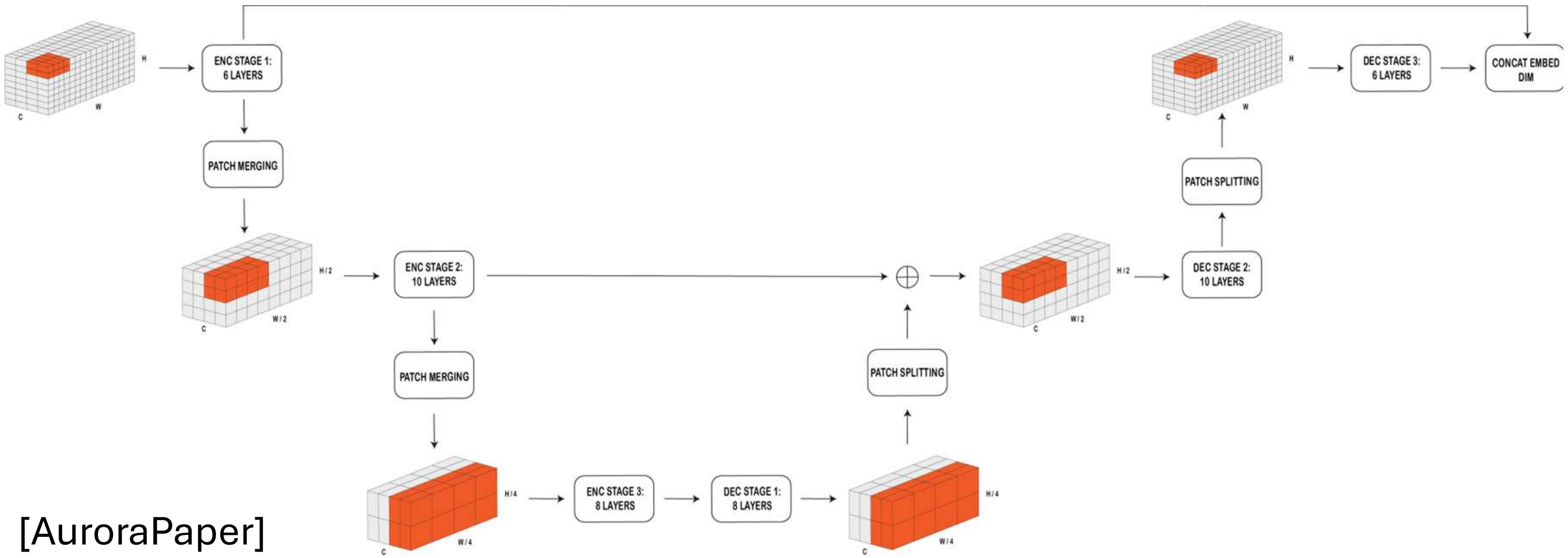
- Further aggregation for full backbone input tensor




[AuroraPaper]

# Backbone

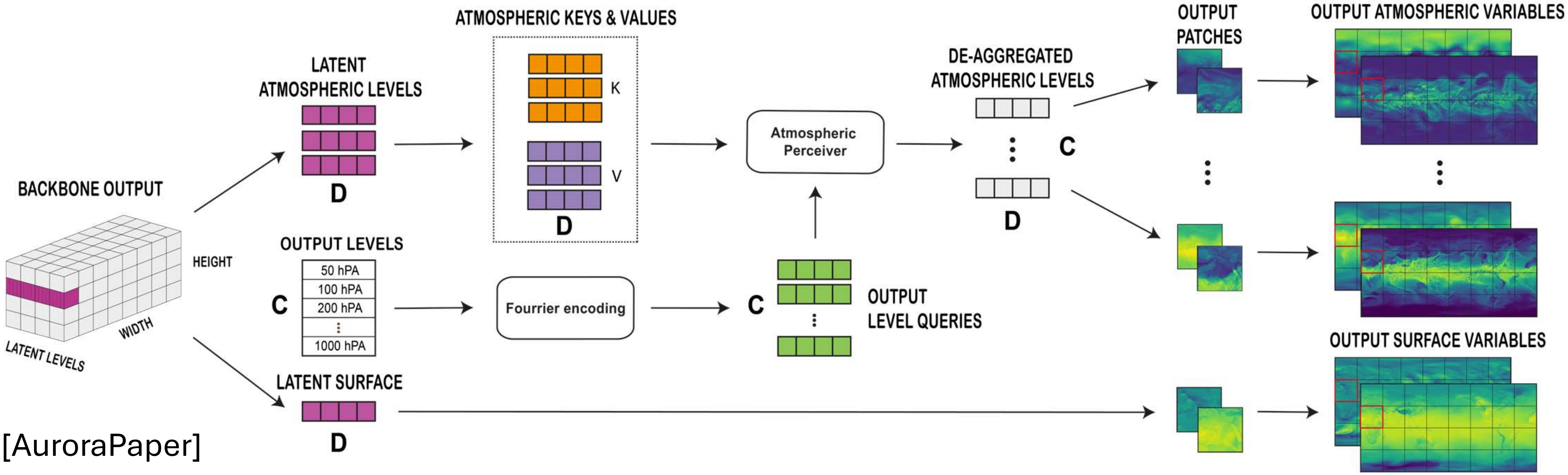
- En/Decoder with 3 stages each halving/doubling the resolution
- Each layer: 3D Swin Transformer layer
  - Dot production attention, attention shift between layers




[AuroraPaper]

# 3D Perceiver Decoder

- Output of encoder – backbone input
- Mirror of encoder




[AuroraPaper]

# Training stages

- Pretraining
  - (1) pretraining,
- Fine tuning
  - (2) short lead-time fine-tuning of pretrained weights
  - (3) long lead-time (rollout) fine-tuning

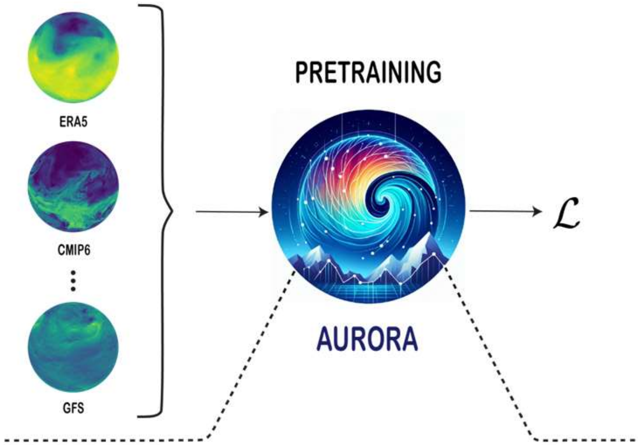
**Pretraining**

- Give network snapshot time $t$
- Get prediction on time $t + 1$ compare

➔Minimise Loss

- 2-3 week on 32 a100 GPUs

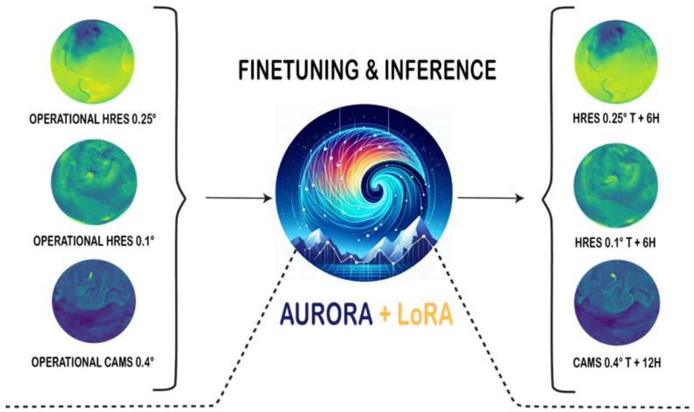
Data: diverse big data




**Finetuning**

- Short lead-time fine tuning:
  - Minimise loss
- Rollout Fine Tuning
  - LoRA (Low Rank Adaptation)
  - Adapt Backbone weights
    5 days on 8 a100 GPUs

Data: limited and sparse



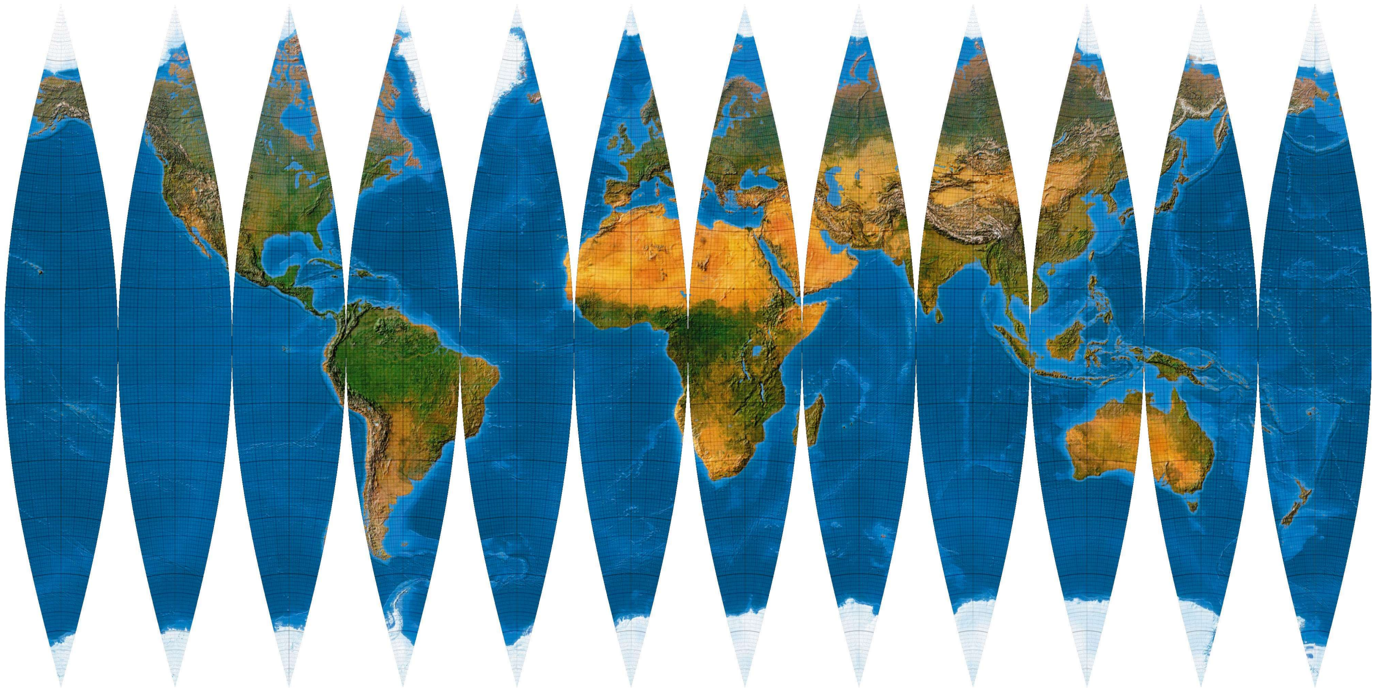
# Loss – Mean absolute error

- Training objective $\mathcal{L}(\hat{X}^t, X^t)$
- Predicted state $\hat{X}^t = (\hat{S}^t, \hat{A}^t)$
- Ground truth state $X^t = (S^t, A^t)$
- Weight associated with surface-level variable $k$: $w_k^S$
- Weight atmospheric variable k at pressure level c: $w_{k,c}^A$
- $\alpha, \beta, \gamma$ – predefined weights

$$\mathcal{L}(\hat{X}^t, X^t) = \frac{\gamma}{V_S + V_A}\left[\alpha\left(\sum_{k=1}^{V_S}\frac{w_k^S}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|\hat{S}_{k,i,j}^t - S_{k,i,j}^t|\right)\right.$$
$$\left. + \beta\left(\sum_{k=1}^{V_A}\frac{1}{C \times H \times W}\sum_{c=1}^{C} w_{k,c}^A \sum_{i=1}^{H}\sum_{j=1}^{W}|\hat{A}_{k,c,i,j}^t - A_{k,c,i,j}^t|\right)\right]$$

# Validation

- Latitude weighting



[Atlas]

$$w(i) = \frac{\cos\big(lat(i)\big)}{\frac{1}{H}\sum_{i'=1}^{H}\cos\big(lat(i')\big)}$$

# RMSE

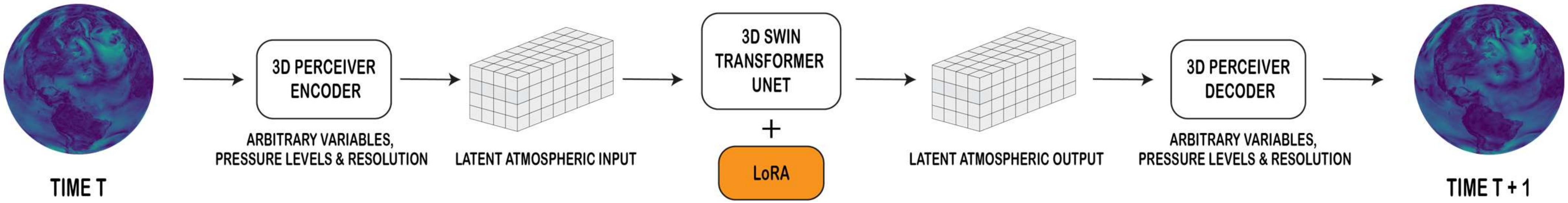
- Measure error between predicitons and “ground truth”

$$\mathrm{RMSE} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} w(i)(\hat{X}_{i,j}^{t} - X_{i,j}^{t})^2},$$

- $t$ sample datasets
- $i, j$ index or latitude and longitude of each image
- $w(i)$ weighing factor

# Low Rank Adaptation

- Knowledge:
  over-parametrized model have low intrinsic dimension [LoRA]
- Advantages:
  - easy switching between models
  - Efficient since no gradients required
  - No inference latency in deployment

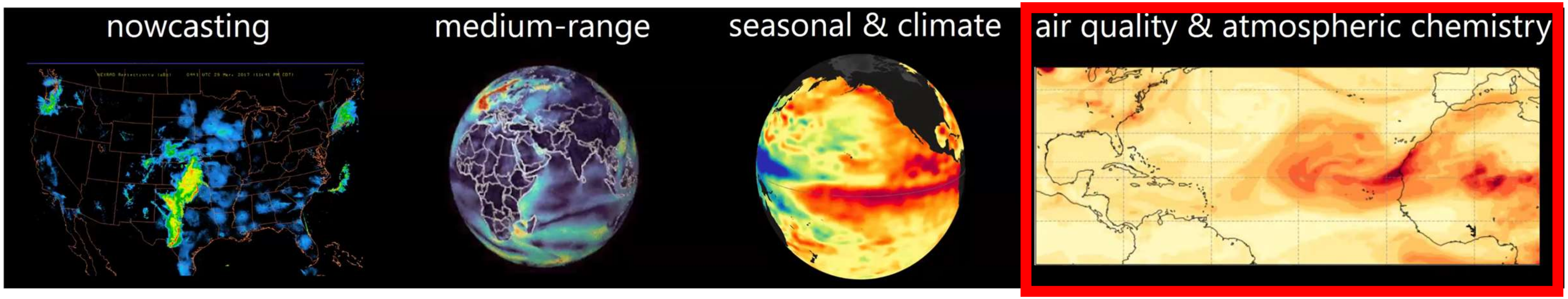


# Low Rank Adaptation: How it works

- $W_0$ is initial weight (~ full rank) $W_0 \in \mathbb{R}^{d\times k}$, $\mathrm{W}_0$ is frozen
- Initialisation: random Gaussian $\mathrm{A}$, zero for $\mathrm{B}$
- Constraint Update:
  - $W_0 + \Delta W = W_0 + BA,$
  - With $\mathrm{B} \in \mathbb{R}^{d\times r}, \mathrm{A} \in \mathbb{R}^{r\times k}$, rank $\mathrm{r} \ll \min(\mathrm{d}, \mathrm{k})$
- New Forward pass
- $\mathrm{h} = \mathrm{W}_0\mathrm{x} + \Delta W x = W_0 x + B\,A\,x = (W_0 + B\,A)\,x$

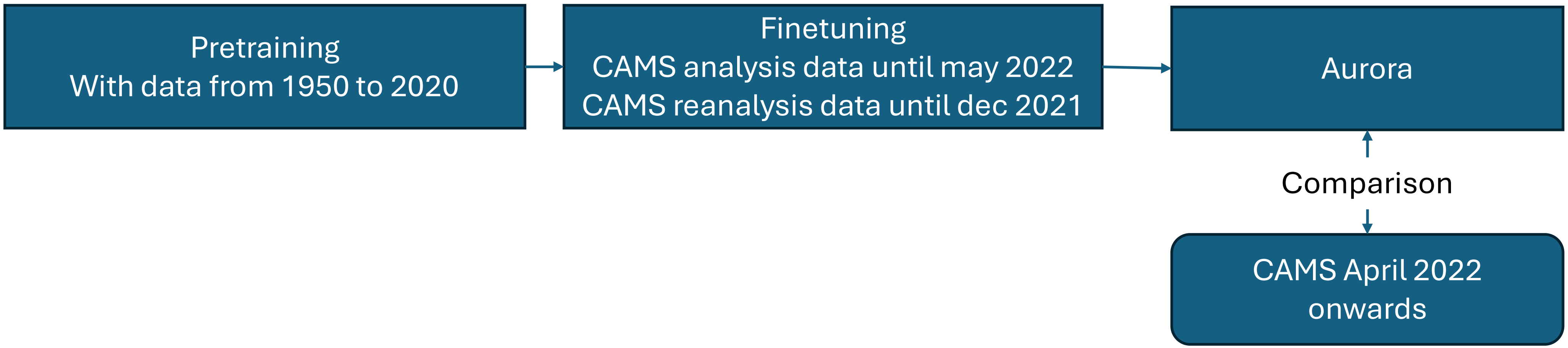
# Scenario 1: Fast prediction of atmospheric chemistry and air pollution

- Meteorological variables, air pollution – concentration values with
  - Air pollution strongly depended to anthropogenic factors (COVID)



# Fast prediction of atmospheric chemistry and air pollution

- Finetuning - data
  - CAMS analysis data
    - Very scarce and Non-stationary
    - Large dynamic range
    - Highly heterogeneous often extremely sparse and skewed



# Setup differences

**CAMS**

- Emissions data as input
    - Natural factors (wildfires, vegetation etc.)
    - Anthropogenic factors (vehicle combustion, energy production)

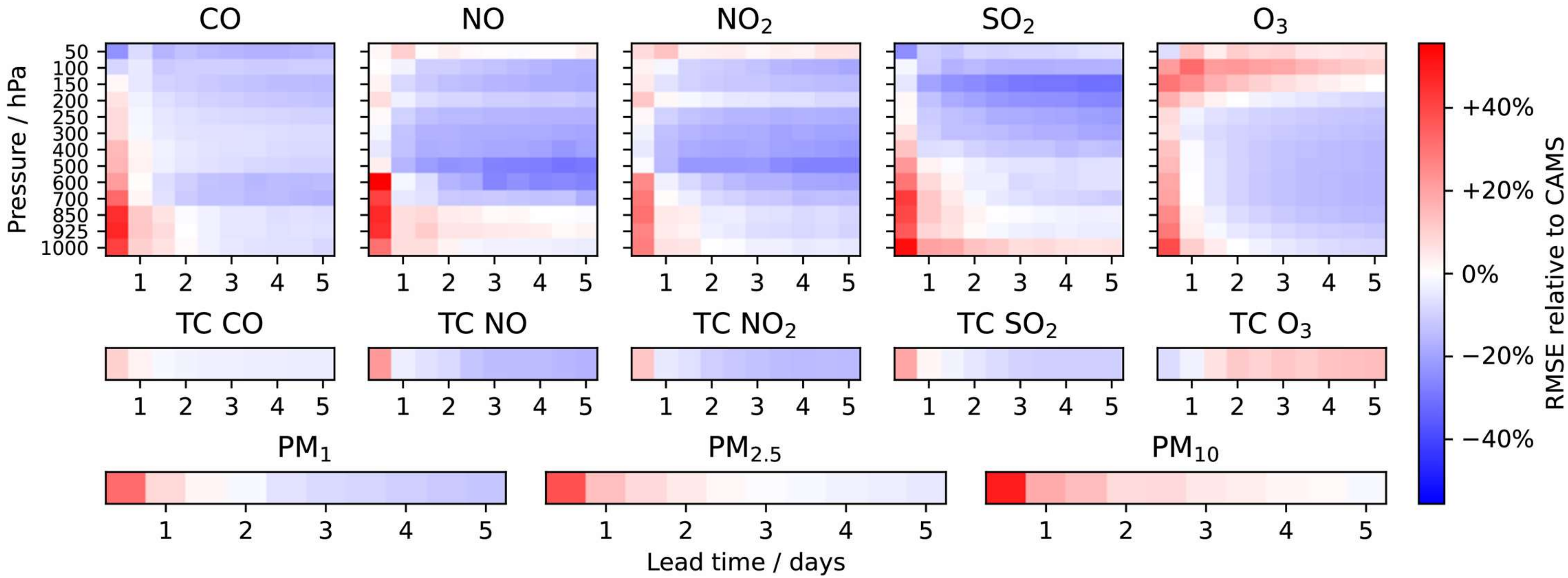
**Aurora**

- No emissions data as input except
- Static variable fixed across all times and experiments

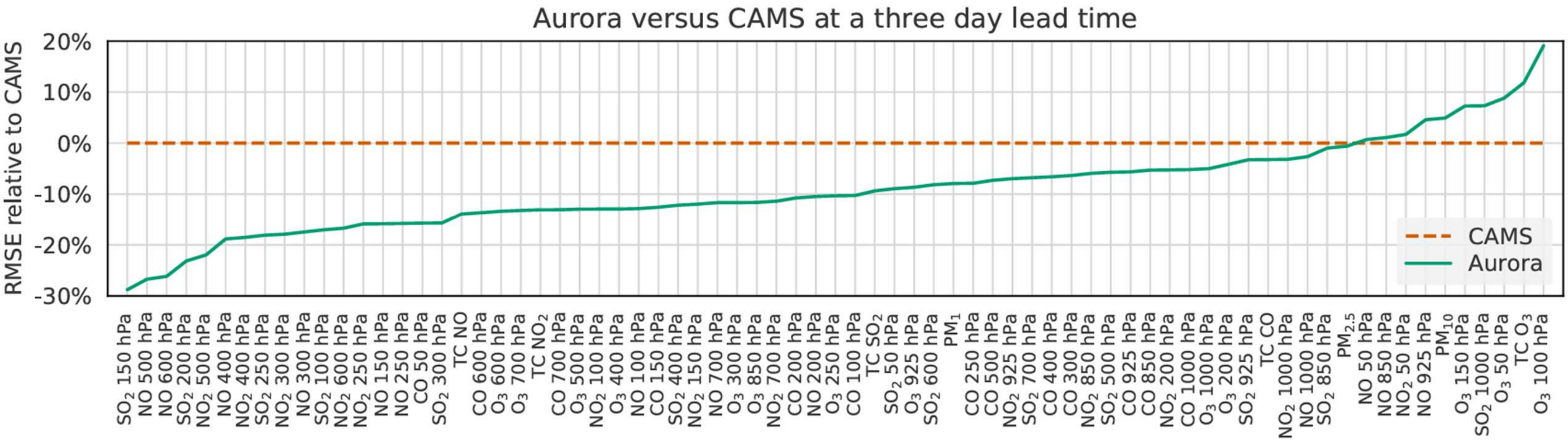
➔learning implicitly from historical data which is affected by natural and anthropogenic factors

# Results b

- Competitive results



# Results c



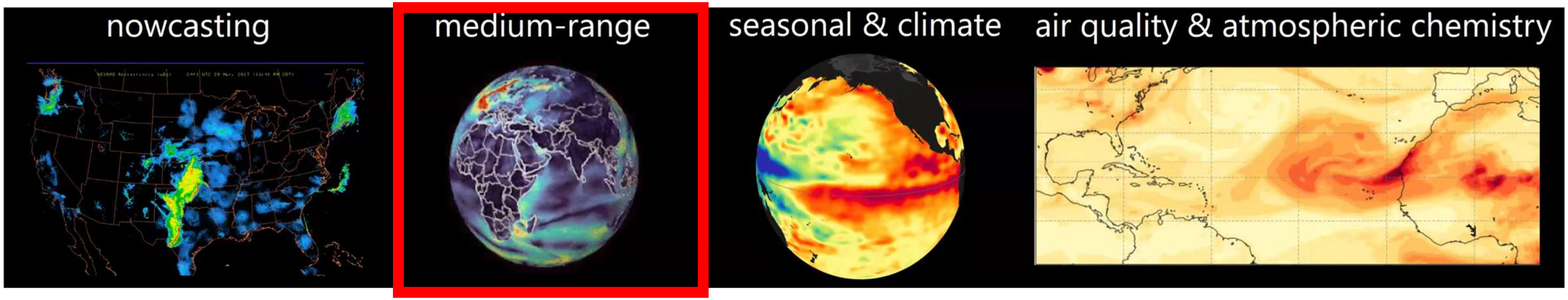
Aurora versus CAMS at a three day lead time
RMSE relative to CAMS
20%
10%
0%
-10%
-20%
-30%
CAMS
Aurora
$SO_2$ 150 hPa
NO 500 hPa
NO 600 hPa
$SO_2$ 200 hPa
$NO_2$ 500 hPa
NO 400 hPa
$NO_2$ 400 hPa
$SO_2$ 250 hPa
$NO_2$ 300 hPa
NO 300 hPa
$SO_2$ 100 hPa
$NO_2$ 600 hPa
$NO_2$ 250 hPa
NO 150 hPa
NO 250 hPa
CO 50 hPa
$SO_2$ 300 hPa
TC NO
CO 600 hPa
$O_3$ 600 hPa
$O_3$ 700 hPa
TC $NO_2$
CO 700 hPa
$O_3$ 500 hPa
$NO_2$ 100 hPa
$O_3$ 400 hPa
NO 100 hPa
CO 150 hPa
$SO_2$ 400 hPa
$NO_2$ 150 hPa
NO 700 hPa
$O_3$ 300 hPa
$O_3$ 850 hPa
$NO_2$ 700 hPa
CO 200 hPa
NO 200 hPa
$O_3$ 250 hPa
CO 100 hPa
TC $SO_2$
$SO_2$ 50 hPa
$O_3$ 925 hPa
$SO_2$ 600 hPa
$PM_1$
CO 250 hPa
CO 500 hPa
$NO_2$ 925 hPa
$SO_2$ 700 hPa
CO 400 hPa
CO 300 hPa
$NO_2$ 850 hPa
$SO_2$ 500 hPa
CO 925 hPa
CO 850 hPa
$NO_2$ 200 hPa
CO 1000 hPa
$O_3$ 1000 hPa
$O_3$ 200 hPa
$SO_2$ 925 hPa
TC CO
$NO_2$ 1000 hPa
NO 1000 hPa
$SO_2$ 850 hPa
$PM_{2.5}$
NO 50 hPa
NO 850 hPa
$NO_2$ 50 hPa
NO 925 hPa
$PM_{10}$
$O_3$ 150 hPa
$SO_2$ 1000 hPa
$O_3$ 50 hPa
TC $O_3$
$O_3$ 100 hPa

# Summary of results

- Aurora competitive with CAMS (within 20 % RMSE) on 94  % of targets
- Match or outperform on 74 % of all targets

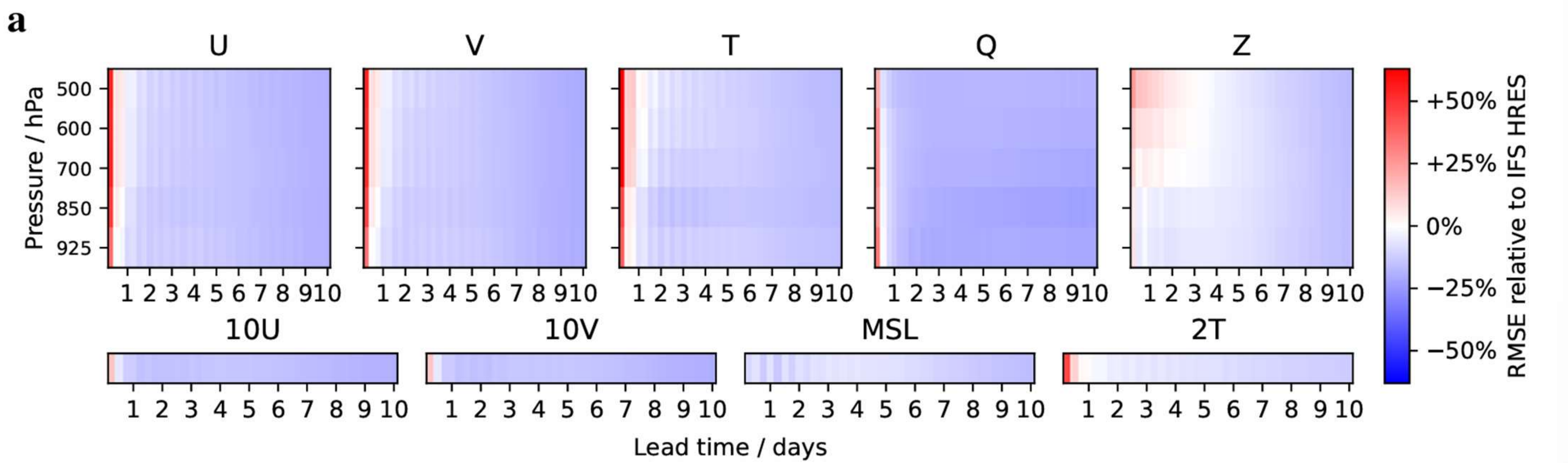
➔ Problem low atmosphere, due to anthropogenic factors which are not accounted in Aurora

# Skillful operational weather forecasting at 0.1∘ resolution (11km)



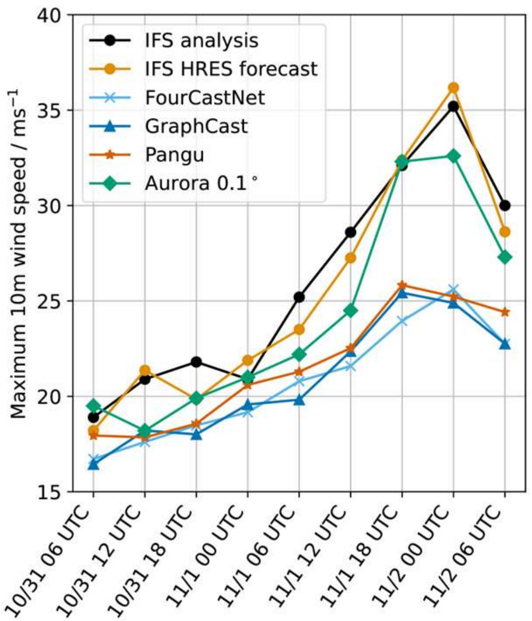
# Skillful operational weather forecasting at 0.1∘ resolution (11km)

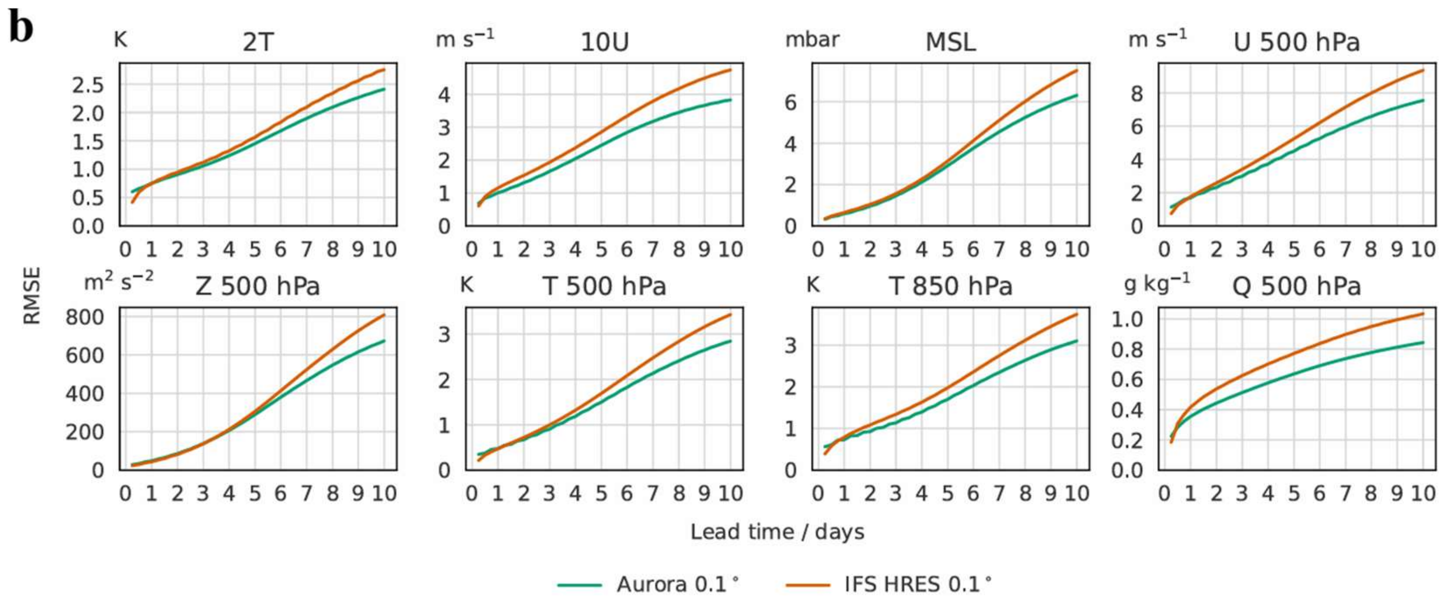
- IFS-HRES – comparison (state-of-the art)
- Better results for long lead-times



# Case Study: Strom Ciaran

- IFS-HRES – comparison (state-of-the art)
- Better results overall



# Comparison Computation

**IFS**

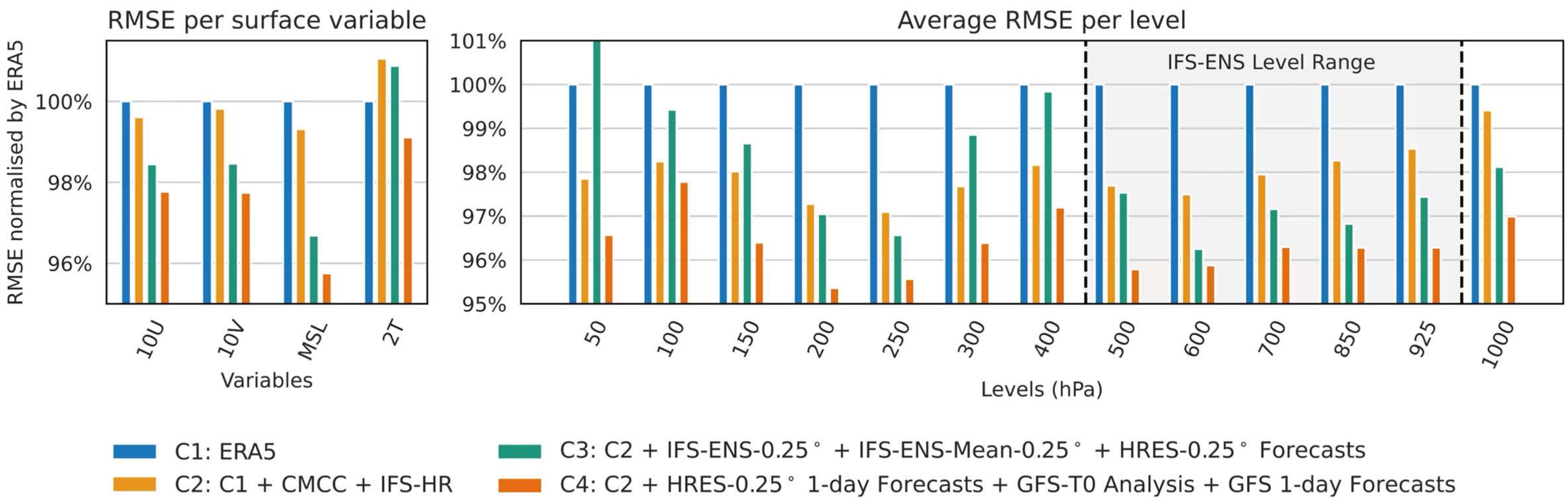
- 10-day forecast takes approximately 65 minutes on 352 high-end CPU nodes with 36 cores each

**Aurora**

- 1.1s per hour lead time on a single A100 GPU,
- ➔ 4,4 minutes for 10-day forecast
- roughly a ×5,000 speedup over IFS

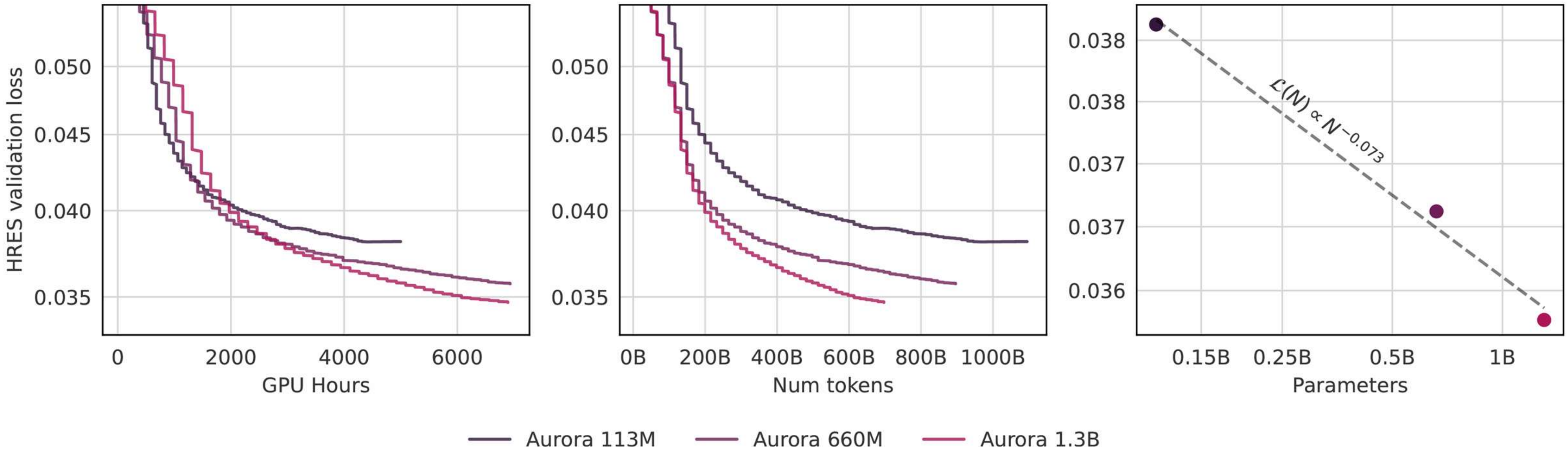
# Diverse inputs

- More diversity on input data improved results



# Model Scaling Pretraining

- Bigger models yield lower validation loss
- 5 % reduction in loss for every doubling of model



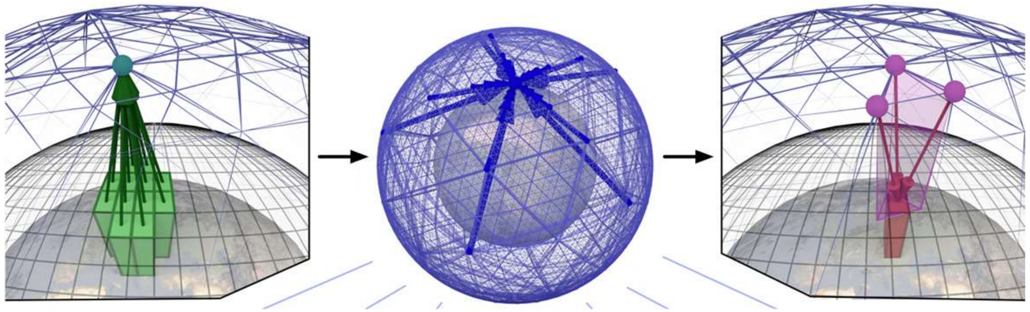
# Excurse: Graphcast Resolution: 0.25 °

**Graphcast**

- World Represented as Graph
- Trained on HRES and ERA5
- Single Task

**Aurora**

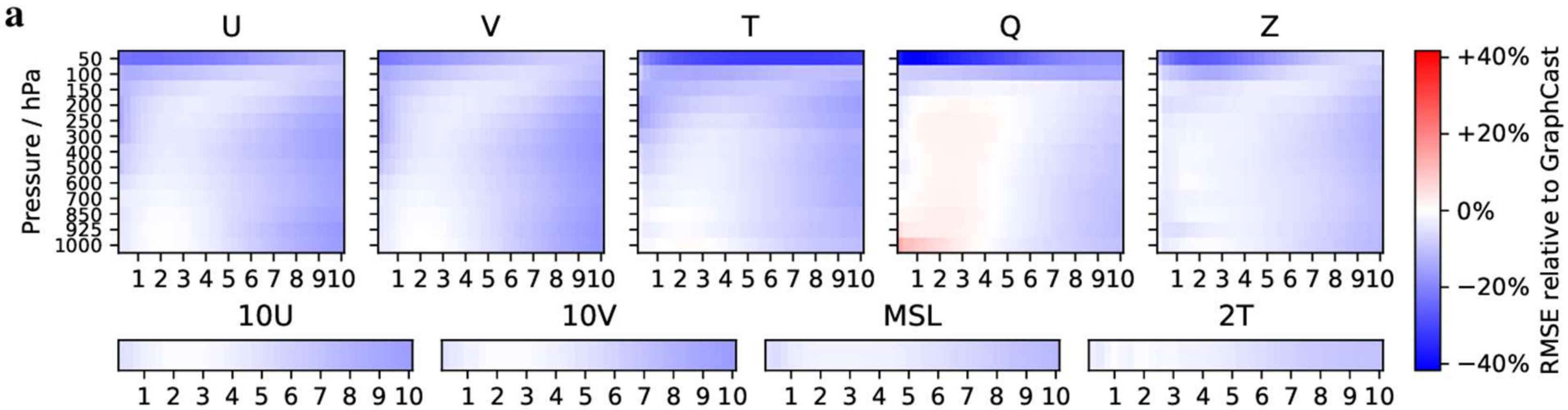
- World represented as images
- Trained on diverse data
- Multiple Tasks with Finetuning



[Graphcast]

# Comparison to Graphcast

- Match or outperform at 94 % targets
- Biggest gains:
  - Upper atmosphere
  - High lead times



# Conclusion

- First time AI models beat NWP
- Only deterministic forecasts
  - ➔ solution: ensemble of models
- No usage of local high-resolution datasets (only global datasets)
  - More potential
- Robustness and verification have to be improved to be used operational

# Personal takeaway

- Huge progress in terms of replacing NWP
- Potential to provide high-res predictions faster
- Better results potential with better GPUs with more memory
- An enough big and complex model can capture everything

# Further Information

Predecessor:

- **ClimaX: A foundation model for weather and climate**

Try it out yourself:

- GitHub - microsoft/aurora: Implementation of the Aurora model for atmospheric forecasting

# Sources


[Aurorapaper][2405.13063] Aurora: A Foundation Model of the Atmosphere

[ButterflyImage] Butterfly Storm Insect - Free photo on Pixabay

[auroraYoutubeVideo] https://youtu.be/OqHlCXcibrg?si=y9loQWIloyzuUxia&t=664

[objectDetection] Object-detection-Real-world-applications-and-benefits.png (1500×1000)

[Protein] https://frontlinegenomics.com/alphafold-2-protein-structure-prediction-software-for-all/

[AIFoundationModels] https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf

[ECMWFLogo] https://th.bing.com/th/id/R.914f4a4d81996063cbfcde7f9c622dc3?rik=uw%2b4z%2bFNY%2bvN4g&riu=http%3a%2f%2fwww.ecmwf.int%2fsites%2fdefault%2ffiles%2fECMWF_Master_Logo_RGB_nostrap.png&ehk=XrDRtwu3snQDAJxRu%2fq3J0%2fVQrixEaxFqvF8E7RHK9s%3d&risl=&pid=ImgRaw&r=0

[NWS] https://th.bing.com/th/id/OIP.zDY1VZwJtBV5Fsr2KYrJtQHaHa?rs=1&pid=ImgDetMain

[Atlas]Pin page

[Climax]2301.10343

[LoRA] https://www.semanticscholar.org/paper/LoRA%3A-Low-Rank-Adaptation-of-Large-Language-Models-Hu-Shen/a8ca46b171467ceb2d7652fbfb67fe701ad86092

[Graphcast] [2212.12794] GraphCast: Learning skillful medium-range global weather forecasting