spcl.ethz.ch
@spcl
@spcl_eth

T. HOEFLER

AI for Climate Data Generation, Assimilation, and Modelin

with contributions by B. Stevens, L. Huang, O. Fuhrer, P. Dueben, S. Schemm, and the whole SPCL team, CSCS, E

Keynote at HPC-CH Meeting, Lugano, Switzerland, Oct. 2024

Seeing hazy skies? Seattle area's air quality was among worst in US Friday morning

Seattle, Kirkland and Bothell were seeing the worst air quality in western Washington the morning after F

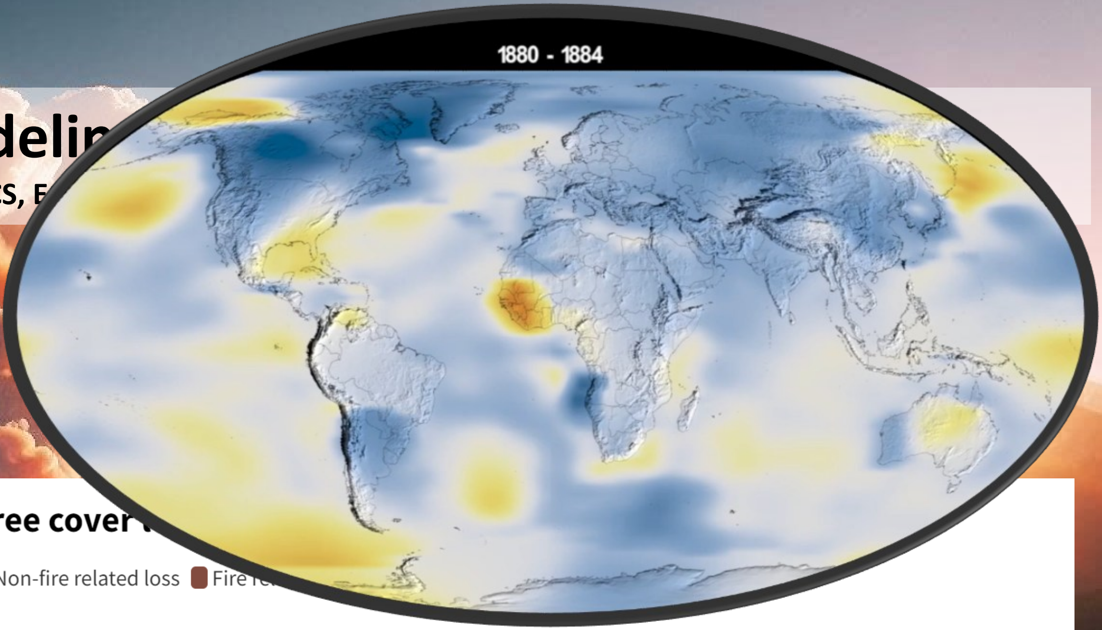
Author: Alex Didion
Published: 7:08 AM PDT July 5, 2024
Updated: 6:43 PM PDT July 5, 2024

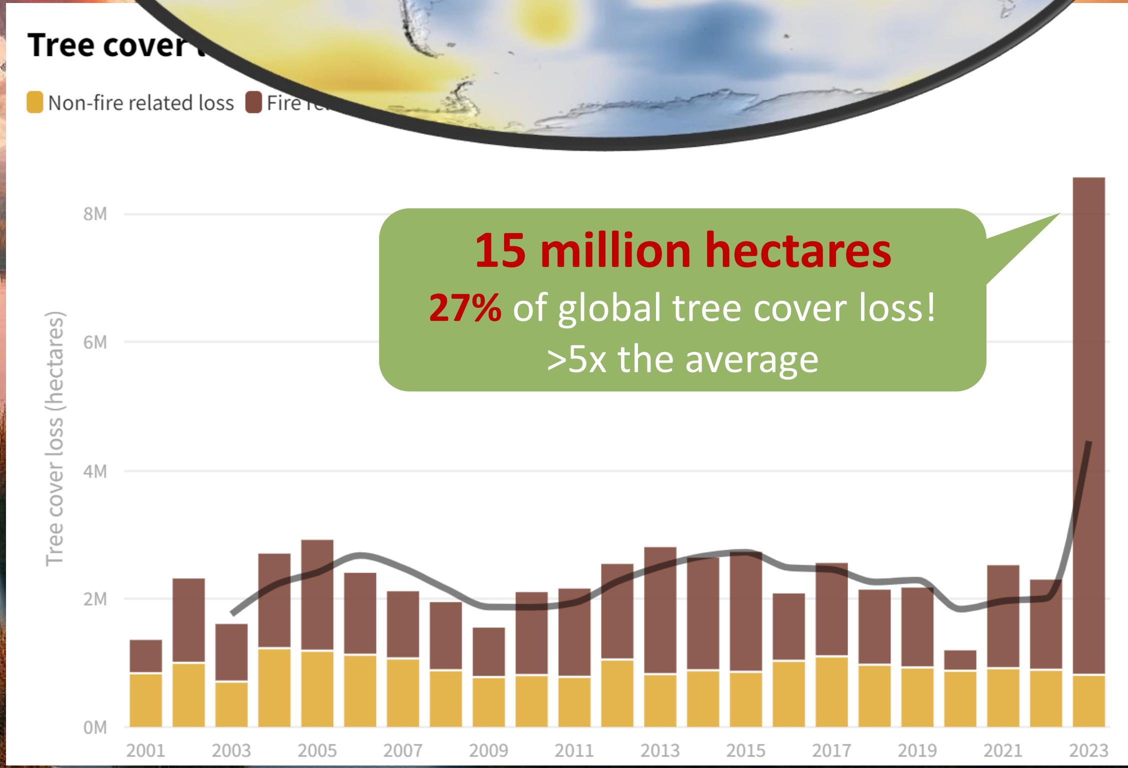
MYNORTHWEST WEATHER

Brace yourself for weeks of poor air quality in 2024, Seattle residents

Apr 16, 2024, 3:29 PM | Updated: Apr 17, 2024, 12:32 pm

Smoke from wildfires fills the air along Alaskan Way on September 12, 2020 in Seattle, Washington. (Photo by Lindsey Wasson, Getty Images)





T. HOEFLER

AI for Climate Data Generation, Assimilation, and Modeling

with contributions by B. Stevens, L. Huang, O. Fuhrer, P. Dueben, S. Schemm, and the whole SPCL team, CSCS, ECMWF, the EVE Summit Attendees, industry friends, and others

Keynote at HPC-CH Meeting, Lugano, Switzerland, Oct. 2024

"Climate simulation is basically impossible today."

"Predicting the average temperature is possible. However, the world doesn't care about average. You care about your own region." (Huang, Nov. 2023)

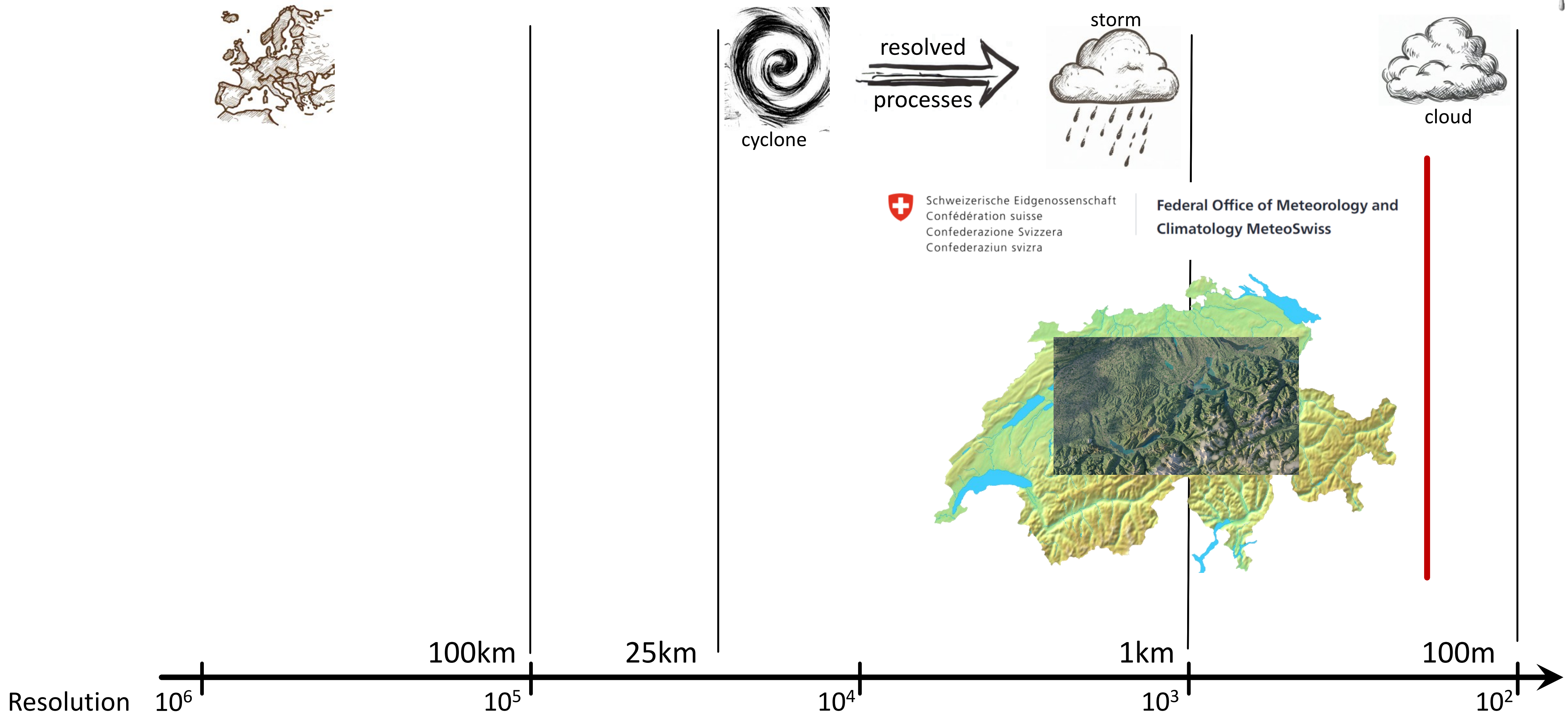
What can WE do about this?

Alps, the 6th fastest publicly known supercomputer on the planet - #2 in Europe (maybe soon #1!)

10,752 H100 GPUs and Grace CPUs	~10 Exaflop (10^{18}) BF16 performance	~40 years of humanity-ops/s
~1 PiB HBM3 + ~1.4 GiB LPDDR5	~48 PiB/s memory BW	~2M full wikipedias/s (24 GiB)
200 Gbps Slingshot interconnect	~250 TiB/s network bandwidth	~2x the total Internet bandwidth

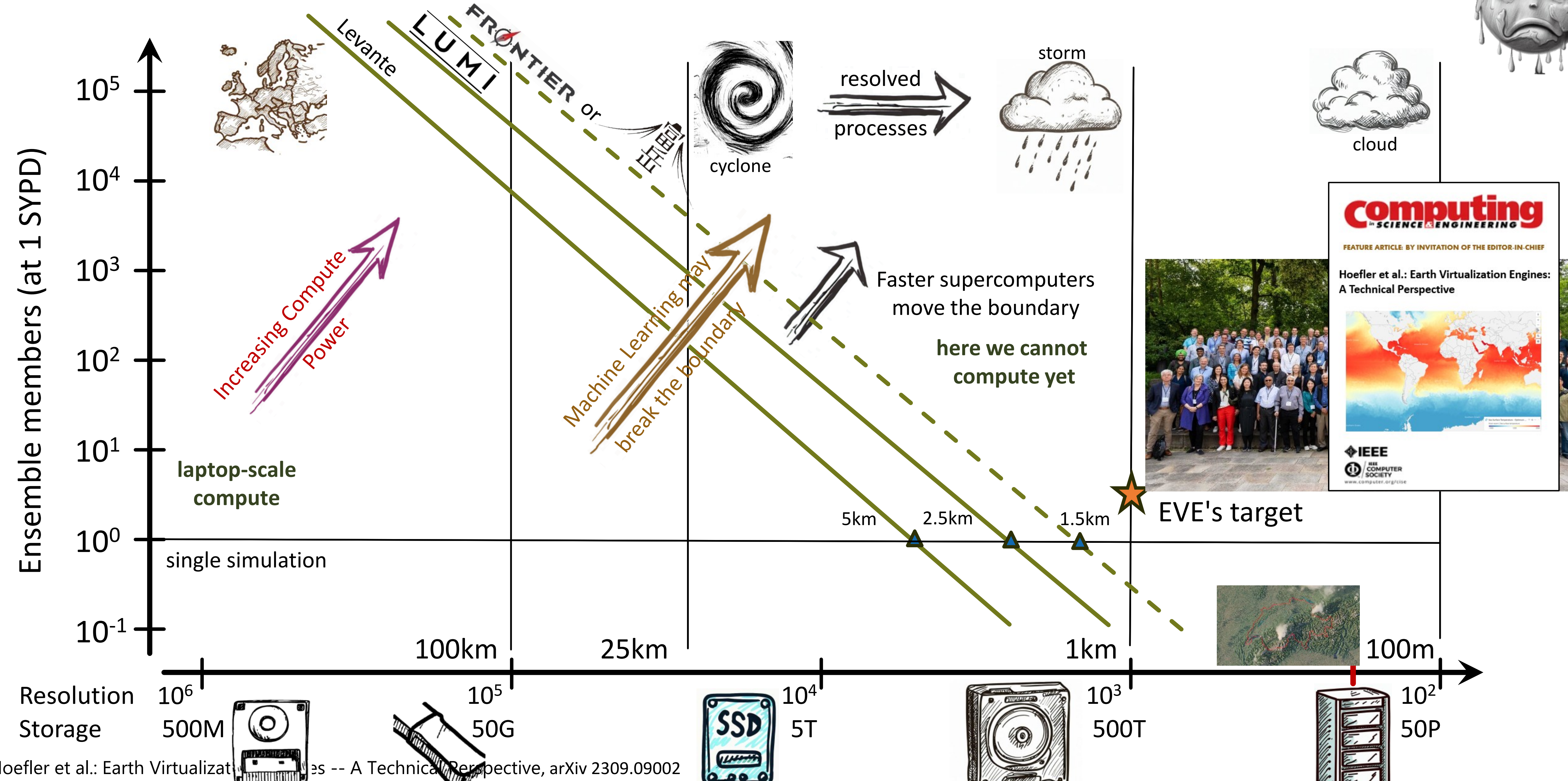
If we can’t do it with this machine – who else could?

Climate prediction is extremely demanding ("impossible" – decades long)



Hoefler et al.: Earth Virtualization Engines -- A Technical Perspective, arXiv 2309.09002

Climate prediction is extremely demanding (“impossible” – decades long)



Hoefler et al.: Earth Virtualization Engines -- A Technical Perspective, arXiv 2309.09002

The three pillars of AI in Climate Sciences

Data

- Unstructured observation data
- Structured simulation data

Combine both to train models

- Learn physics and data-driven prediction

Compute

- AI models require accelerated high-performance computing for training

Accelerate AI computations

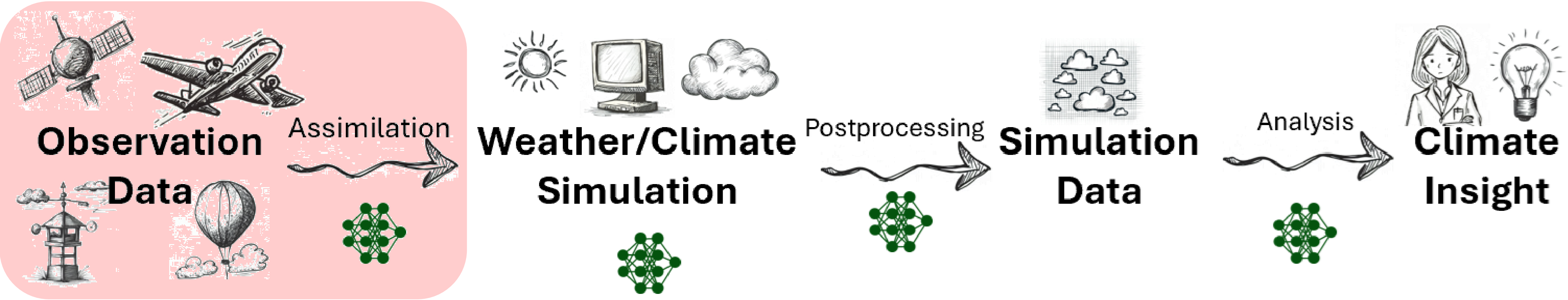
- Re-use infrastructure from LLMs and related generative AI methods (GNN, CNN)

Models

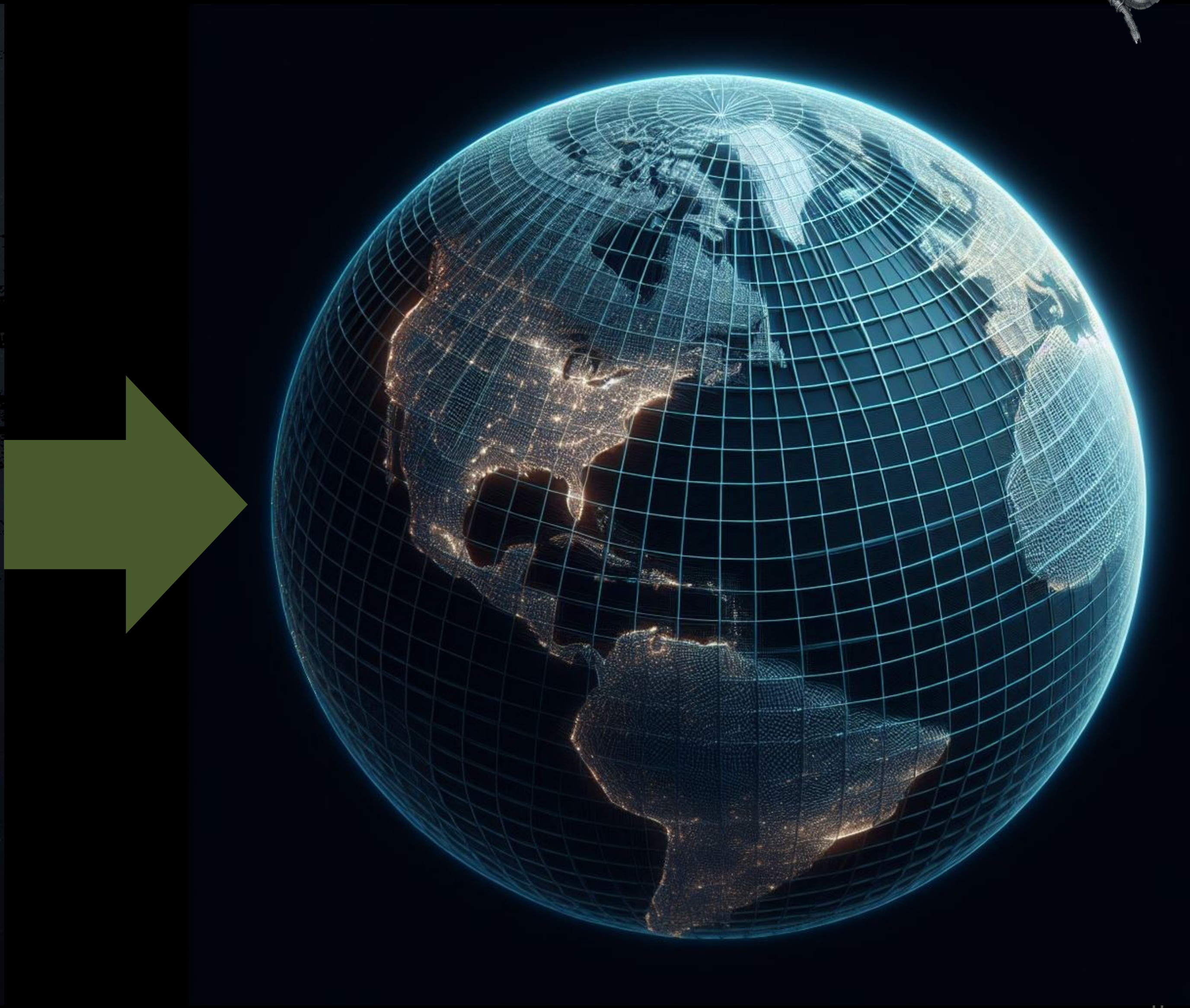
- Models need to provide the right structural bias/prior

Develop better AI methods

- Step 1: use generative AI models: tformer, CNN, GNN, Diffusion, etc.
- Step 2: use automatically parameterized physics-based models encoding equations

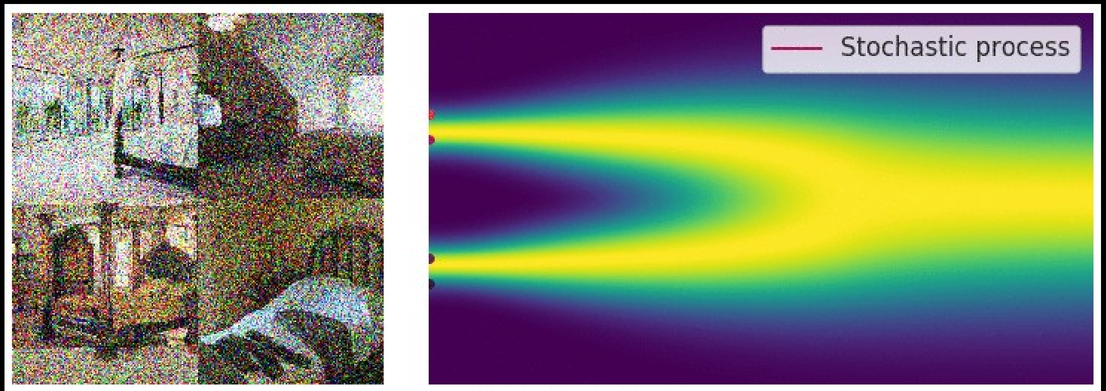


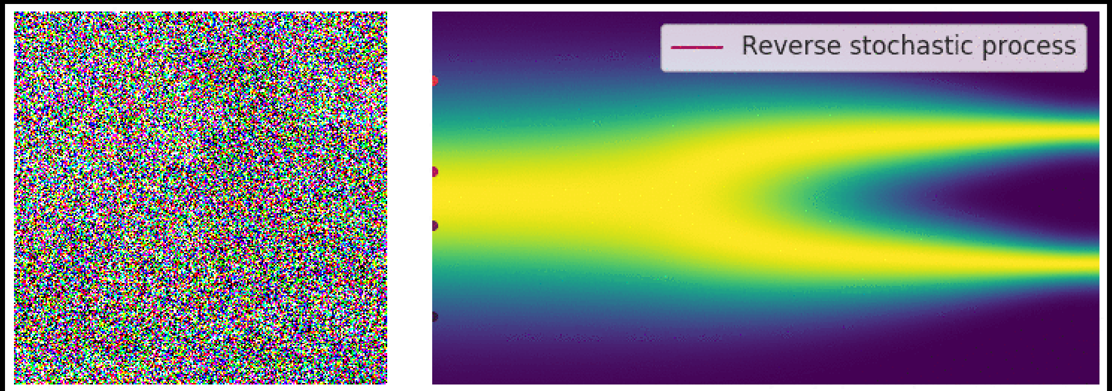
Diffusion-based Data Assimilation



Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D. Dueben, Torsten Hoefler: DiffDA: a Diffusion model for weather-scale Data Assimilation arXiv:2401.05932 + ICML’24, Jul. 2024

How do diffusion models work?





source: medium.com – Vadim Titko

DiffDA: A Diffusion Model for Weather-scale Data Assimilation

DiffDA: A Diffusion Model for Weather-scale Data Assimilation

Langwen Huang [1] Lukas Gianinazzi [1] Yuejiang Yu [1] Peter D. Dueben [2] Torsten Hoefler [1]

Abstract

The generation of initial conditions via accurate data assimilation is crucial for weather forecasting and climate modeling. We propose DiffDA as a denoising diffusion model capable of assimilating atmospheric variables using predicted states and sparse observations. Exploiting the similarity between a weather forecasting model and a denoising diffusion model dedicated to weather applications, we adapt the pretrained GraphCast neural network as the backbone of the diffusion model. Through experiments based on simulated observations from the ERA5 reanalysis dataset, our method can produce assimilated global atmospheric data consistent with observations at 0.25° (≈30km) resolution globally. This marks the highest resolution achieved by ML data assimilation models. The experiments also show that the initial conditions assimilated from sparse observations (less than 0.96% of gridded data) and 48-hour forecast can be used for forecast models with a loss of lead time of at most 24 hours compared

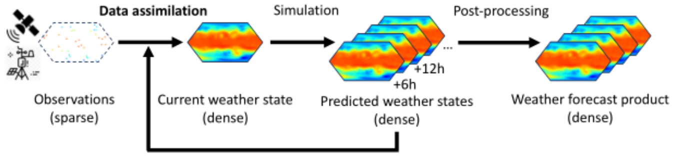
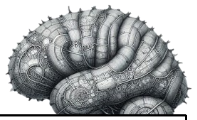
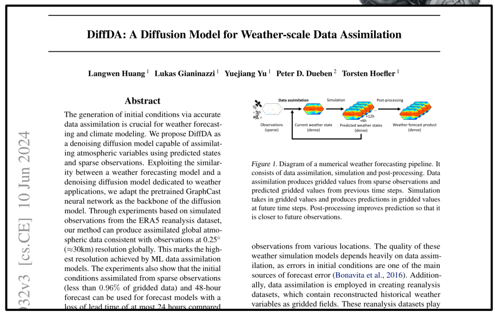


Figure 1. Diagram of a numerical weather forecasting pipeline. It consists of data assimilation, simulation and post-processing. Data assimilation produces gridded values from sparse observations and predicted gridded values from previous time steps. Simulation takes in gridded values and produces predictions in gridded values at future time steps. Post-processing improves prediction so that it is closer to future observations.

observations from various locations. The quality of these weather simulation models depends heavily on data assimilation, as errors in initial conditions are one of the main sources of forecast error (Bonavita et al., 2016). Additionally, data assimilation is employed in creating reanalysis datasets, which contain reconstructed historical weather variables as gridded fields. These reanalysis datasets play

Case study: Assimilated 2m temperature at 2022-01-03 06z





DiffDA: A Diffusion Model for Weather-scale Data Assimilation

Langwen Huang [1] Lukas Gianinazzi [1] Yuejiang Yu [1] Peter D. Dueben [2] Torsten Hoefler [1]

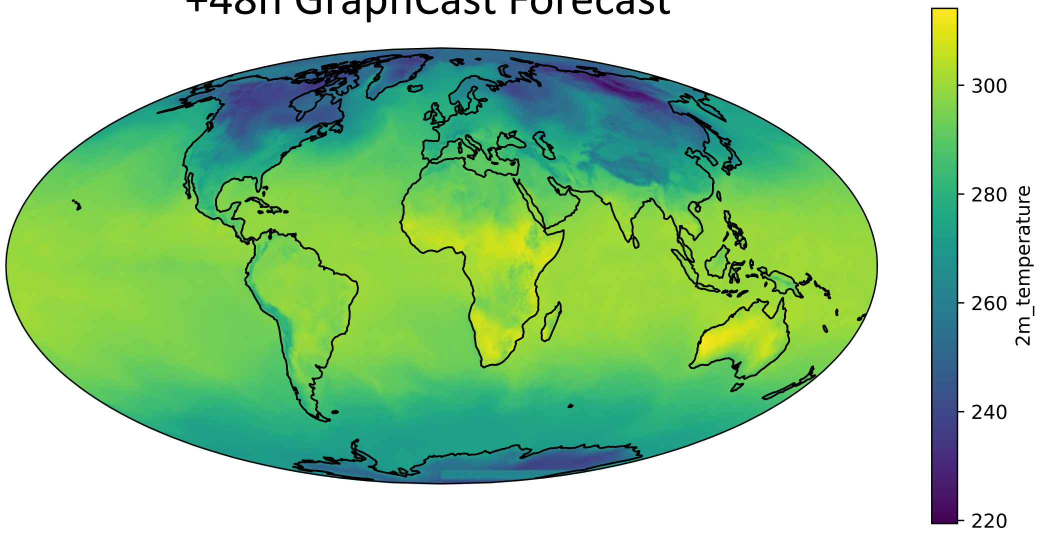
Abstract

The generation of initial conditions via accurate data assimilation is crucial for weather forecasting and climate modeling. We propose DiffDA as a denoising diffusion model capable of assimilating atmospheric variables using predicted states and sparse observations. Exploiting the similarity between a weather forecasting model and a denoising diffusion model dedicated to weather applications, we adapt the pretrained GraphCast neural network as the backbone of the diffusion model. Through experiments based on simulated observations from the ERA5 reanalysis dataset, our method can produce assimilated global atmospheric data consistent with observations at 0.25° (≈30km) resolution globally. This marks the highest resolution achieved by ML data assimilation models. The experiments also show that the initial conditions assimilated from sparse observations (less than 0.96% of gridded data) and 48-hour forecast can be used for forecast models with a

Figure 1. Diagram of a numerical weather forecasting pipeline. It consists of data assimilation, simulation and post-processing. Data assimilation produces gridded values from sparse observations and predicted gridded values from previous time steps. Simulation takes in gridded values and produces predictions in gridded values at future time steps. Post-processing improves prediction so that it is closer to future observations.

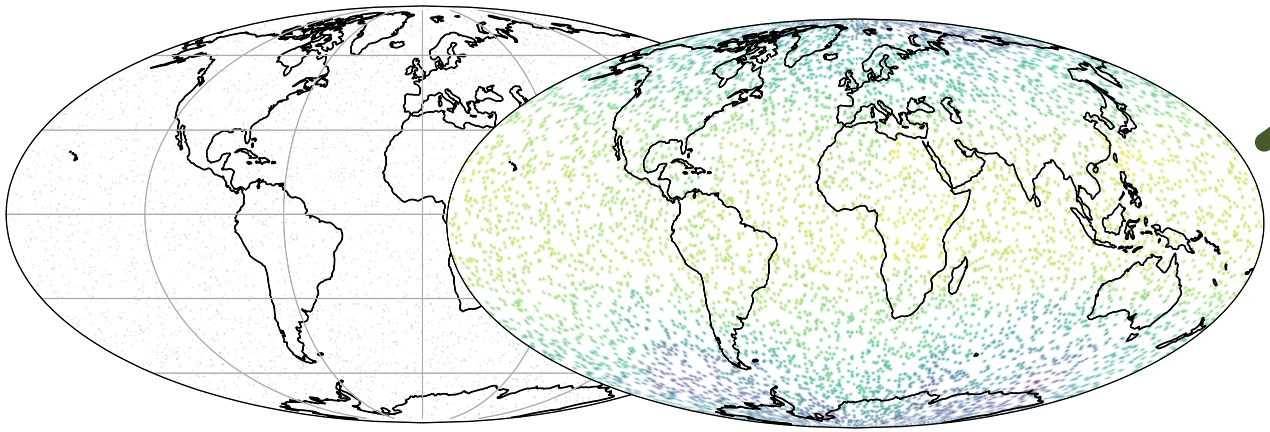
observations from various locations. The quality of these weather simulation models depends heavily on data assimilation, as errors in initial conditions are one of the main sources of forecast error (Bonavita et al., 2016). Additionally, data assimilation is employed in creating reanalysis datasets, which contain reconstructed historical weather variables as gridded fields. These reanalysis datasets play

+48h GraphCast Forecast

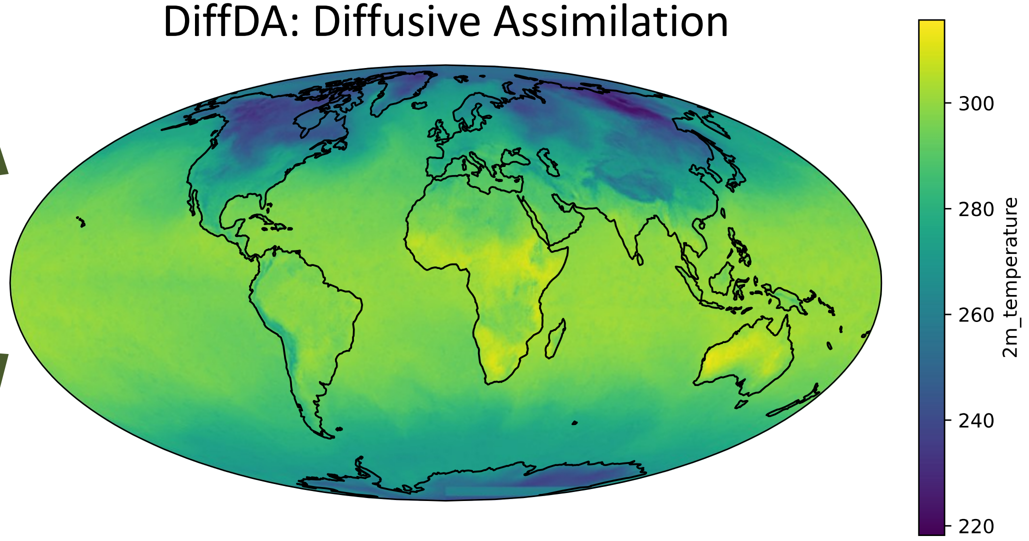


+48h Sparse Observations

+softmask



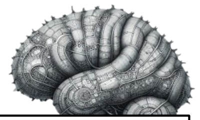
DiffDA: Diffusive Assimilation

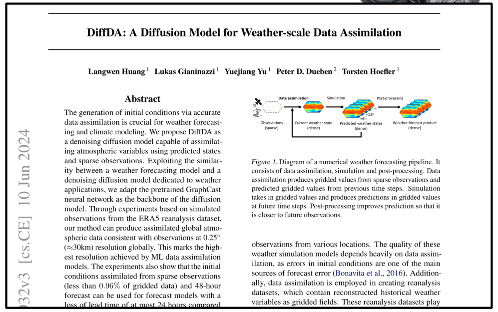


How good is this really?
Let’s look at Errors!

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D. Dueben, Torsten Hoefler: DiffDA: a Diffusion model for weather-scale Data Assimilation arXiv:2401.05932 + ICML’24, Jul. 2024

Case study: Assimilated 2m temperature at 2022-01-03 06z




DiffDA: A Diffusion Model for Weather-scale Data Assimilation

Langwen Huang[1] Lukas Gianinazzi[1] Yuejiang Yu[1] Peter D. Dueben[2] Torsten Hoefler[1]

Abstract

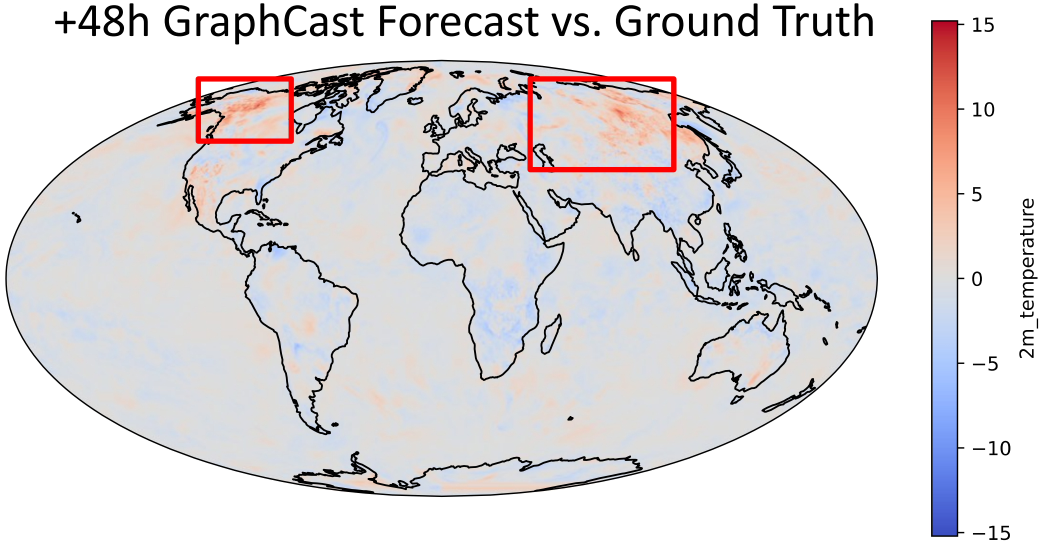
The generation of initial conditions via accurate data assimilation is crucial for weather forecasting and climate modeling. We propose DiffDA as a denoising diffusion model capable of assimilating atmospheric variables using predicted states and sparse observations. Exploiting the similarity between a weather forecasting model and a denoising diffusion model dedicated to weather applications, we adapt the pretrained GraphCast neural network as the backbone of the diffusion model. Through experiments based on simulated observations from the ERA5 reanalysis dataset, our method can produce assimilated global atmospheric data consistent with observations at 0.25° (≈30km) resolution globally. This marks the highest resolution achieved by ML data assimilation models. The experiments also show that the initial conditions assimilated from sparse observations (less than 0.96% of gridded data) and 48-hour forecast can be used for forecast models with a loss of lead time of at most 24 hours compared

Figure 1. Diagram of a numerical weather forecasting pipeline. It consists of data assimilation, simulation and post-processing. Data assimilation produces gridded values from sparse observations and predicted gridded values from previous time steps. Simulation takes in gridded values and produces predictions in gridded values at future time steps. Post-processing improves prediction so that it is closer to future observations.

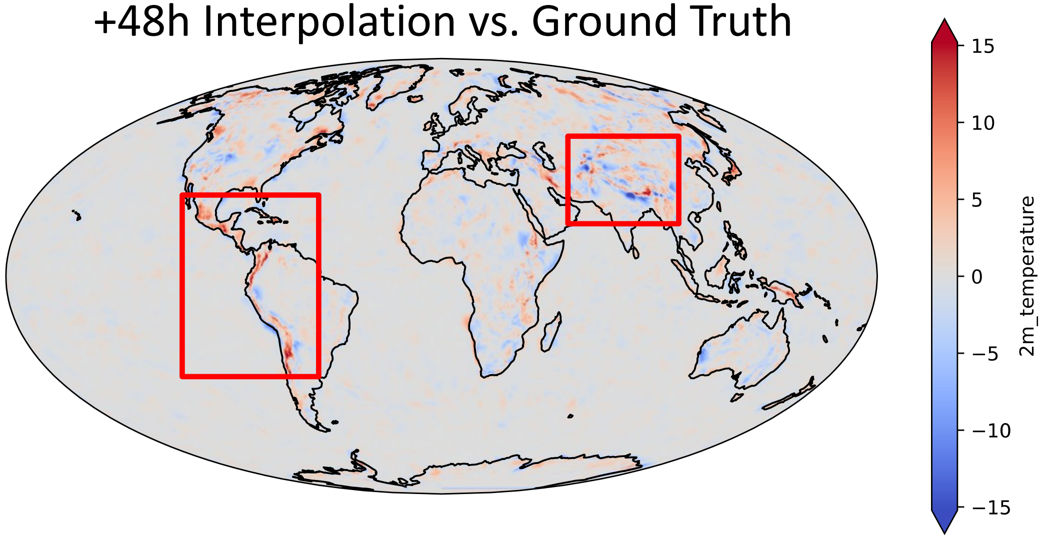
observations from various locations. The quality of these weather simulation models depends heavily on data assimilation, as errors in initial conditions are one of the main sources of forecast error (Bonavita et al., 2016). Additionally, data assimilation is employed in creating reanalysis datasets, which contain reconstructed historical weather variables as gridded fields. These reanalysis datasets play

...32v3 [cs.CE] 10 Jun 2024

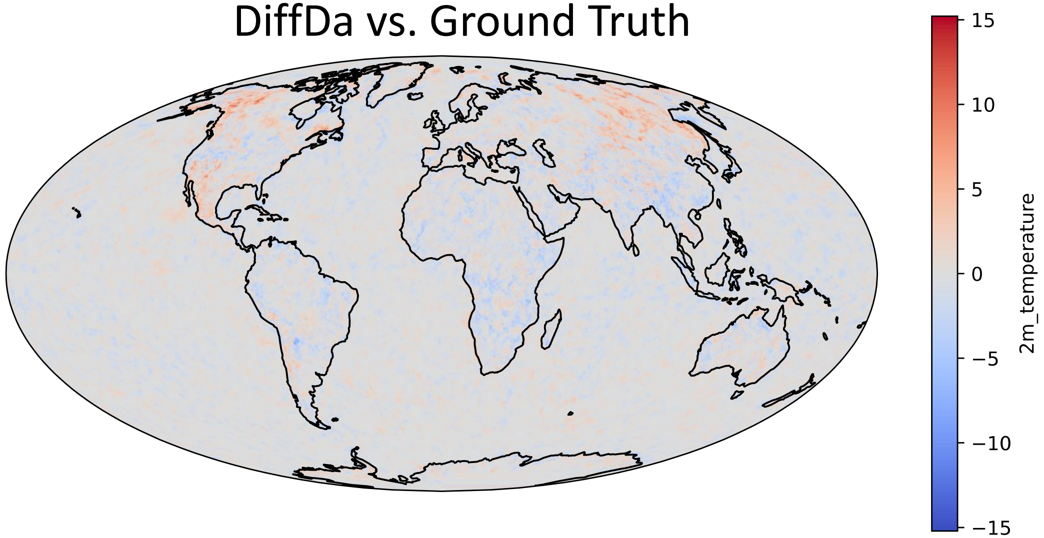
+48h GraphCast Forecast vs. Ground Truth



+48h Interpolation vs. Ground Truth

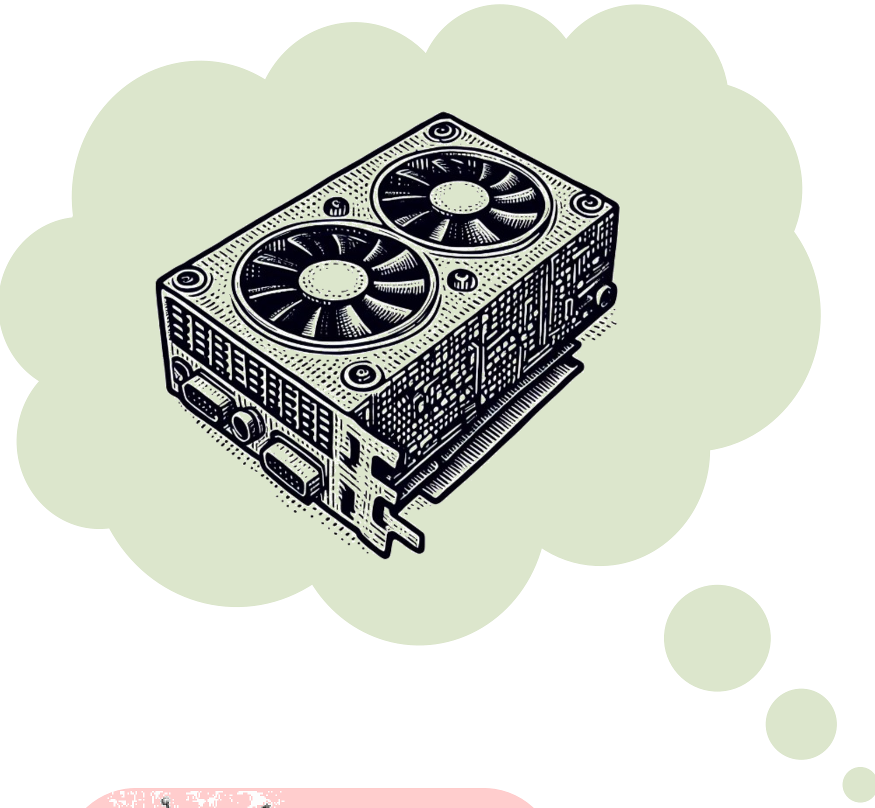


DiffDa vs. Ground Truth



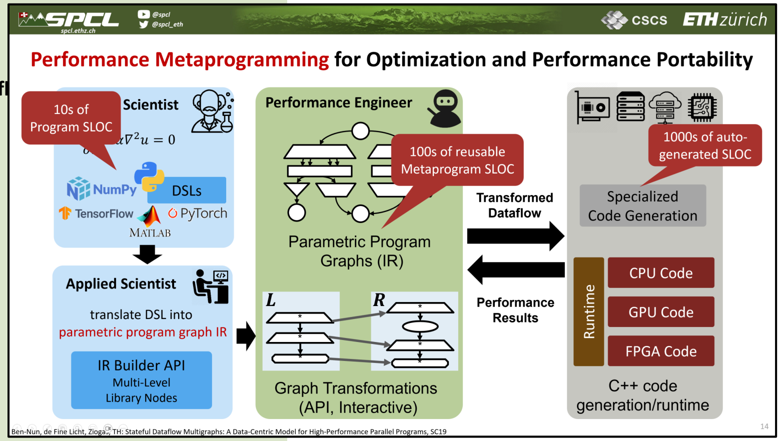
DiffDa-Assimilated data is better than both, forecast and interpolated observations

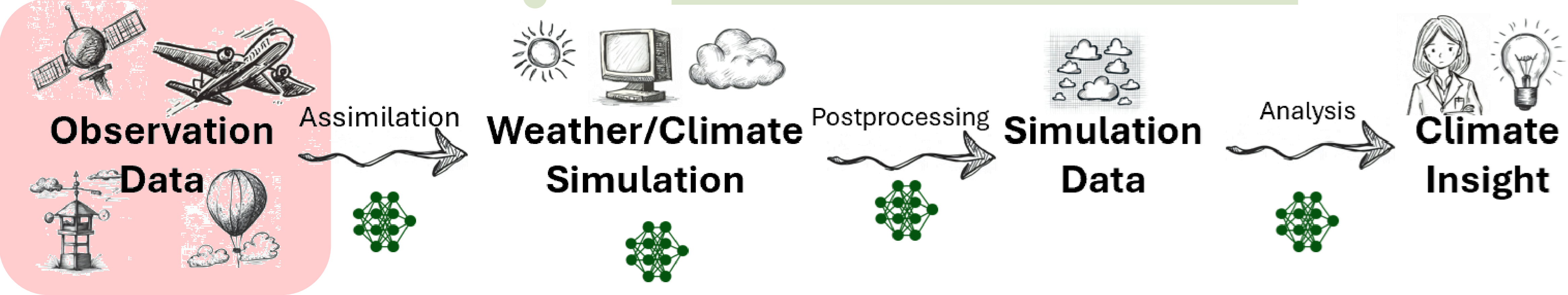
Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D. Dueben, Torsten Hoefler: DiffDA: a Diffusion model for weather-scale Data Assimilation arXiv:2401.05932 + ICML’24, Jul. 2024



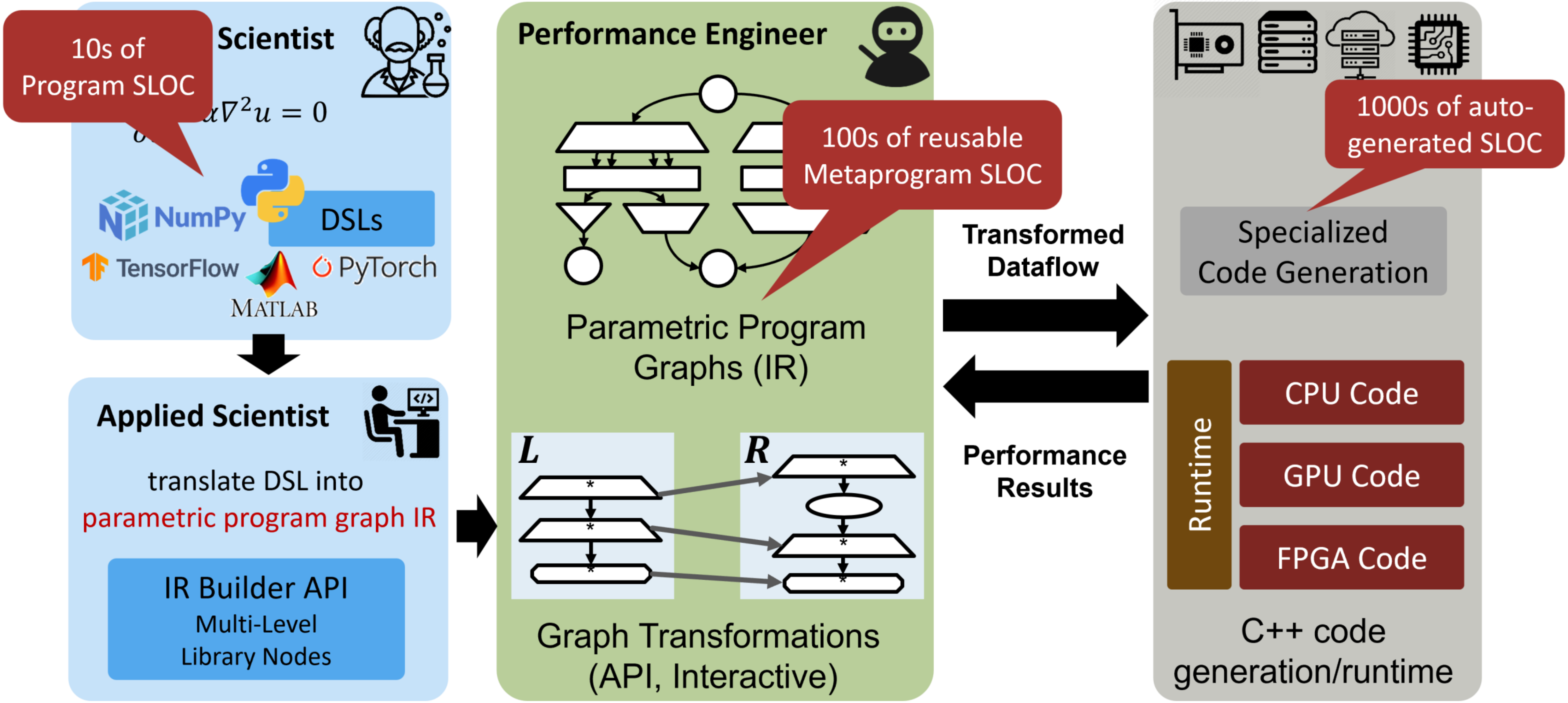
```
!$ACC DATA &
!$ACC PRESENT(density1,energy1) &
!$ACC PRESENT(vol_flux_x,vol_flux_y,volume,mass_flux_x,mass_flux_y,vertexdx,vertexdy) &
!$ACC PRESENT(pre_vol,post_vol,ener_flux)

!$ACC KERNELS

  IF(dir.EQ.g_xdir) THEN

   IF(sweep_number.EQ.1)THEN

!$ACC LOOP INDEPENDENT
     DO k=y_m    2,y_max+2
!$ACC LOOP IN      NDENT
      DO j=x_min-      max+2
       pre_vol(j,k)=vo      e(j,k)+(vol_flux      1,k  )-vol_flux_x(j,k)+vol_fl
       post_vol(j,k)=pre     i,k)-(vol_f      x(j+1,k  )-vol_flux_x(j,k))
      ENDDO
     ENDDO
   ELSE
!$ACC LOOP INDEPENDEN
     DO k=y_min-2,y_m
!$ACC LOOP INDEPE      NT
      DO j=x_min      max+2
       pre_vol      volume(j,k)+vol_flux_x(j+1,      l_flux_x(j,k)
       post_vo  j,k)=volume(j,k)
      ENDDO
     ENDDO

   ENDIF
```

AI?





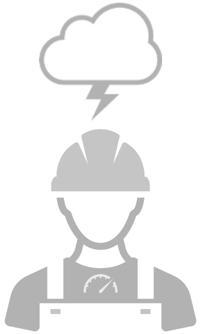
Performance Metaprogramming for Optimization and Performance Portability



Ben-Nun, de Fine Licht, Ziogas, TH: Stateful Dataflow Multigraphs: A Data-Centric Model for High-Performance Parallel Programs, SC19

Pace in DaCe for Performance Metaprogramming – 12k SLOC Python

AI-based Transfer Tuning to the Rescue!

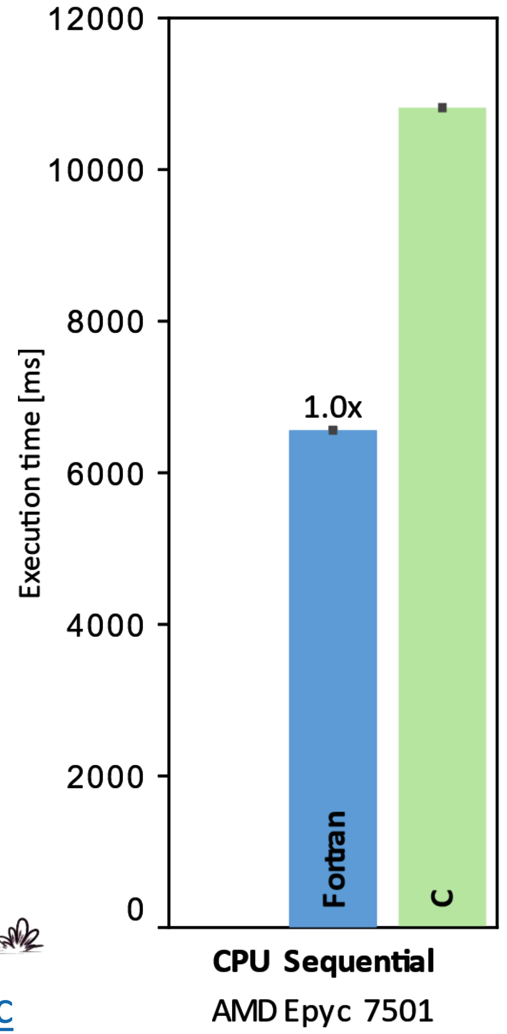


Another real production code … ECMWF's CLOUDSC

```fortran
9
10  SUBROUTINE CLOUDSC &
11   !---input
12   & (KIDIA,    KFDIA,    KLON,    KLEV, &
13   & PTSPHY,&
14   & PT, PQ, tendency_cml,tendency_tmp,tendency_loc, &
15   & PVFA, PVFL, PVFI, PDYNA, PDYNL, PDYNI, &
16   & PHRSW,    PHRLW,&
17   & PVERVEL,  PAP,      PAPH,&
18   & PLSM,     LDCUM,    KTYPE, &
19   & PLU,      PLUDE,    PSNDE,    PMFU,     PMFD,&
20   !---prognostic fields
21   & PA,&
22   & PCLV,  &
23   & PSUPSAT,&
24  !-- arrays for aerosol-cloud interactions
25  !!! & PQAER,    KAER, &
26   & PLCRIT_AER,PICRIT_AER,&
27   & PRE_ICE,&
28   & PCCN,     PNICE,&
29   !---diagnostic output
30   & PCOVPTOT, PRAINFRAC_TOPRFZ,&
31   !---resulting fluxes
32   & PFSQLF,   PFSQIF ,  PFCQNNG,  PFCQLNG,&
33   & PFSQRF,   PFSQSF ,  PFCQRNG,  PFCQSNG,&
34   & PFSQLTUR, PFSQITUR , &
35   & PFPLSL,   PFPLSN,   PFHPSL,   PFHPSN, KFLDX, &
36   & YDCST, YDTHF, YDECLDP)
```

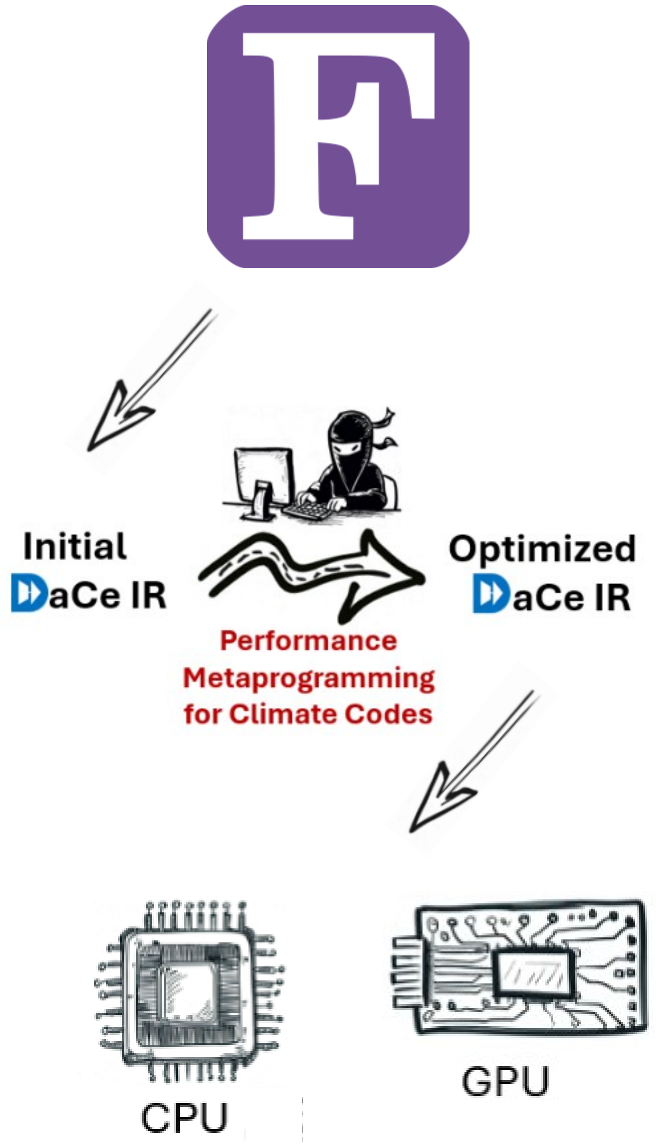
… variable setup/initialization until line 500 ;-)

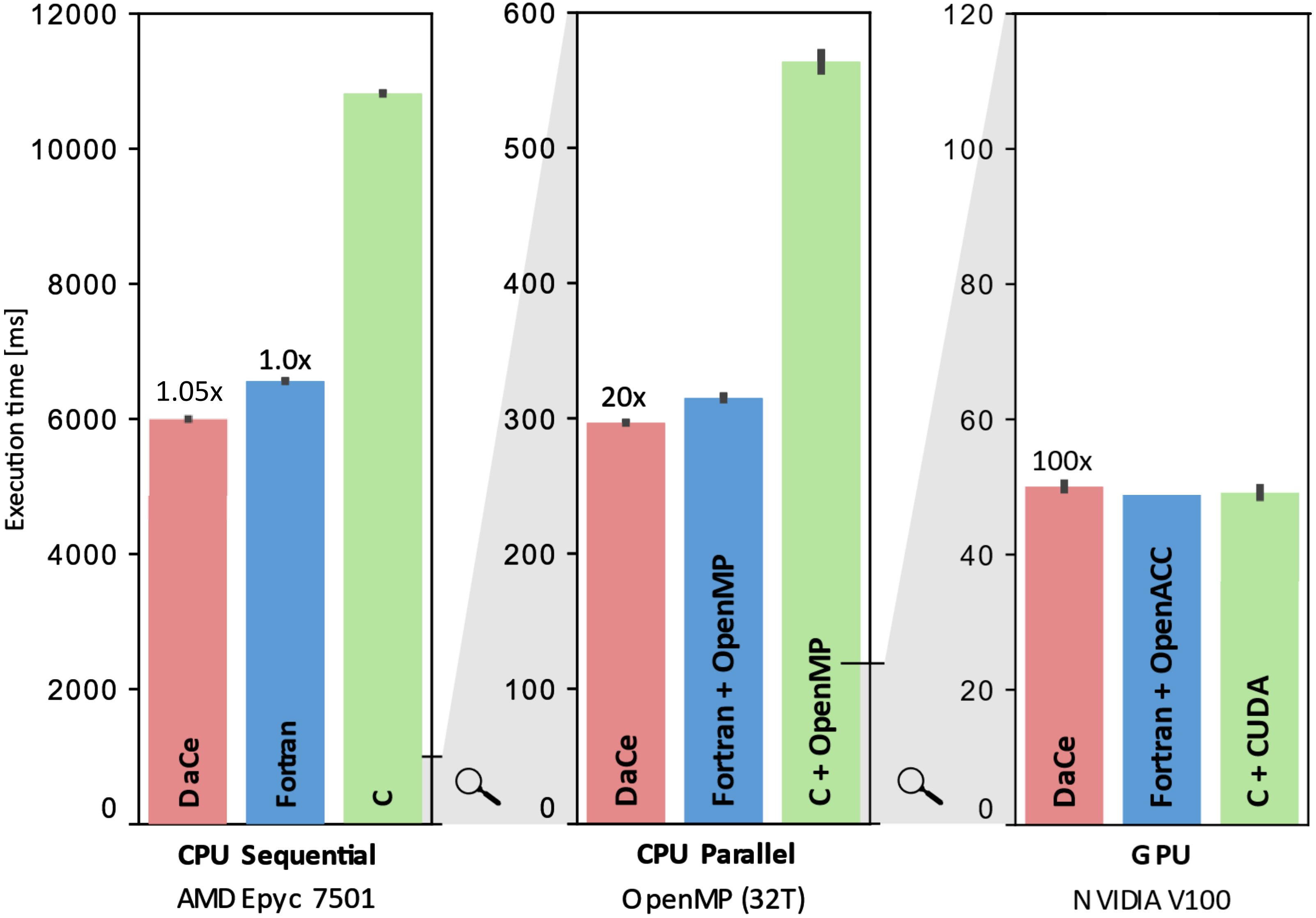
- **Cloud Microphysics of IFS**
  - Resolve sub-grid features
  - Original 2,525 SLOC of **Fortran 95**
- **Rewritten for performance portability benchmarking (optimization took months!)**
  - 2,635 SLOC C
  - 2,610 SLOC C++/CUDA



https://github.com/ecmwf-ifs/dwarf-p-cloudsc

Performance Metaprogramming – from the unchanged CLOUDSC Fortran code!





The three pillars of AI in Climate Sciences

Data

- Unstructured observation data
- Structured simulation data

Combine both to train models

- Learn physics and data-driven prediction

Compute

- AI models require accelerated high-performance computing for training

Accelerate AI computations

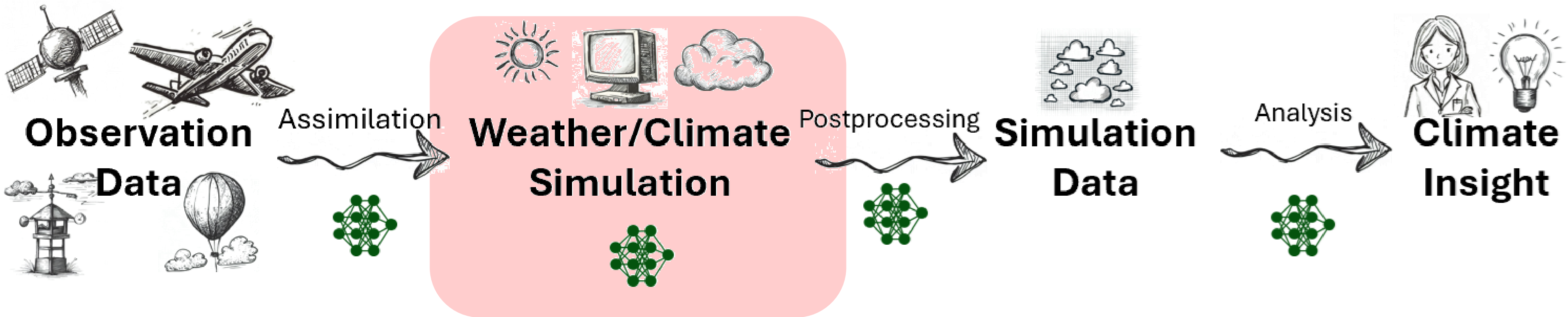
- Re-use infrastructure from LLMs and related generative AI methods (GNN, CNN)

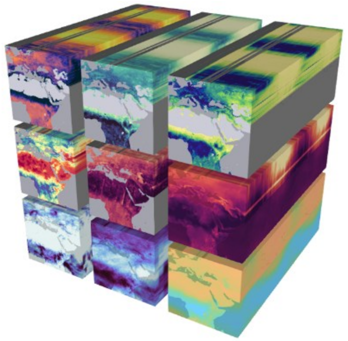
Models

- Models need to provide the right structural bias/prior

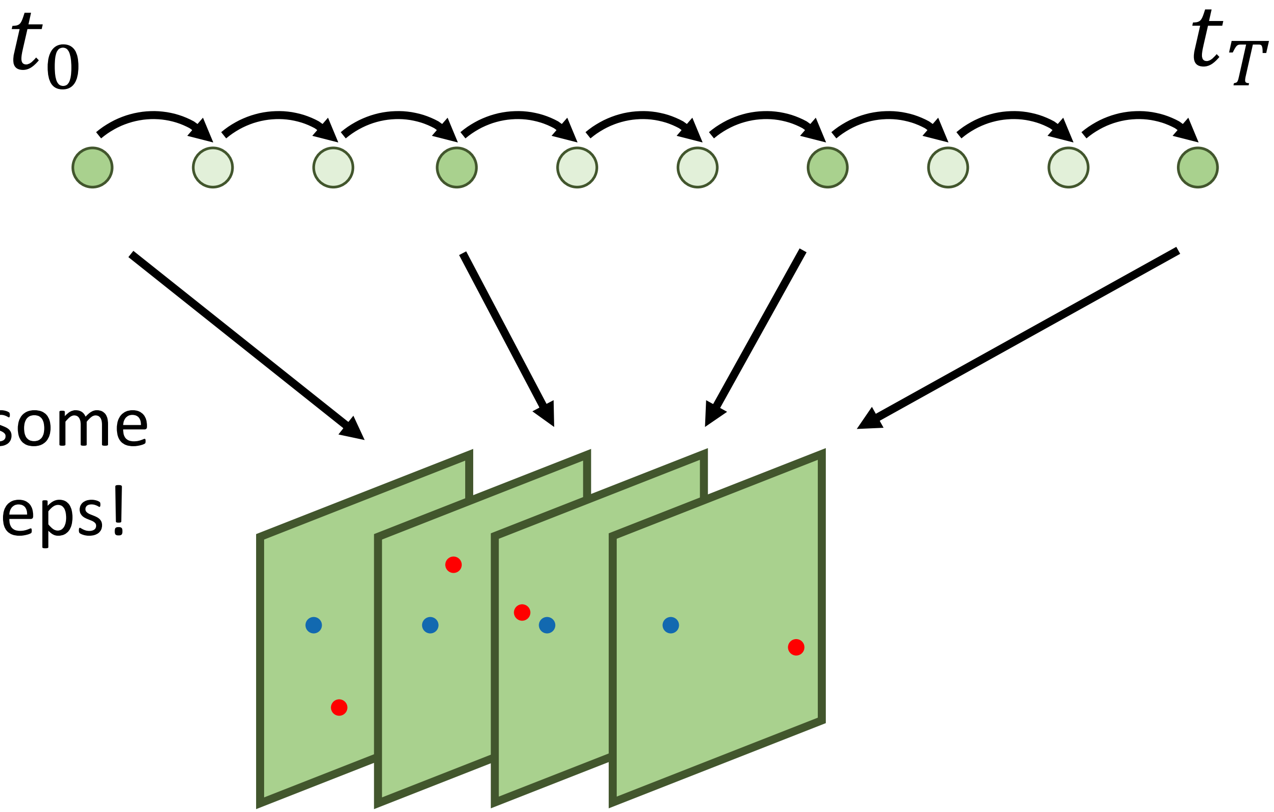
Develop better AI methods

- Step 1: use generative AI models: tformer, CNN, GNN, Diffusion, etc.
- Step 2: use automatically parameterized physics-based models encoding equations



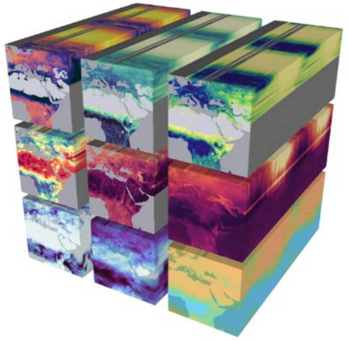


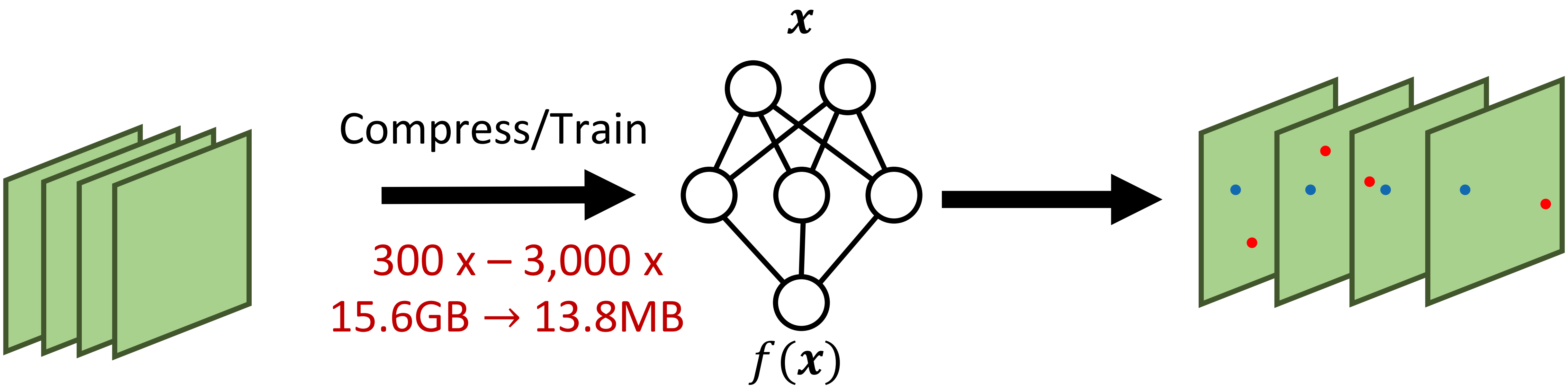
Simulation runs time-stepping forward



Store some timesteps!

Analysis access pattern is often strided or even random





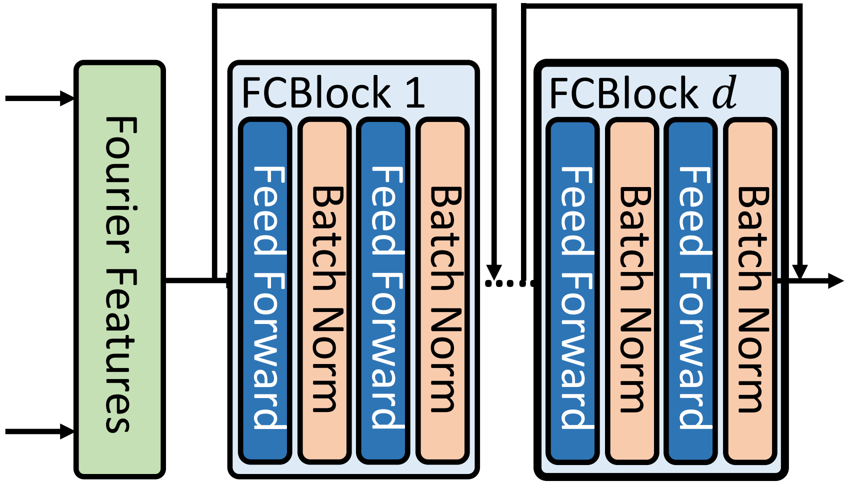
Multidimensional Data

Neural Representation

Analysis Queries

Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

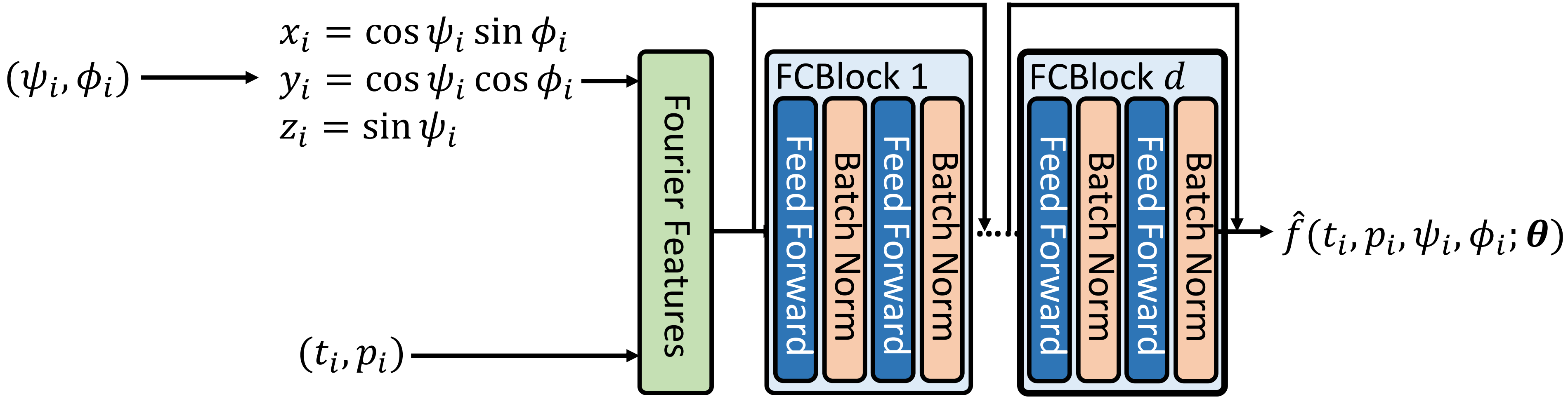
Neural Network Structure



Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

Neural Network Structure

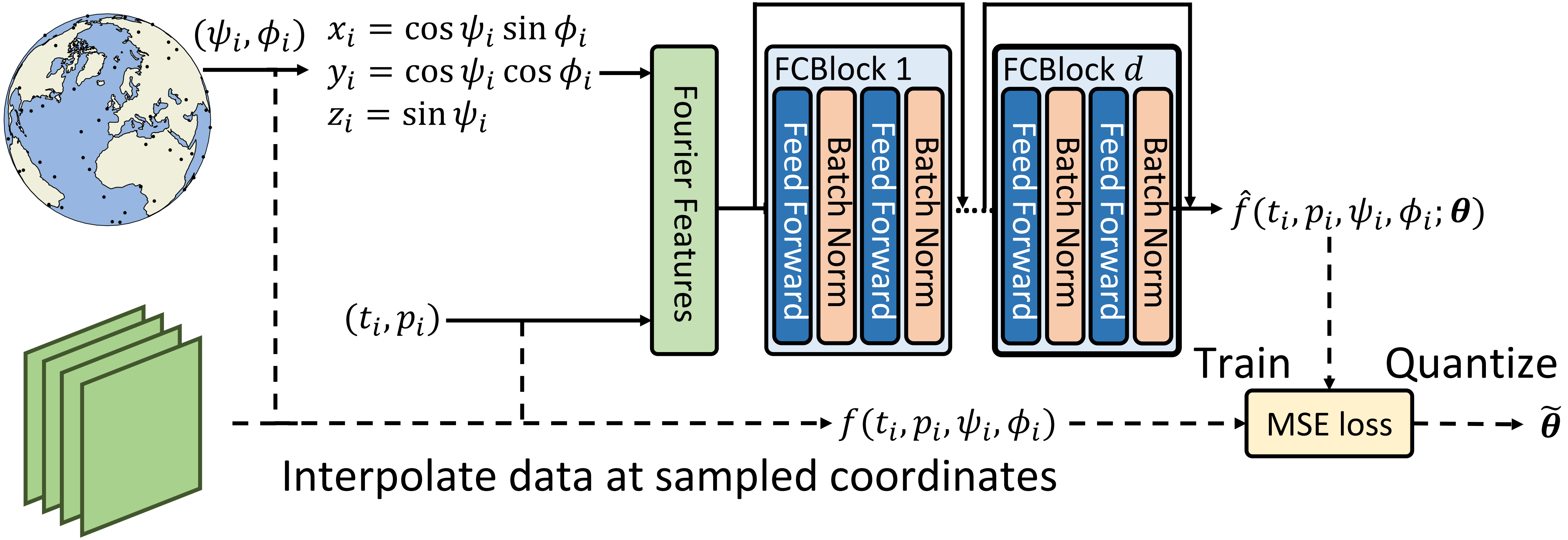
Decompression / Inference



- On-demand decompression
- Fully utilize GPUs

Neural Network Structure

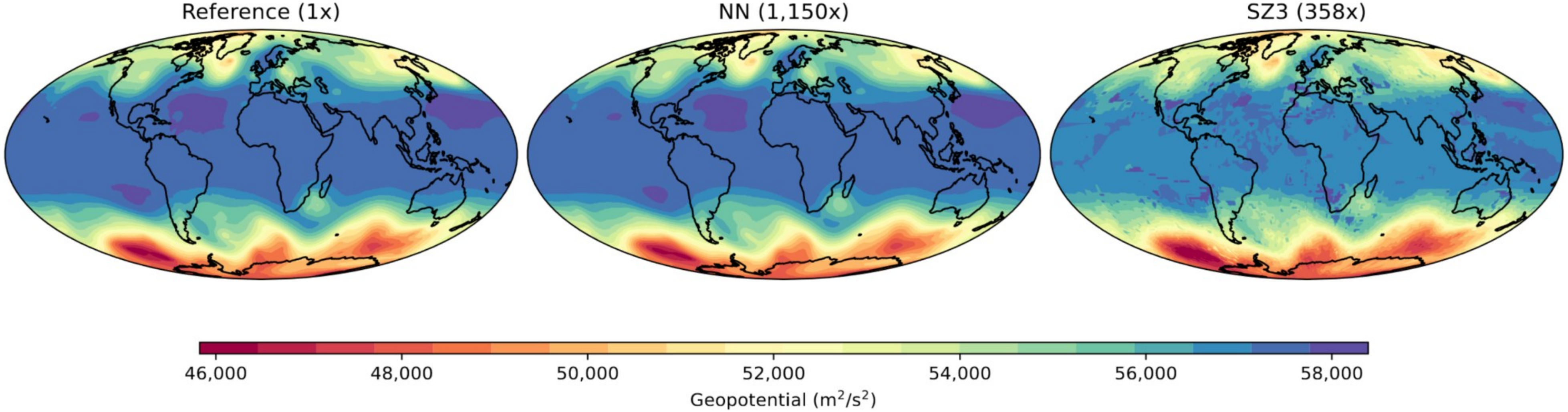
Compression / Training



Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

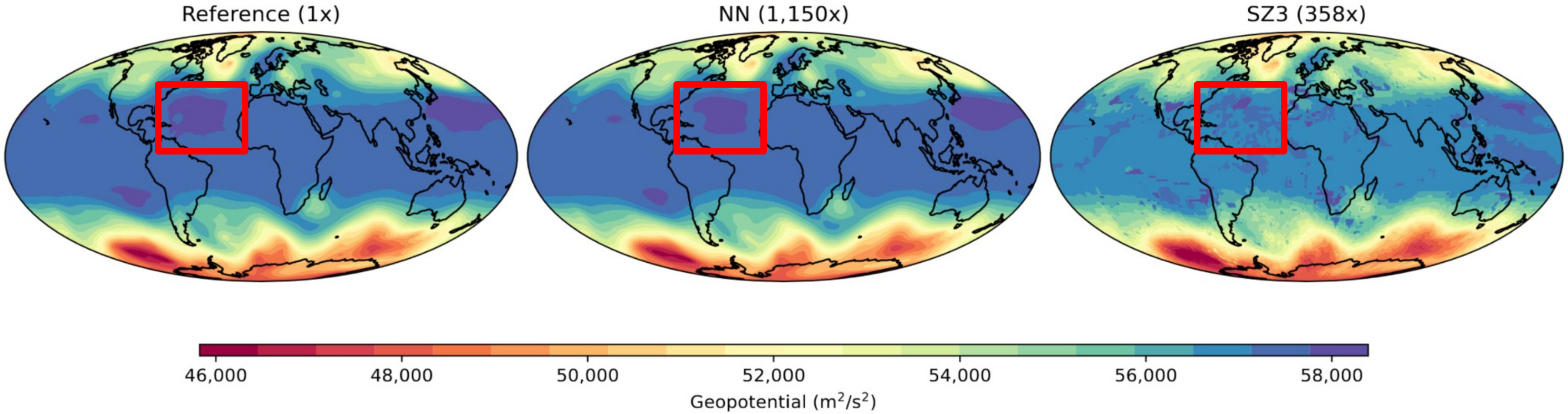
Evaluation: Case Study

Geopotential at 500hPa, 2016 Oct 5th



Evaluation: Case Study

Geopotential at 500hPa, 2016 Oct 5th

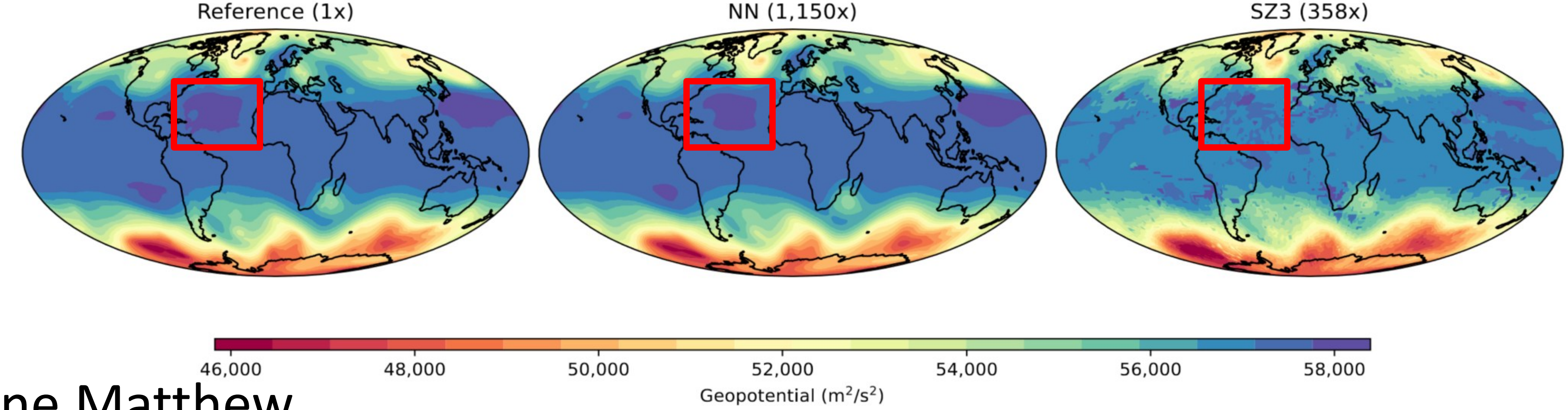


Preserves general shapes of important events and average values without introducing significant artifacts

Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

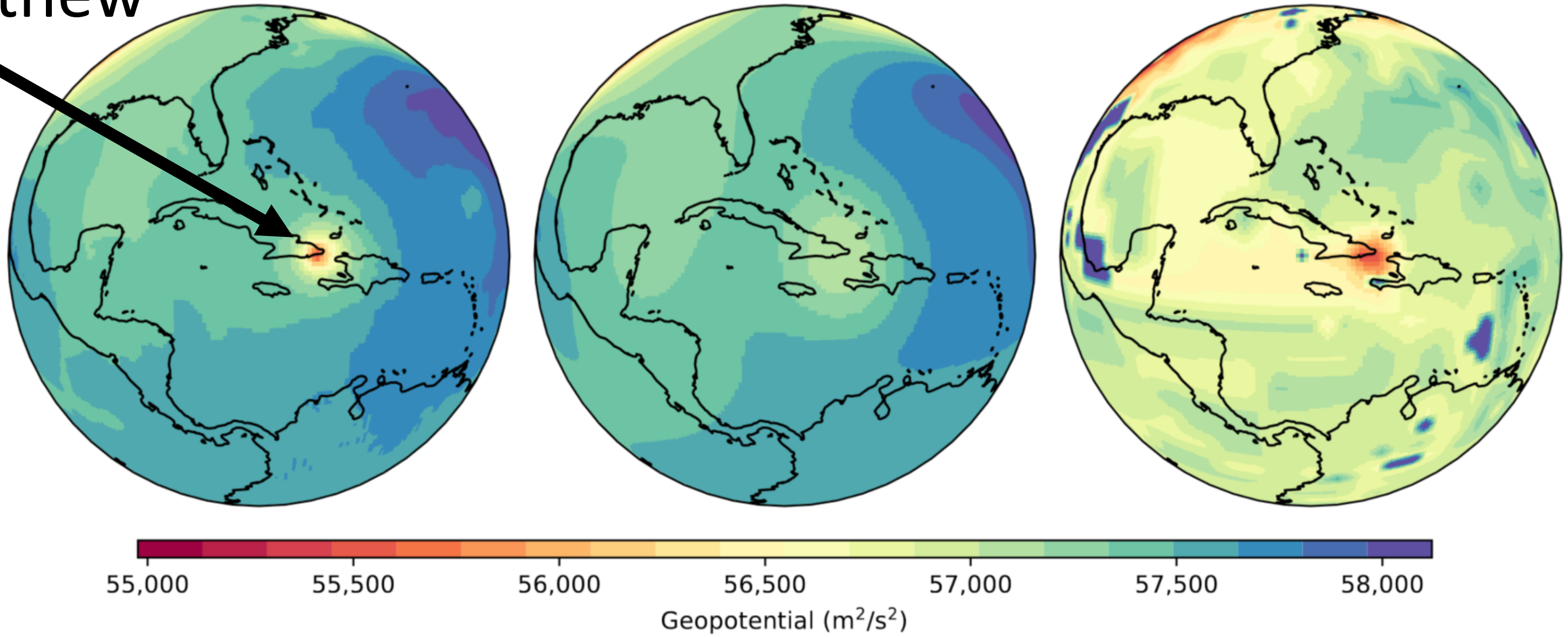
Evaluation: Case Study

Geopotential at 500hPa, 2016 Oct 5th



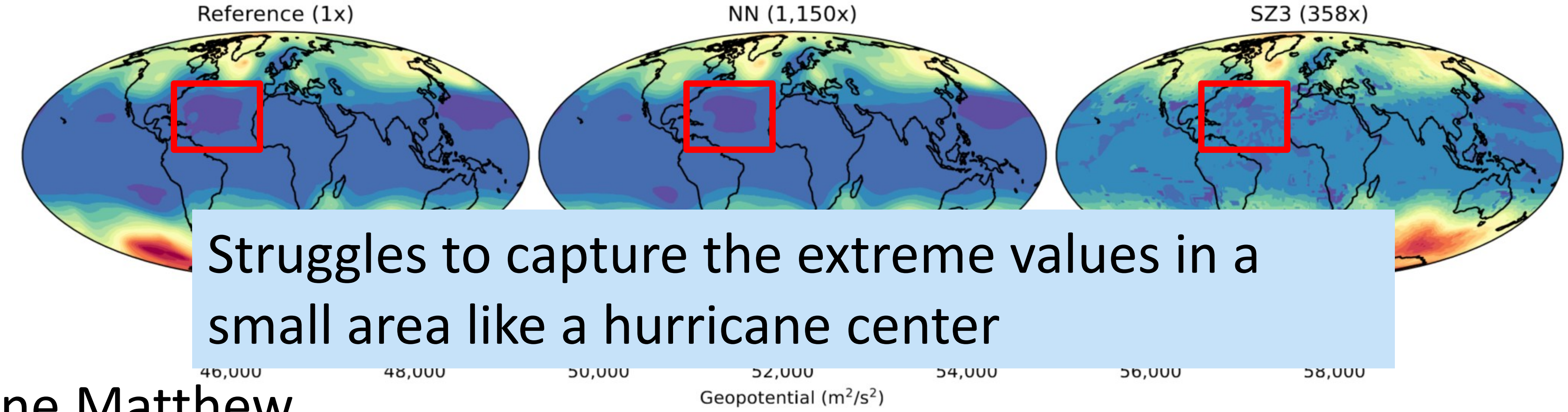
Hurricane Matthew

16.5bn damage
603 fatalities



Evaluation: Case Study

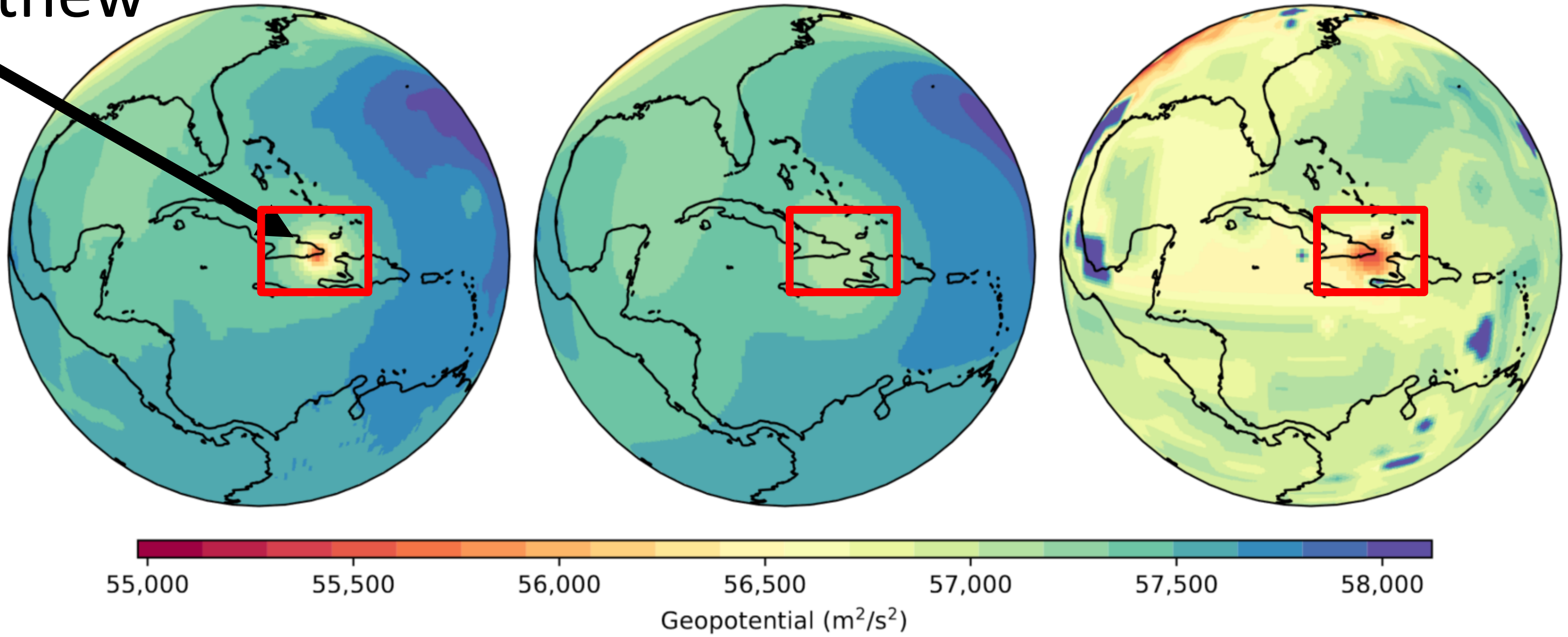
Geopotential at 500hPa, 2016 Oct 5th



Struggles to capture the extreme values in a small area like a hurricane center

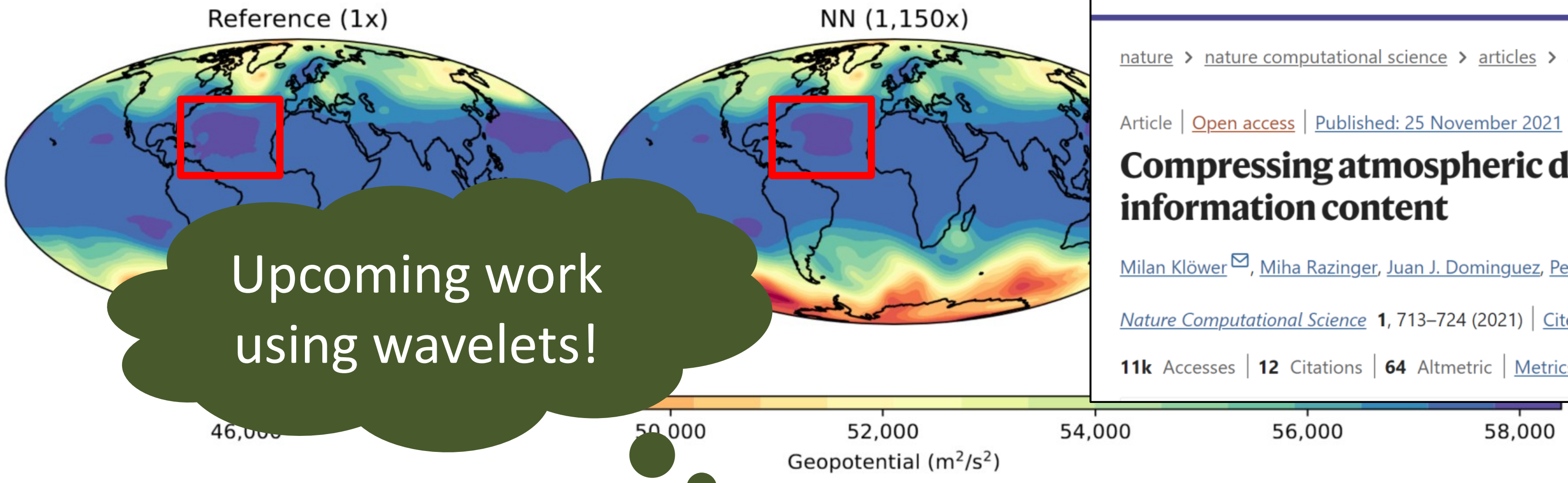
Hurricane Matthew

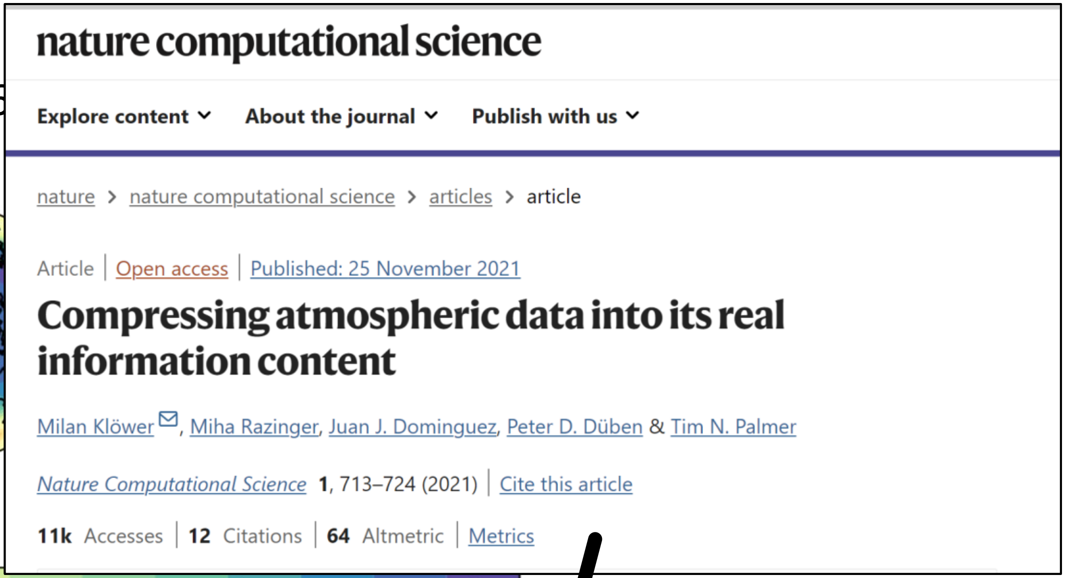
16.5bn damage
603 fatalities



Evaluation: Case Study

Geopotential at 500hPa, 2016





nature computational science

Explore content About the journal Publish with us

nature > nature computational science > articles > article

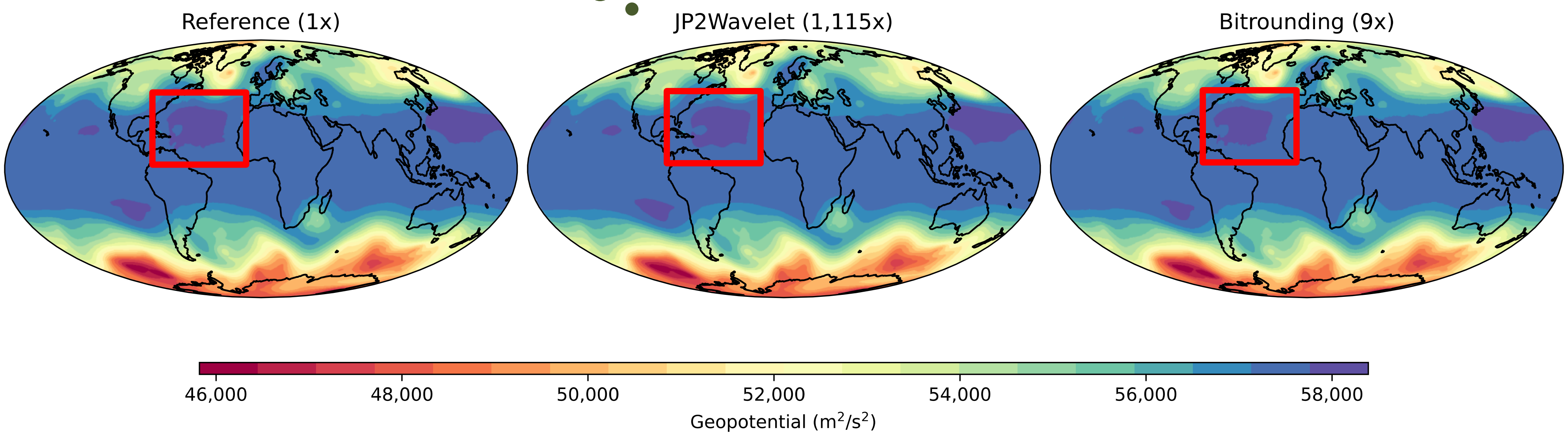
Article | Open access | Published: 25 November 2021

Compressing atmospheric data into its real information content

Milan Klöwer, Miha Razinger, Juan J. Dominguez, Peter D. Düben & Tim N. Palmer

Nature Computational Science **1**, 713–724 (2021) | Cite this article

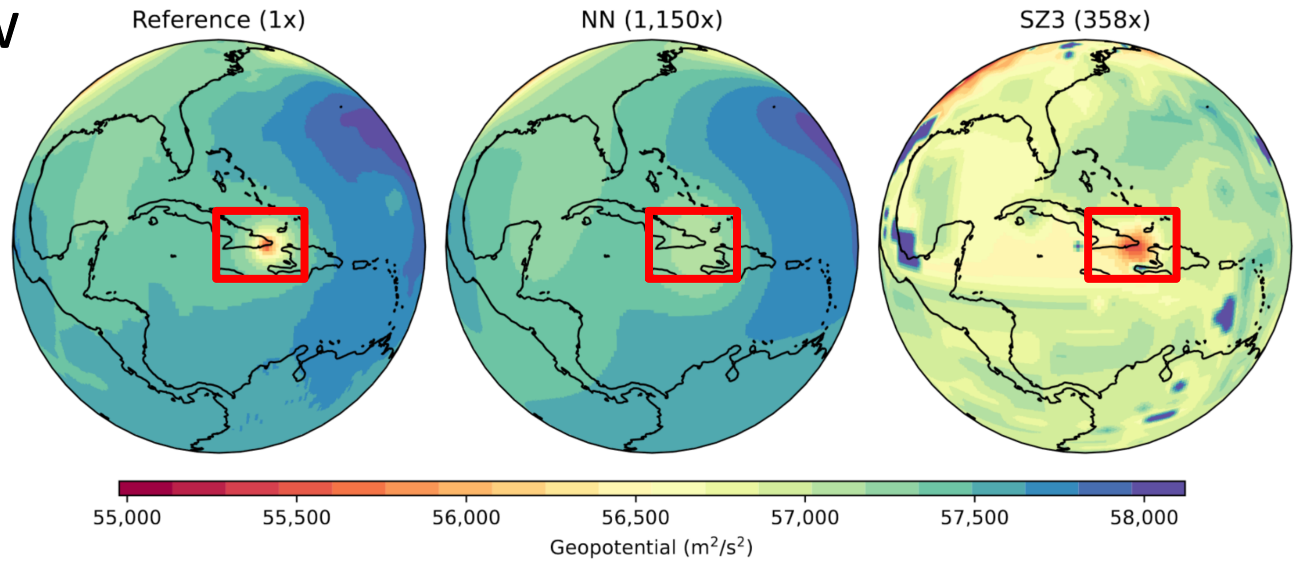
11k Accesses | **12** Citations | **64** Altmetric | Metrics

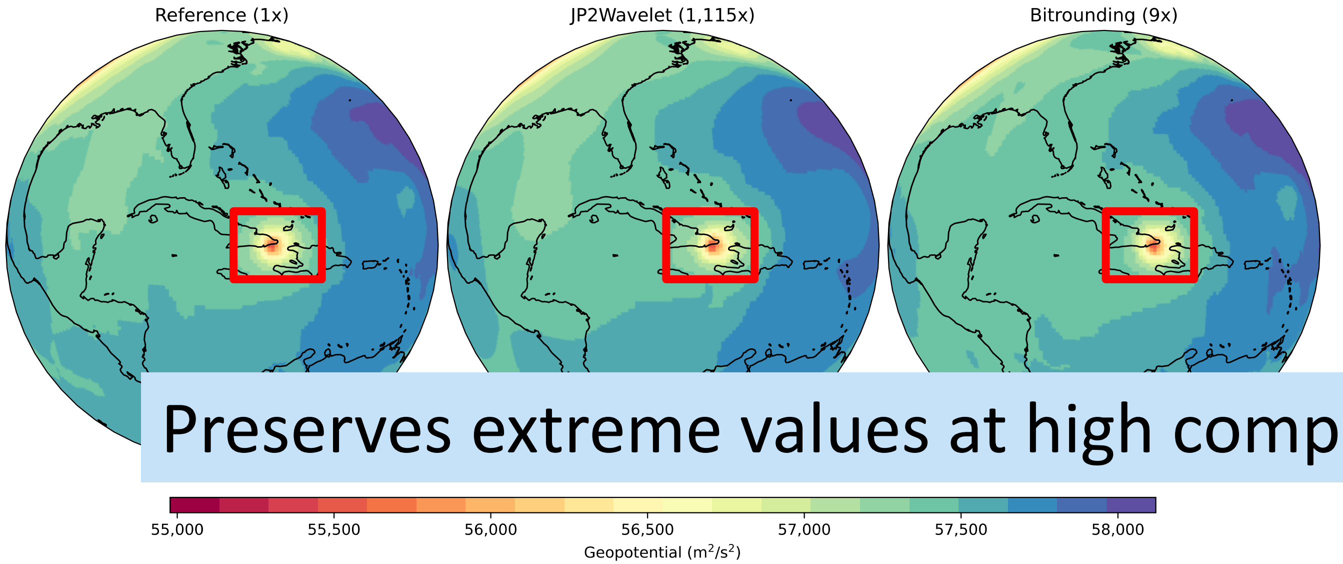


Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

Evaluation: Case Study

Hurricane Matthew





Preserves extreme values at high compression ratio

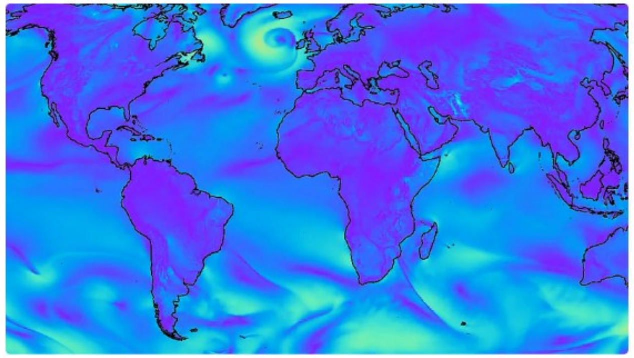
Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR’23

Is AI based weather and climate prediction a solved problem?

Google DeepMind's AI Weather Forecaster Handily Beats a Global Standard

Machine learning algorithms that digested decades of weather data were able to fo... atmospheric measures more accurately than Europe's top weather center.

Wired, Nov. 23 **GraphCastNet**



Google DeepMind's GraphCast AI software produces weather forecasts for weather variables like wind speed much faster than traditional simulations. COURTESY OF GOOGLE

Pangu Weather Model

AI Weather Prediction More Accurate, in Just Seconds

Pangu Weather

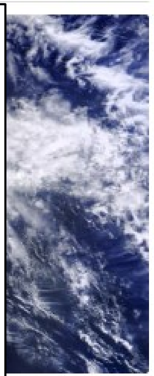
Huawei Cloud's Pangu-Weathe... is the ...t AI mode... at can predict weather more accurate... (NWP) methods, and at a sp... tha... 10,00... mes fa...r. In the past, predicting the path of a typ... on a high-performance clus... of 3,00... rve... Now, P...u-Weather can do that in 10 seconds on... global weather forecast within seconds, with all factors taken into consideration, such as geopotent... pressure. Besides typhoon paths, it can also accurately predict precipitation, cold and heat waves, a...

AI?

NVIDIA's New AI Model Revolutionizes Extreme Weather Forecasting with Unprecedented Accuracy

August 20, 2024 by Ali Azhar

FourCastNet

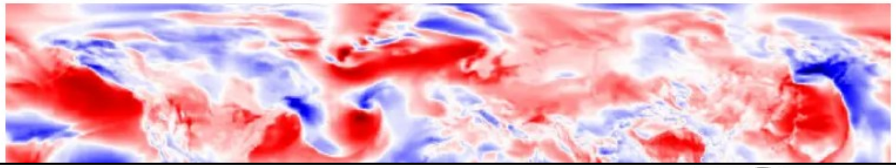


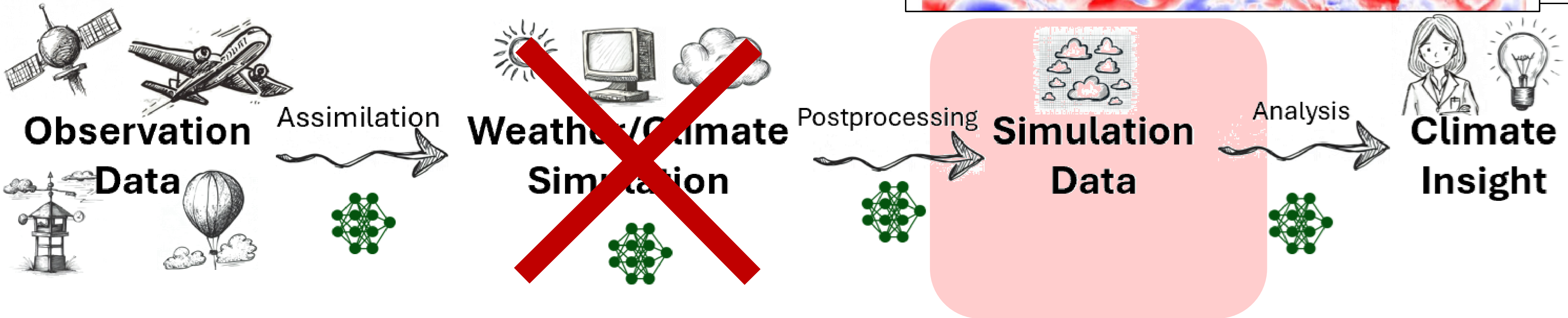
...d-breaking heat ...unities and ...age annually in the

Introducing Prithvi WxC, a new general-purpose AI model for weather and climate

Sept. 2024

In collaboration with NASA, IBM is releasing an open-source foundation model that can be customized for a variety of weather and climate-related applications — and run on a desktop computer.





Out of Distribution (Future) Case Study: Storm Ciarán (Nov. 2023)

npj | climate and atmospheric science Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00638-w

Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán

Check for updates

Andrew J. Charlton-Perez[1], Helen F. Dacre[1], Simon Driscoll[1,2], Suzanne L. Gray[1], Ben Harvey[1,3], Natalie J. Harvey[1], Kieran M. R. Hunt[1,3], Robert W. Lee[1], Ranjini Swaminathan[1,2], Remy Vandaele[1,2] & Ambrogio Volonté[1,3]

There has been huge recent interest in the potential of making operational weather forecasts using machine learning techniques. As they become a part of the weather forecasting toolbox, there is a

GraphCastNet Pangu Weather

FourCastNet

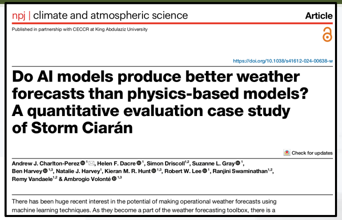
vs.

ECMWF OpenIFS

16 dead, gusts over 185 km/h
likely strongest since 1954
1.2m households w/o power
1m residents cut off phone
airports closed (e.g., AMS)
…

Photograph: Getty Images

Out of Distribution (Future) Case Study: Storm Ciarán (Nov. 2023)

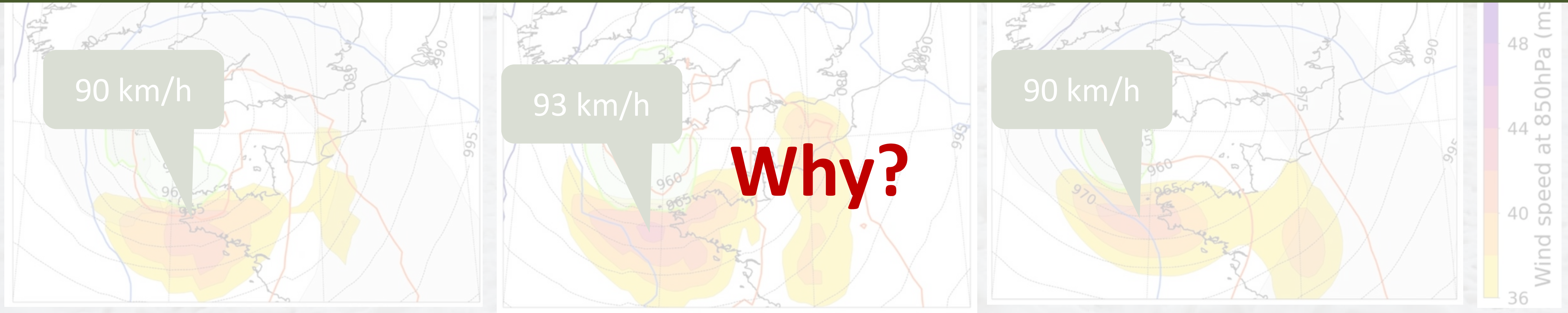

npj | climate and atmospheric science — Article

Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán

There has been huge recent interest in the potential of making operational weather forecasts using machine learning techniques. As they become a part of the weather forecasting toolbox, there is a

“more than 48 h before Storm Ciarán […], forecasts of the rapid MSLP deepening and track of the storms produced by the ML models were essentially indistinguishable”

“when considering the damaging winds associated with Storm Ciarán in detail, forecasts from the ML models had significant errors and poorer performance than conventional NWP models”



Photograph: Getty Images

Why? A deeper look inside some physics in ML-based Weather Prediction

Physics Case Study: Typhoon Doksuri, July 2023 (SE Asia, 137 dead, $28.4bn damage – strongest in region in 2023)

AGU ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL107377

On Some Limitations of Current Machine Learning Weather Prediction Models

Massimo Bonavita[1]

[1]ECMWF, Reading, UK

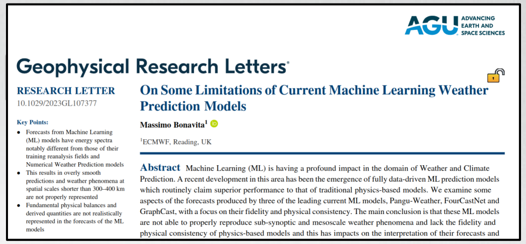
Key Points:

- Forecasts from Machine Learning (ML) models have energy spectra notably different from those of their training reanalysis fields and Numerical Weather Prediction models
- This results in overly smooth predictions and weather phenomena at spatial scales shorter than 300–400 km are not properly represented
- Fundamental physical balances and derived quantities are not realistically represented in the forecasts of the ML models

Abstract Machine Learning (ML) is having a profound impact in the domain of Weather and Climate Prediction. A recent development in this area has been the emergence of fully data-driven ML prediction models which routinely claim superior performance to that of traditional physics-based models. We examine some aspects of the forecasts produced by three of the leading current ML models, Pangu-Weather, FourCastNet and GraphCast, with a focus on their fidelity and physical consistency. The main conclusion is that these ML models are not able to properly reproduce sub-synoptic and mesoscale weather phenomena and lack the fidelity and physical consistency of physics-based models and this has impacts on the interpretation of their forecasts and

Photograph: wikipedia

Why? A deeper look inside some physics in ML-based Weather


AGU ADVANCING EARTH AND SPACE SCIENCES

Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL107377

On Some Limitations of Current Machine Learning Weather Prediction Models

Massimo Bonavita[1]

[1]ECMWF, Reading, UK

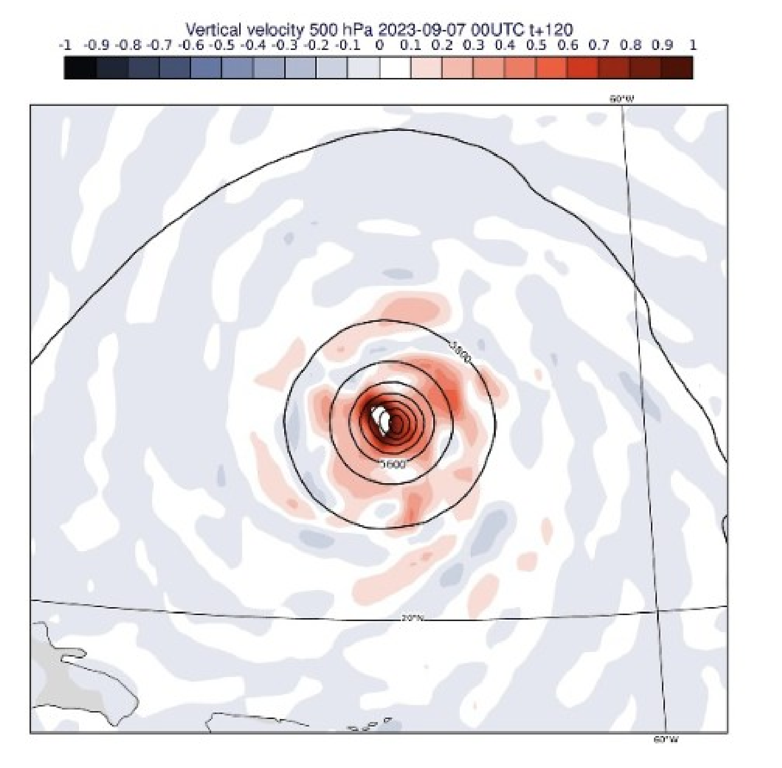
Key Points:

- Forecasts from Machine Learning (ML) models have energy spectra notably different from those of their training reanalysis fields and Numerical Weather Prediction models
- This results in overly smooth predictions and weather phenomena at spatial scales shorter than 300–400 km are not properly represented
- Fundamental physical balances and derived quantities are not realistically represented in the forecasts of the ML models

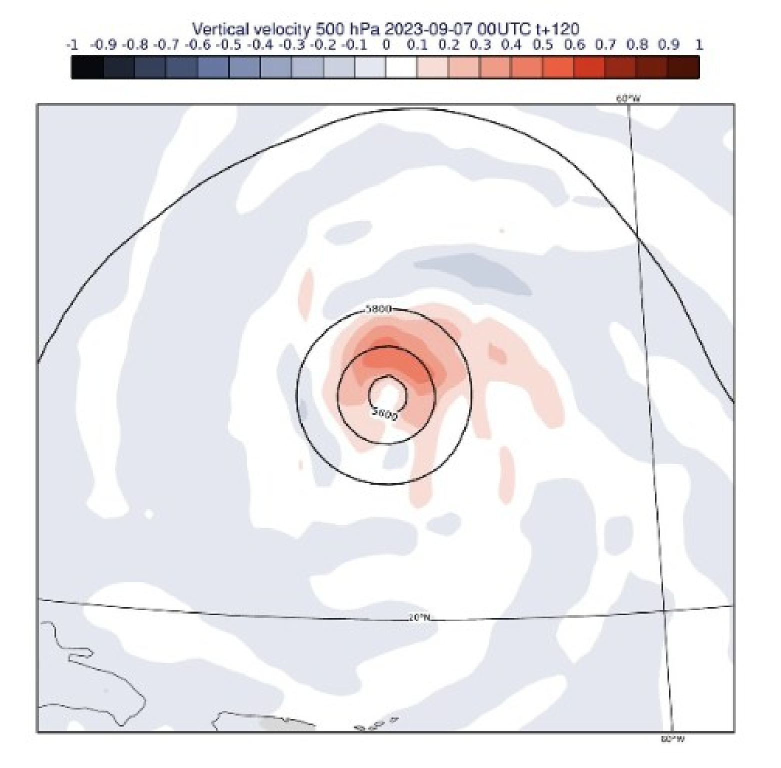
Abstract Machine Learning (ML) is having a profound impact in the domain of Weather and Climate Prediction. A recent development in this area has been the emergence of fully data-driven ML prediction models which routinely claim superior performance to that of traditional physics-based models. We examine some aspects of the forecasts produced by three of the leading current ML models, Pangu-Weather, FourCastNet and GraphCast, with a focus on their fidelity and physical consistency. The main conclusion is that these ML models are not able to properly reproduce sub-synoptic and mesoscale weather phenomena and lack the fidelity and physical consistency of physics-based models and this has impacts on the interpretation of their forecasts and

Physics Case Study: Typhoon Doksuri, July 2023 (SE Asia, 137 dead, $28.4bn damage – strongest in

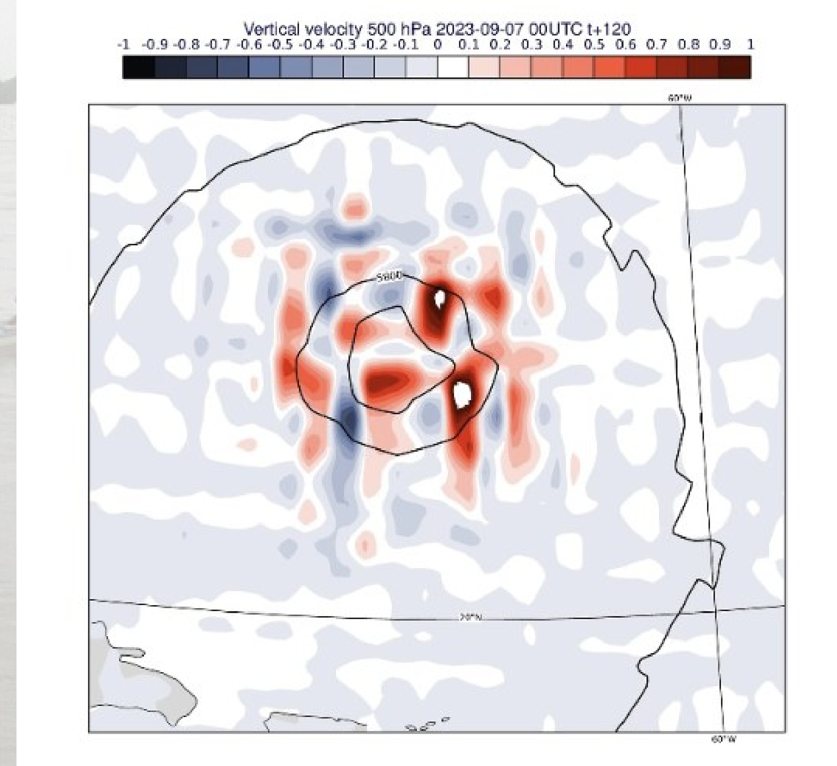
Vertical wind from continuity (mass conservation) $$\omega(p) = \omega(p_s) - \int_{p_s}^{p} \mathrm{div}(u)_p dp$$



ERA-5 data
(ground truth)



ECMWF's IFS
(simulation forecast)

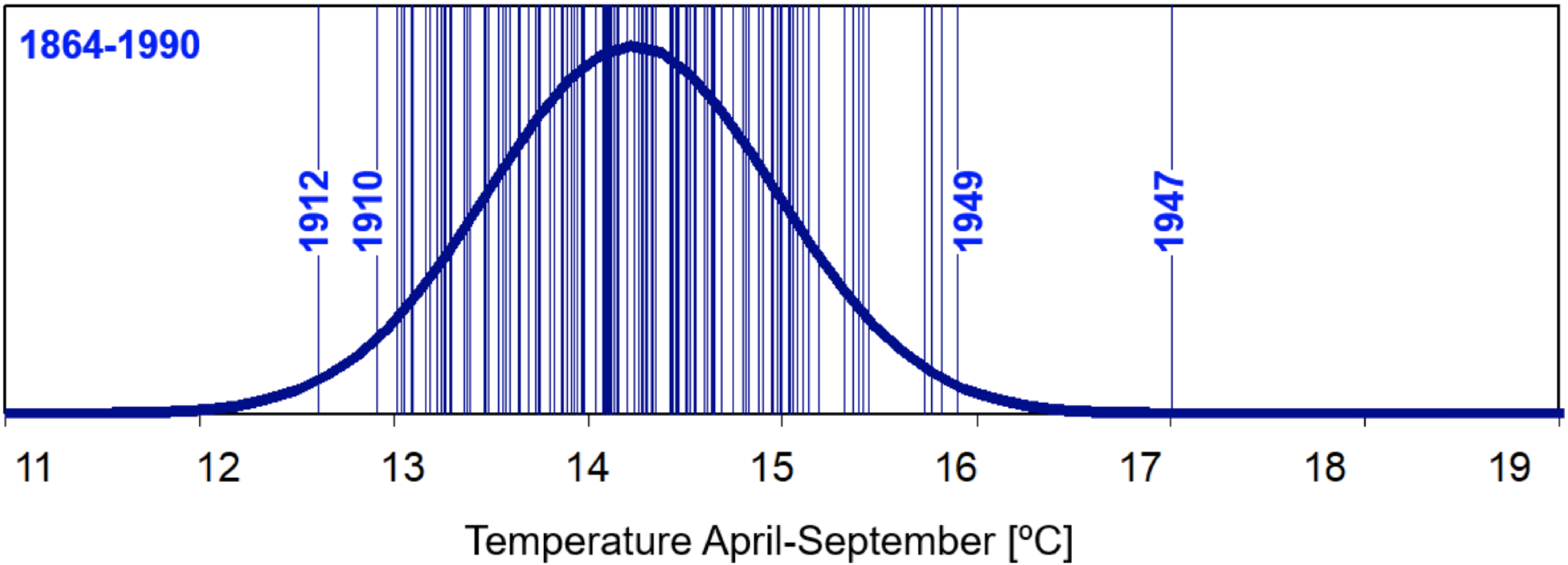


Pangu Weather
(AI based forecast)

Photograph: wikipedia

What about global warming? Can data-driven methods predict the future?

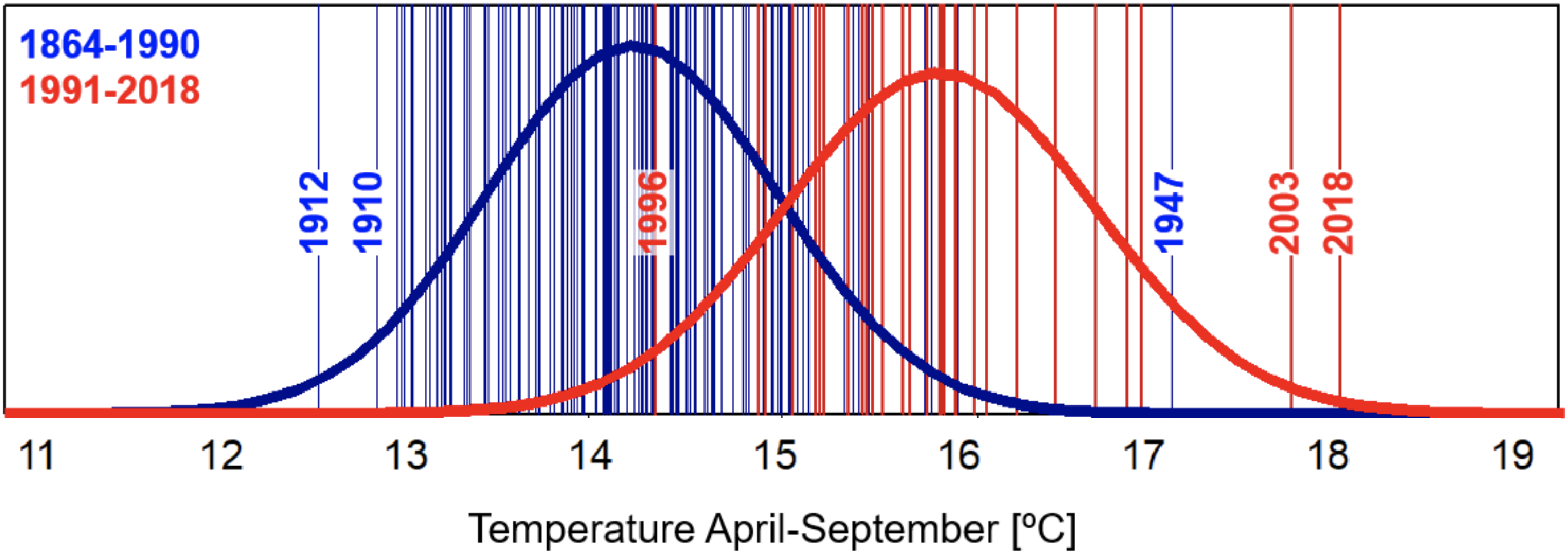
Observed temperatures April-September (credit: C. Schaer)



[in collaboration with Schemm, Fuhrer, Mishra]

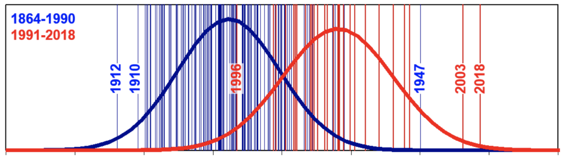
What about global warming? Can data-driven methods predict the future?

Observed temperatures April-September (credit: C. Schaer)



[in collaboration with Schemm, Fuhrer, Mishra]

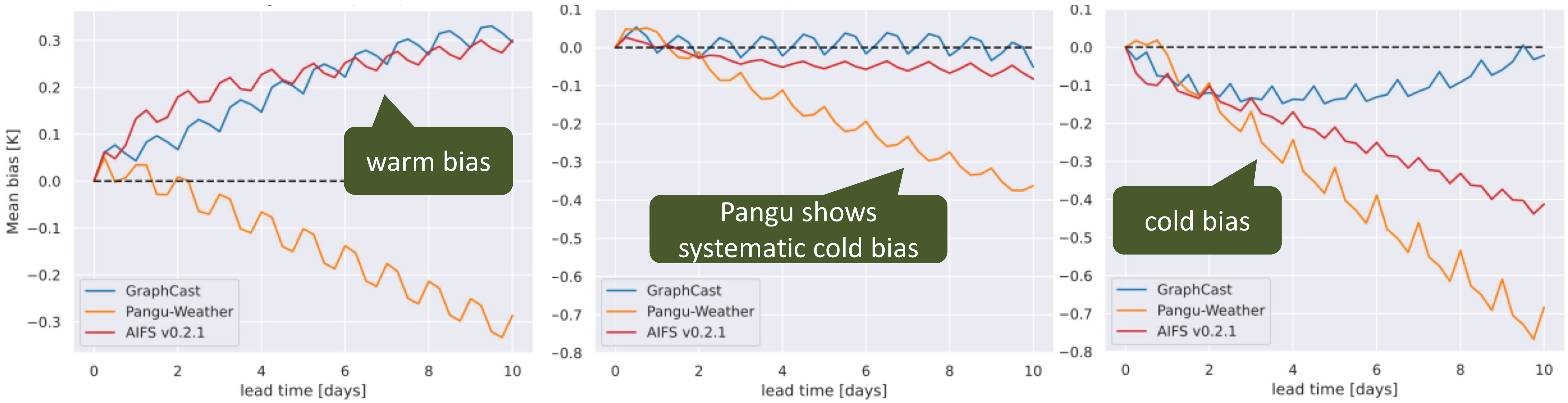
What about global warming? Can data-driven methods predict the future?



forecast bias for 2m temperature

with different starting conditions

starting in 1955 | starting in 2023 | starting in 2049



[in collaboration with Schemm, Fuhrer, Mishra]

N. Koldunov, T. Rackow, C. Lessig: Using AI-based numerical weather prediction models for climate applications https://epic.awi.de/id/eprint/59156/

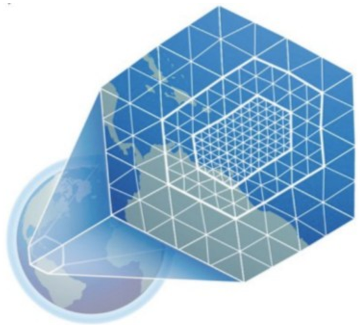
Swiss AI Climate Approach: More and Better Data and Computation

- **More data, more climate models, more variables at more vertical levels and higher resolution**
  - Data from 19 different climate models at 1.0° (~1000 simulated years): ~2.9B equivalent images
  - Potentially early 1km-resolution data (TBD)
  - Fine tuning with ERA-5 0.25° reanalysis data (~40 years of reanalysis): ~ 1.0B equivalent images
  - >4T tokens for pre-training a 160B parameter model
  - Fine tuning with km-scale regional data at 1-10 km resolution: data still under production at MeteoSwiss

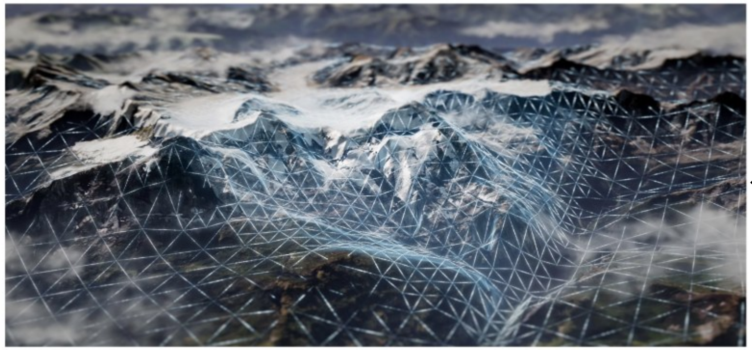
MPI-M

1. Climate model data

2. High-Res model data



3. Reanalysis data



MeteoSwiss

4. km-scale observation-constrained local data
(….because we don't want smooth fields)

Your tail network

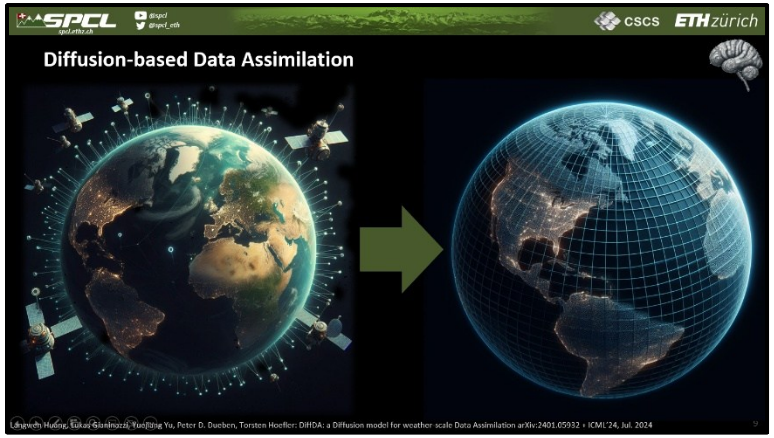
Your tail network

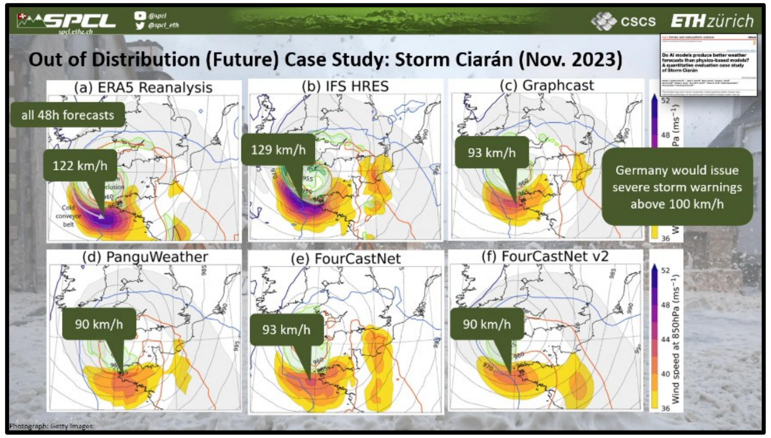
Your tail network

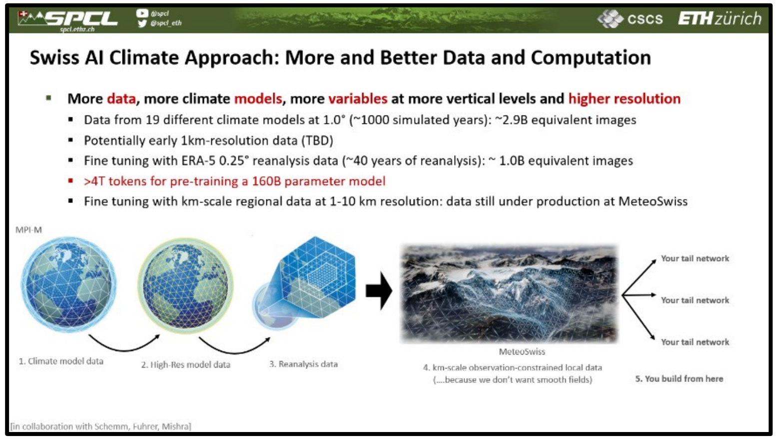
5. You build from here

[in collaboration with Schemm, Fuhrer, Mishra]

Summary and Key Points







All of ERA-5 (Earth' s 40-year history) on a USB-drive! Run your own analyses on your laptop!

More of SPCL's research:

youtube.com/@spcl — 150+ Talks

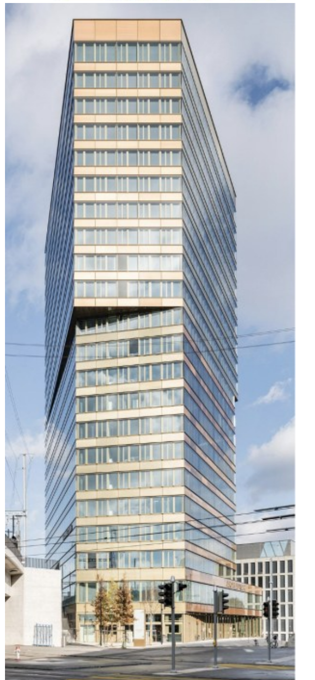
twitter.com/spcl_eth — 1.2K+ Followers

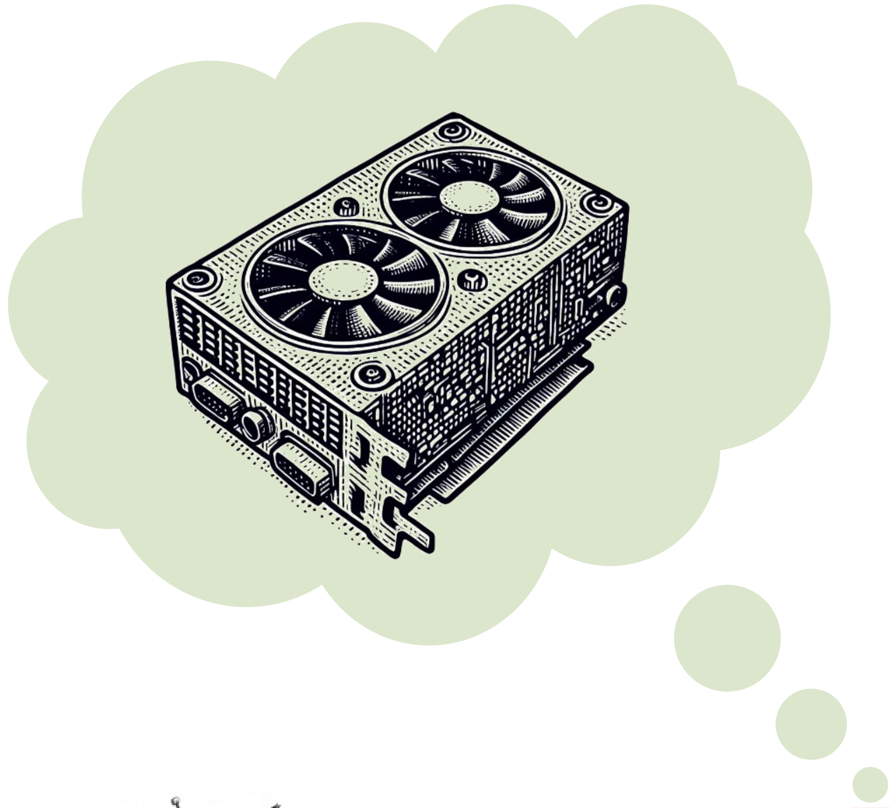
github.com/spcl — 2K+ Stars

… or spcl.ethz.ch

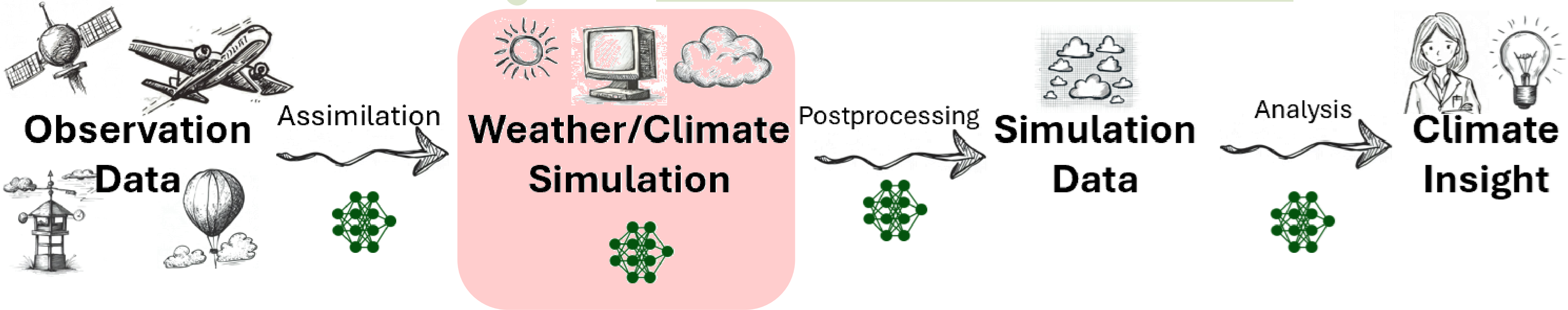
We are looking forward to a fruitful exchange!

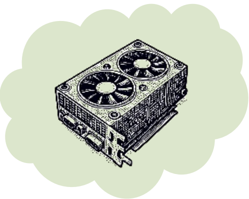
http://spcl.ethz.ch/Jobs/
http://spcl.ethz.ch/Visit/





```
!$ACC DATA &
!$ACC PRESENT(density1,energy1) &
!$ACC PRESENT(vol_flux_x,vol_flux_y,volume,mass_flux_x,mass_flux_y,vertexdx,vertexdy) &
!$ACC PRESENT(pre_vol,post_vol,ener_flux)

!$ACC KERNELS

  IF(dir.EQ.g_xdir) THEN

   IF(sweep_number.EQ.1)THEN

!$ACC LOOP INDEPENDENT
    DO k=y_min-2,y_max+2
!$ACC LOOP INDEPENDENT
     DO j=x_min-2,x_max+2
      pre_vol(j,k)=volume(j,k)+(vol_flux_x(j+1,k  )-vol_flux_x(j,k)+vol_flux_y(j  ,k+1)-vol_flux_y(j,k))
      post_vol(j,k)=pre_vol(j,k)-(vol_flux_x(j+1,k  )-vol_flux_x(j,k))
     ENDDO
    ENDDO
   ELSE
!$ACC LOOP INDEPENDENT
    DO k=y_min-2,y_max+2
!$ACC LOOP INDEPENDENT
     DO j=x_min-2,x_max+2
      pre_vol(j,k)=volume(j,k)+vol_flux_x(j+1,k)-vol_flux_x(j,k)
      post_vol(j,k)=volume(j,k)
     ENDDO
    ENDDO

   ENDIF
```



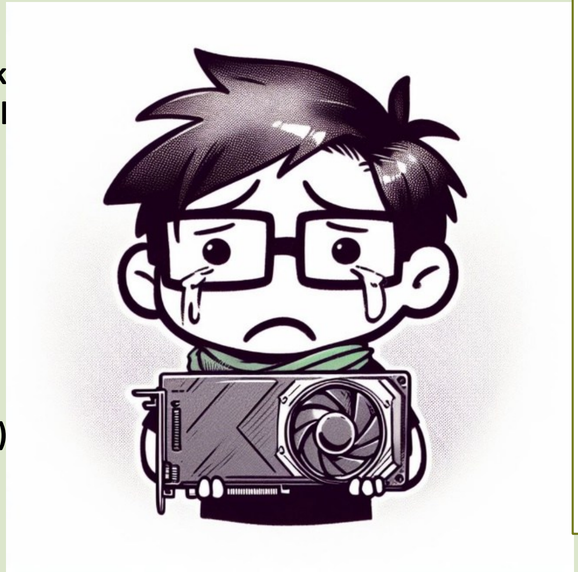


```
!$ACC DATA &
   !$ACC COPY(chunk%tiles(1)%field%density0)   &
   !$ACC COPY(chunk%tiles(1)%field%density1)   &
   !$ACC COPY(chunk%tiles(1)%field%energy0)    &
   !$ACC COPY(chunk%tiles(1)%field%energy1)    &
   !$ACC COPY(chunk%tiles(1)%field%pressure)   &
   !$ACC COPY(chunk%tiles(1)%field%soundspeed) &
   !$ACC COPY(chunk%tiles(1)%field%viscosity)  &
   !$ACC COPY(chunk%tiles(1)%field%xvel0)      &
   !$ACC COPY(chunk%tiles(1)%field%yvel0)      &
   !$ACC COPY(chunk%tiles(1)%field%xvel1)      &
   !$ACC COPY(chunk%tiles(1)%field%yvel1)      &
   !$ACC COPY(chunk%tiles(1)%field%vol_flux_x) &
   !$ACC COPY(chunk%tiles(1)%field%vol_flux_y) &
   !$ACC COPY(chunk%tiles(1)%field%mass_flux_x)&
   !$ACC COPY(chunk%tiles(1)%field%mass_flux_y)&
   !$ACC COPY(chunk%tiles(1)%field%volume)     &
   !$ACC COPY(chunk%tiles(1)%field%work_array1)&
   !$ACC COPY(chunk%tiles(1)%field%work_array2)&
   !$ACC COPY(chunk%tiles(1)%field%work_array3)&
   !$ACC COPY(chunk%tiles(1)%field%work_array4)&
   !$ACC COPY(chunk%tiles(1)%field%work_array5)&
   !$ACC COPY(chunk%tiles(1)%field%work_array6)&
   !$ACC COPY(chunk%tiles(1)%field%work_array7)&
   !$ACC COPY(chunk%tiles(1)%field%cellx)      &
   !$ACC COPY(chunk%tiles(1)%field%celly)      &
   !$ACC COPY(chunk%tiles(1)%field%celldx)     &
   !$ACC COPY(chunk%tiles(1)%field%celldy)     &
   !$ACC COPY(chunk%tiles(1)%field%vertexx)    &
   !$ACC COPY(chunk%tiles(1)%field%vertexdx)   &
   !$ACC COPY(chunk%tiles(1)%field%vertexy)    &
   !$ACC COPY(chunk%tiles(1)%field%vertexdy)   &
   !$ACC COPY(chunk%tiles(1)%field%xarea)      &
   !$ACC COPY(chunk%tiles(1)%field%yarea)      &
   !$ACC COPY(chunk%left_snd_buffer)   &
   !$ACC COPY(chunk%left_rcv_buffer)   &
   !$ACC COPY(chunk%right_snd_buffer)  &
   !$ACC COPY(chunk%right_rcv_buffer)  &
   !$ACC COPY(chunk%bottom_snd_buffer) &
   !$ACC COPY(chunk%bottom_rcv_buffer) &
   !$ACC COPY(chunk%top_snd_buffer)    &
   !$ACC COPY(chunk%top_rcv_buffer)
```

```
,mass_flux_x,mass_flux_y,vertexdx,vertexdy) &
1,k
+1,l
l,k)
```

Sloccount *f90: 6,440

!$ACC: 833 (13%)

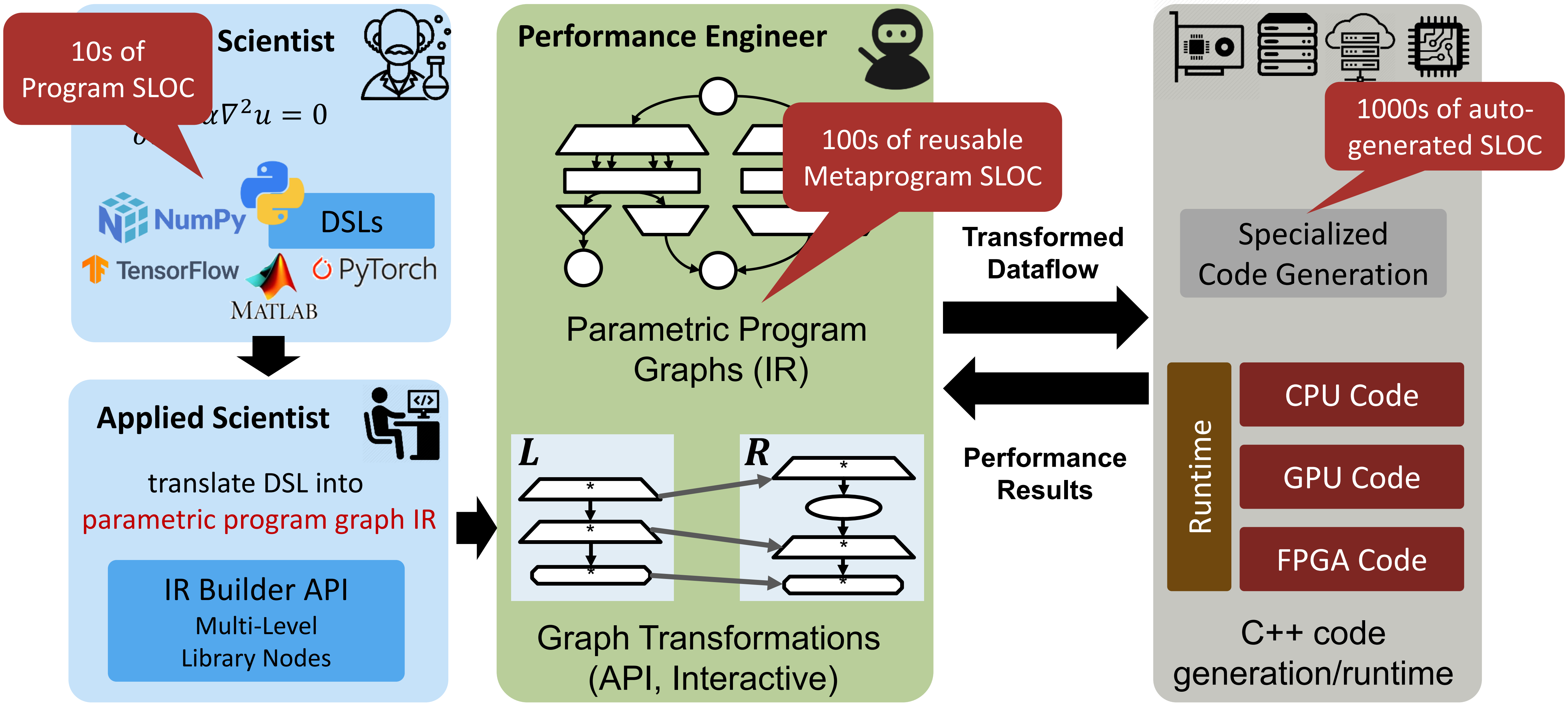


Heitlager et al.: A Practical Model for Measuring Maintainability

ISO 9126 maintainability	source code properties: volume	complexity per unit	duplication	unit size	unit testing
analysability	x		x	x	x
changeability		x	x		
stability					x
testability		x		x	x

ICON: about 1% OpenMP + 3% OpenACC

Performance Metaprogramming for Optimization and Performance Portability



Ben-Nun, de Fine Licht, Ziogas, TH: Stateful Dataflow Multigraphs: A Data-Centric Model for High-Performance Parallel Programs, SC19

Pace in DaCe for Performance Metaprogramming – 12k SLOC Python

AI-based Transfer Tuning to the Rescue!

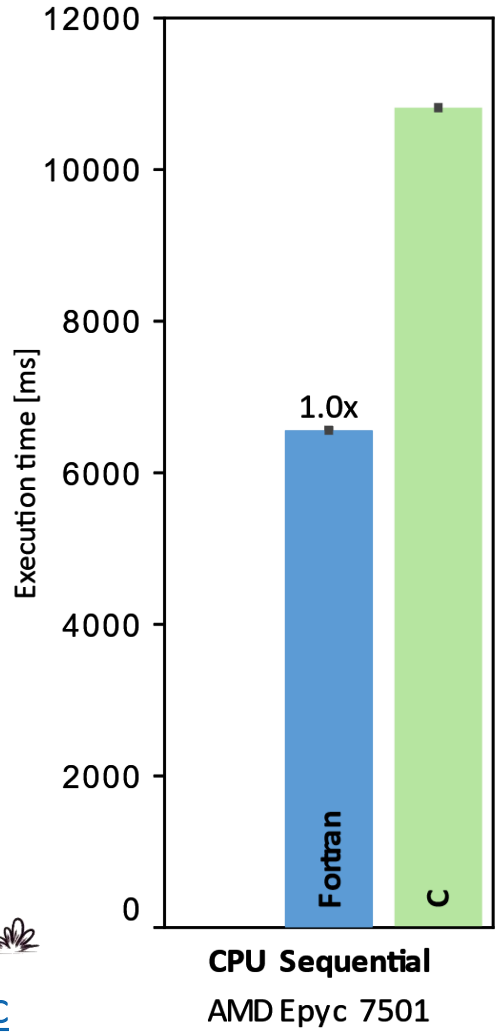
Ben-Nun et al.: Productive Performance Engineering for Weather and Climate Modeling with Python, SC22

Another real production code … ECMWF's CLOUDSC

```fortran

SUBROUTINE CLOUDSC &
 !---input
 & (KIDIA,    KFDIA,    KLON,    KLEV, &
 & PTSPHY,&
 & PT, PQ, tendency_cml,tendency_tmp,tendency_loc, &
 & PVFA, PVFL, PVFI, PDYNA, PDYNL, PDYNI, &
 & PHRSW,    PHRLW,&
 & PVERVEL,  PAP,      PAPH,&
 & PLSM,     LDCUM,    KTYPE, &
 & PLU,      PLUDE,    PSNDE,    PMFU,     PMFD,&
 !---prognostic fields
 & PA,&
 & PCLV,  &
 & PSUPSAT,&
!-- arrays for aerosol-cloud interactions
!!! & PQAER,    KAER, &
 & PLCRIT_AER,PICRIT_AER,&
 & PRE_ICE,&
 & PCCN,     PNICE,&
 !---diagnostic output
 & PCOVPTOT, PRAINFRAC_TOPRFZ,&
 !---resulting fluxes
 & PFSQLF,   PFSQIF ,  PFCQNNG,  PFCQLNG,&
 & PFSQRF,   PFSQSF ,  PFCQRNG,  PFCQSNG,&
 & PFSQLTUR, PFSQITUR , &
 & PFPLSL,   PFPLSN,   PFHPSL,   PFHPSN, KFLDX, &
 & YDCST, YDTHF, YDECLDP)
```

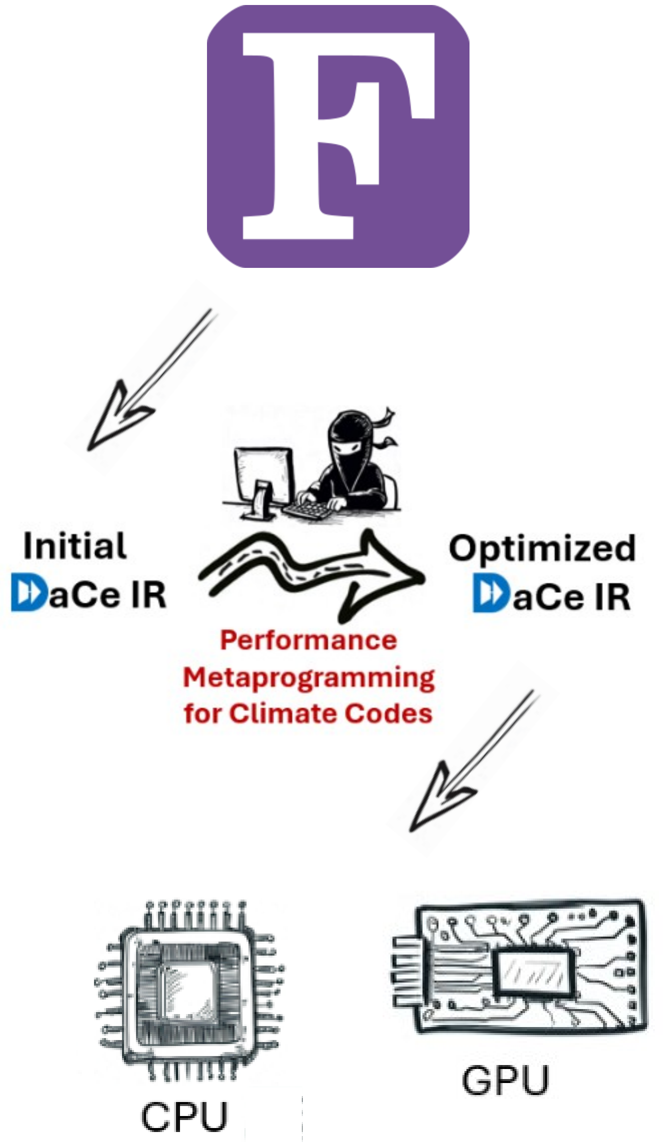
… variable setup/initialization until line 500 ;-)

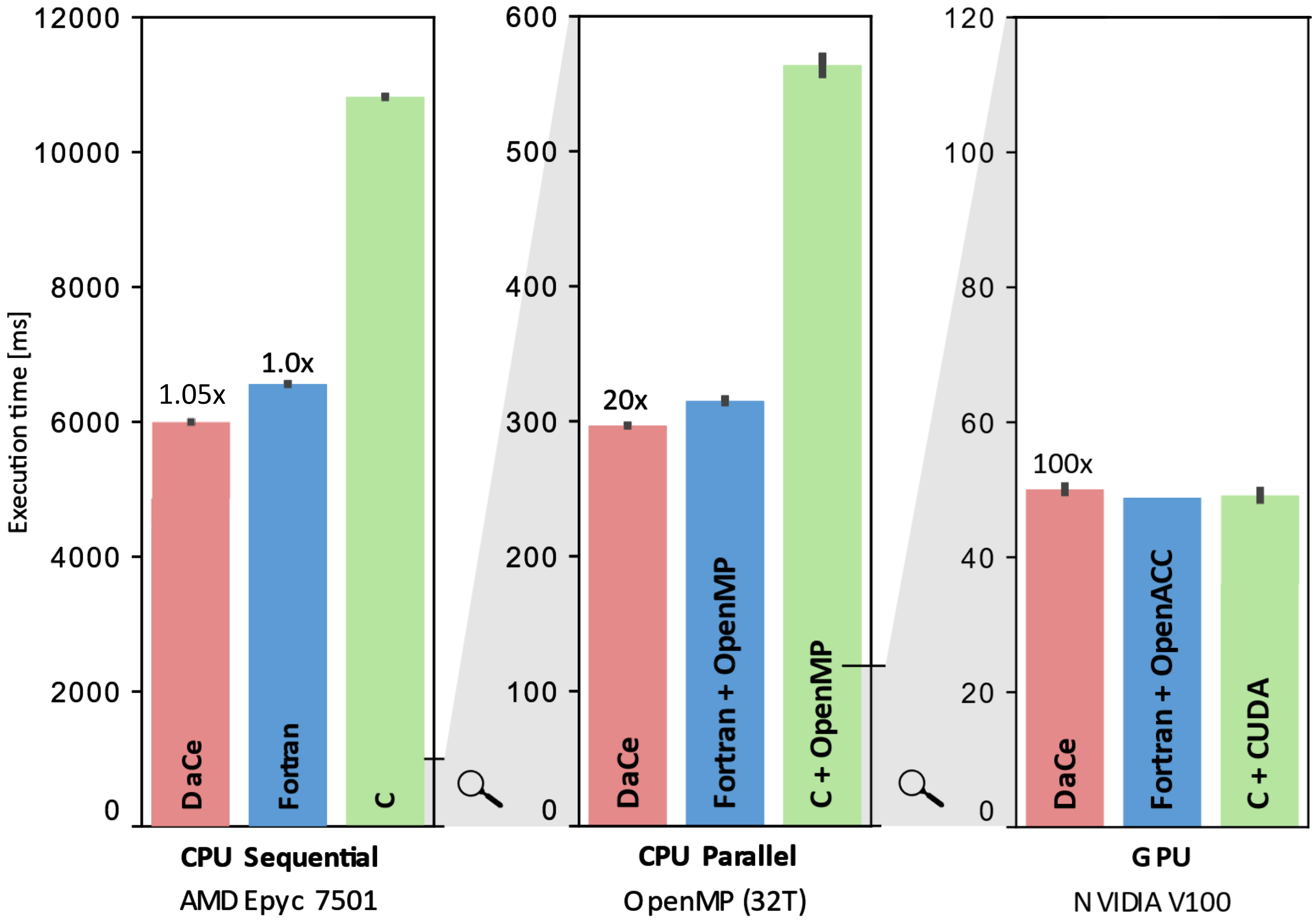
- **Cloud Microphysics of IFS**
  - Resolve sub-grid features
  - Original 2,525 SLOC of **Fortran 95**
- **Rewritten for performance portability benchmarking (optimization took months!)**
  - 2,635 SLOC C
  - 2,610 SLOC C++/CUDA



https://github.com/ecmwf-ifs/dwarf-p-cloudsc

Performance Metaprogramming – from the unchanged CLOUDSC Fortran code!





The three pillars of AI in Climate Sciences

Data

- Unstructured observation data
- Structured simulation data

Combine both to train models

- Learn physics and data-driven prediction

Compute

- AI models require accelerated high-performance computing for training

Accelerate AI computations

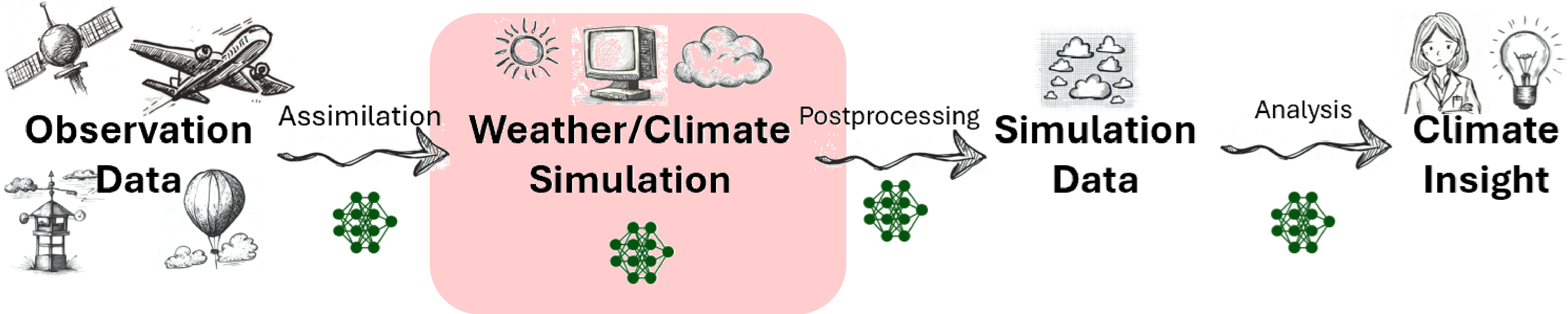
- Re-use infrastructure from LLMs and related generative AI methods (GNN, CNN)

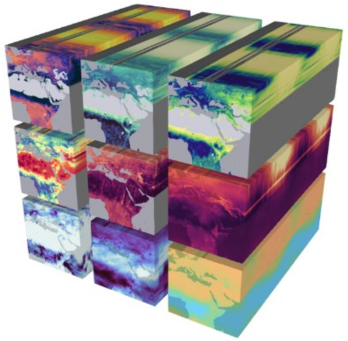
Models

- Models need to provide the right structural bias/prior

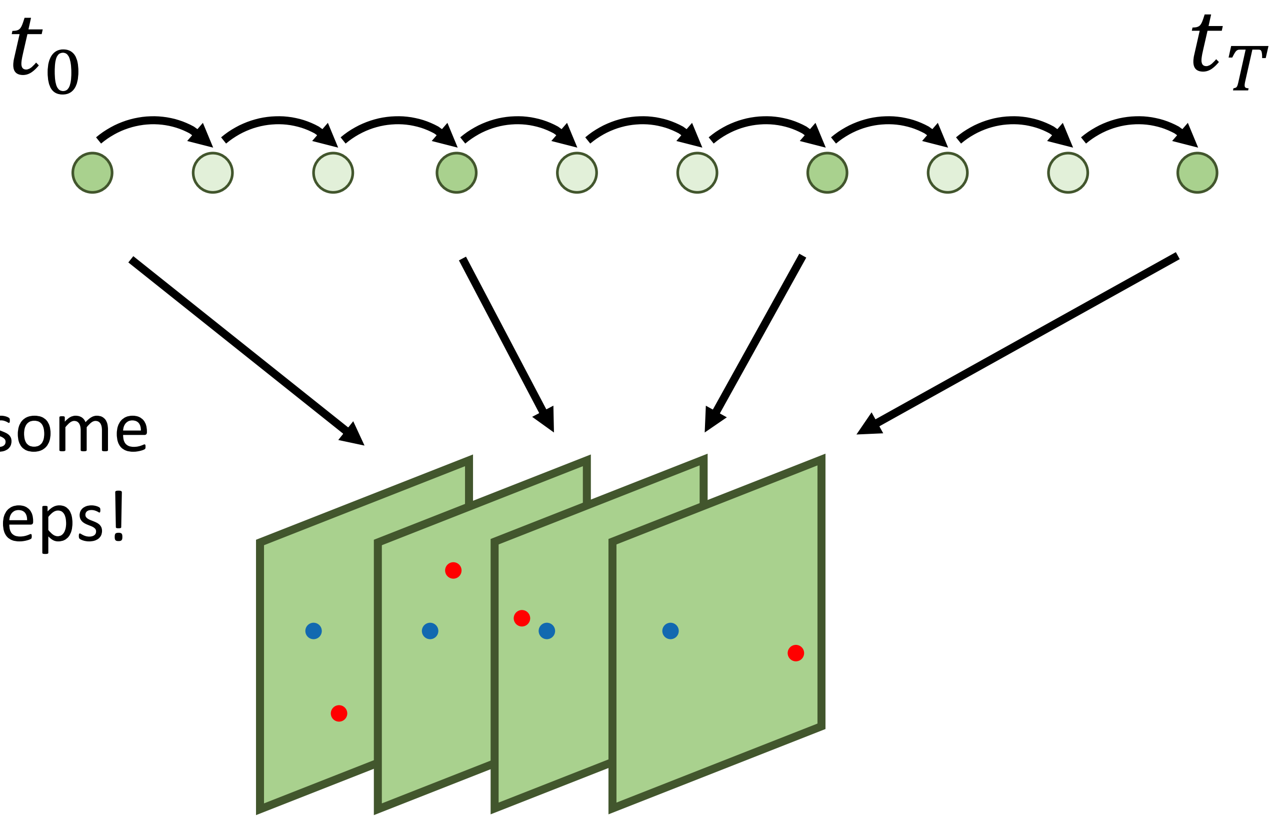
Develop better AI methods

- Step 1: use generative AI models: tformer, CNN, GNN, Diffusion, etc.
- Step 2: use automatically parameterized physics-based models encoding equations





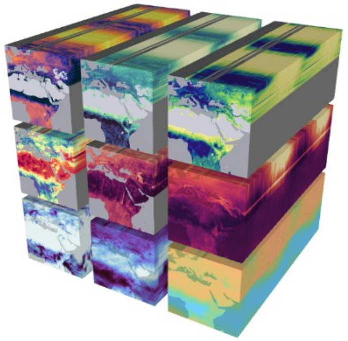
Simulation runs time-stepping forward

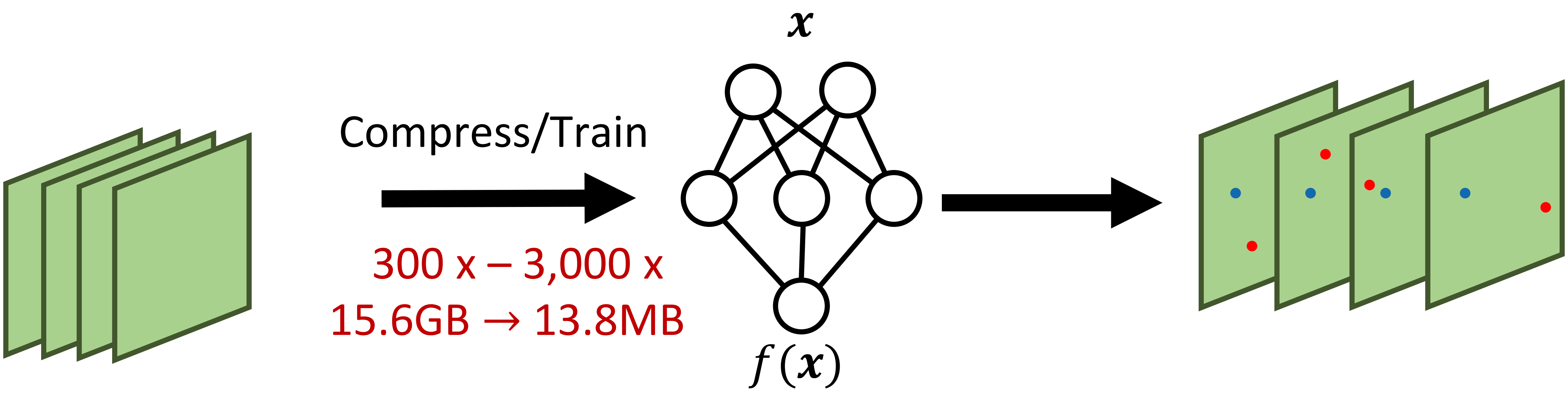


Store some timesteps!

Analysis access pattern is often strided or even random

Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23



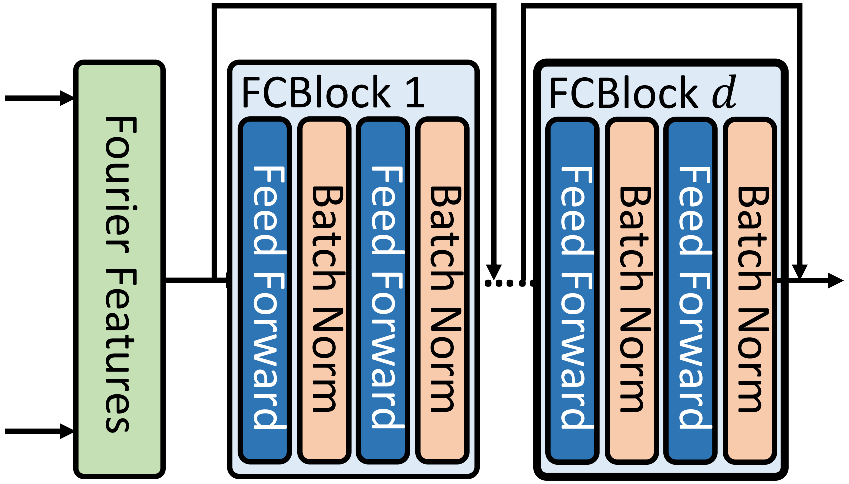


Multidimensional Data

Neural Representation

Analysis Queries

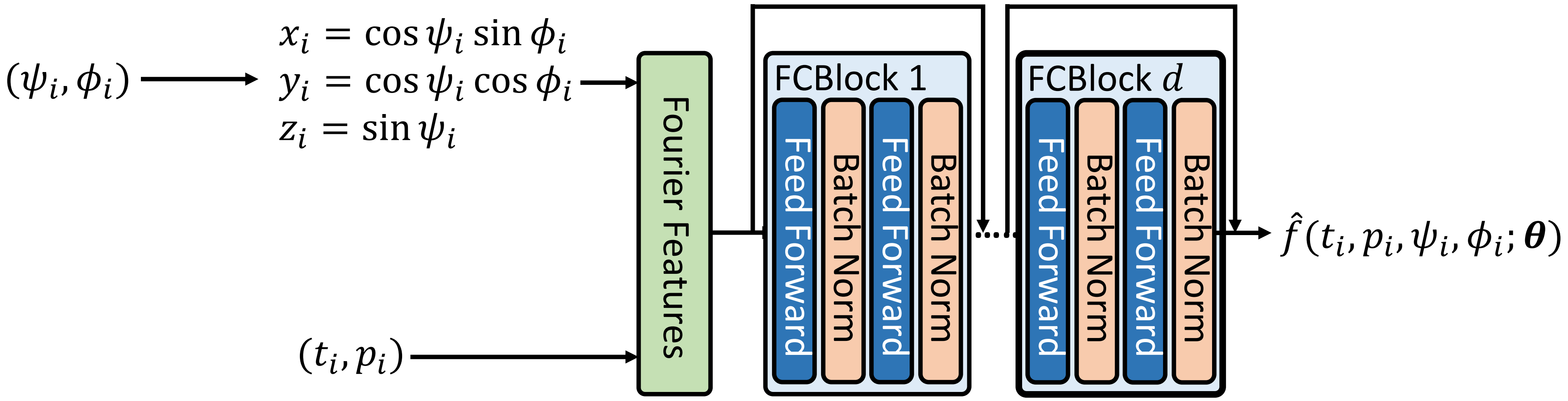
Neural Network Structure



Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

Neural Network Structure

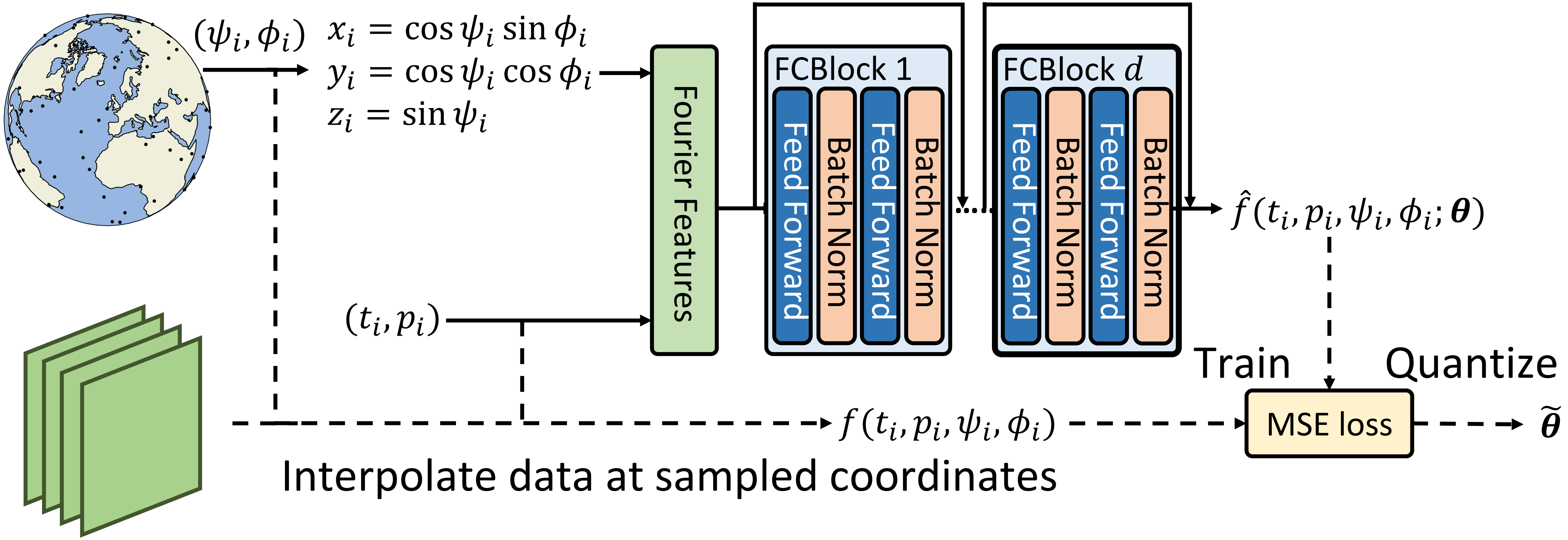
Decompression / Inference



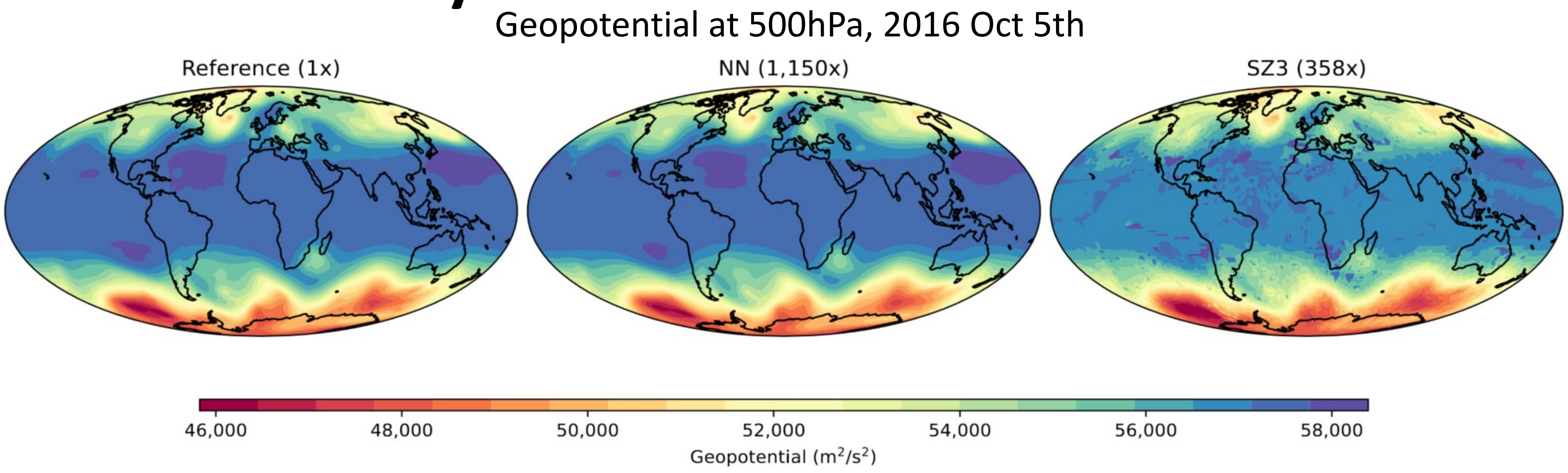
- On-demand decompression
- Fully utilize GPUs

Neural Network Structure

Compression / Training



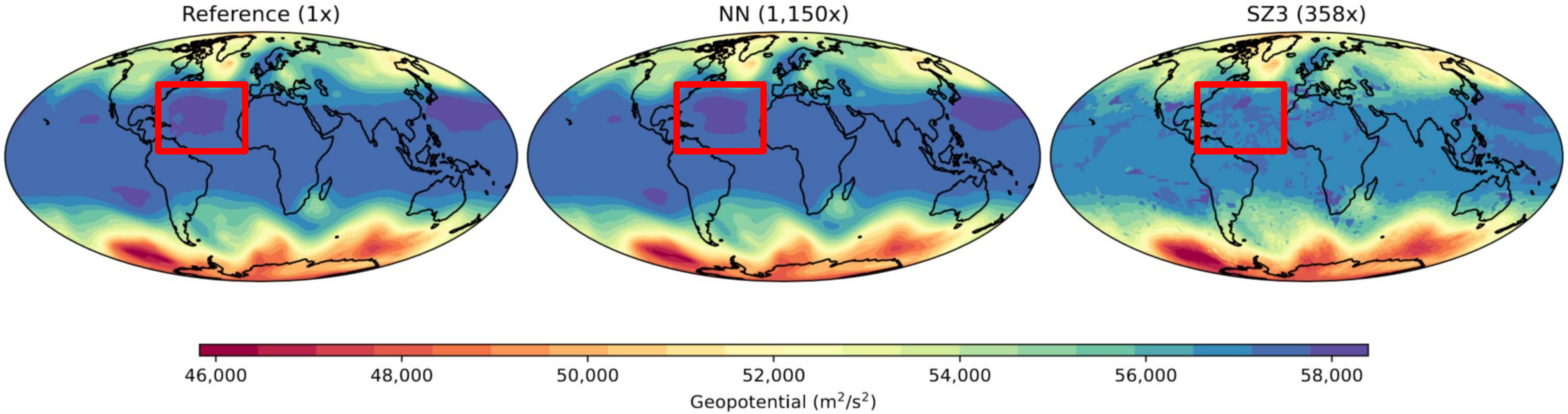
Evaluation: Case Study



Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

Evaluation: Case Study

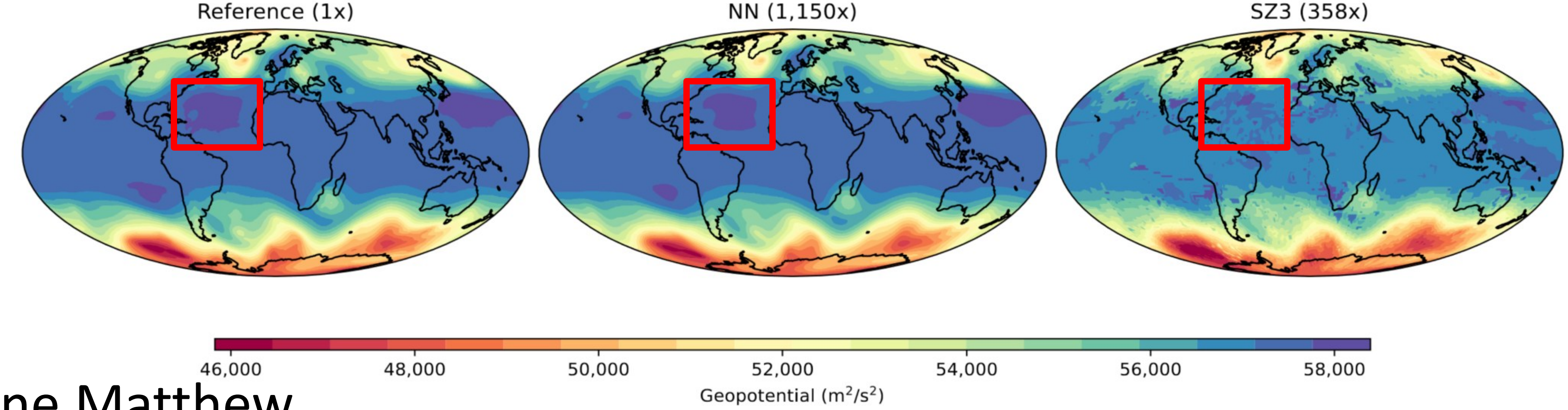
Geopotential at 500hPa, 2016 Oct 5th



Preserves general shapes of important events and average values without introducing significant artifacts

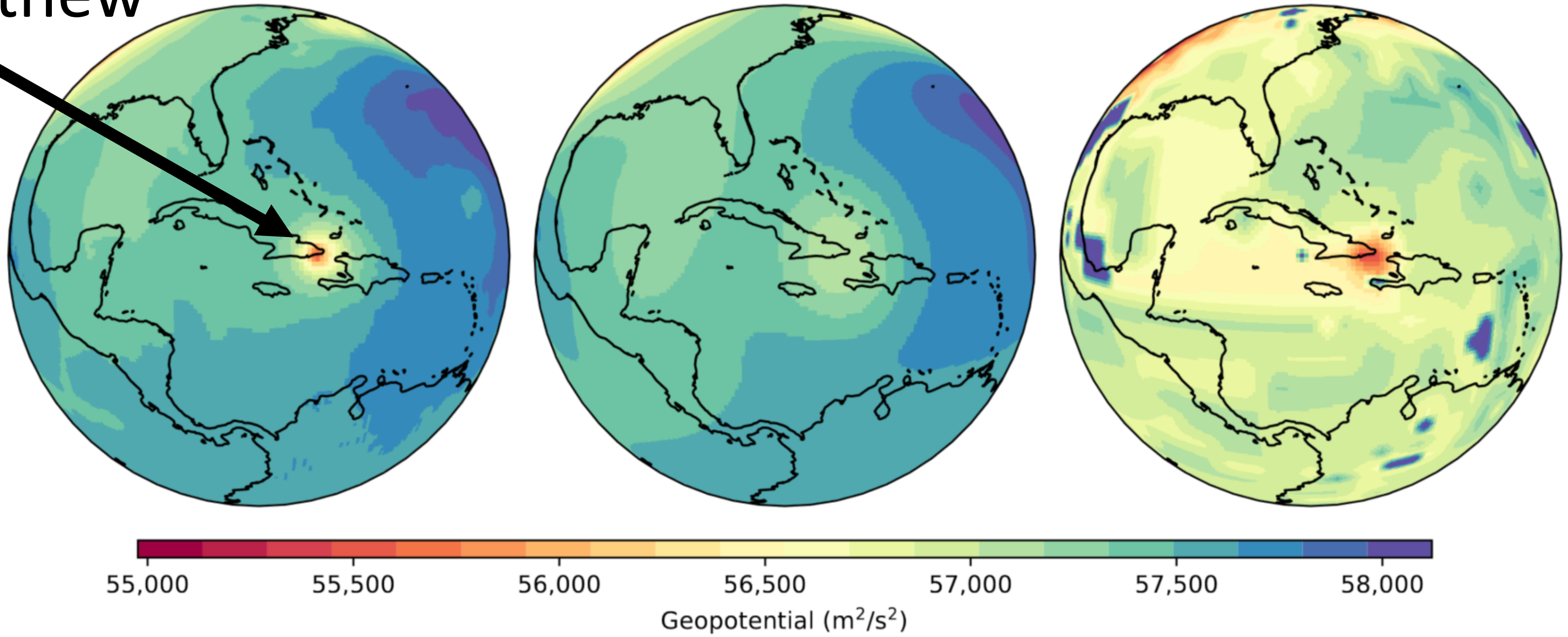
Evaluation: Case Study

Geopotential at 500hPa, 2016 Oct 5th



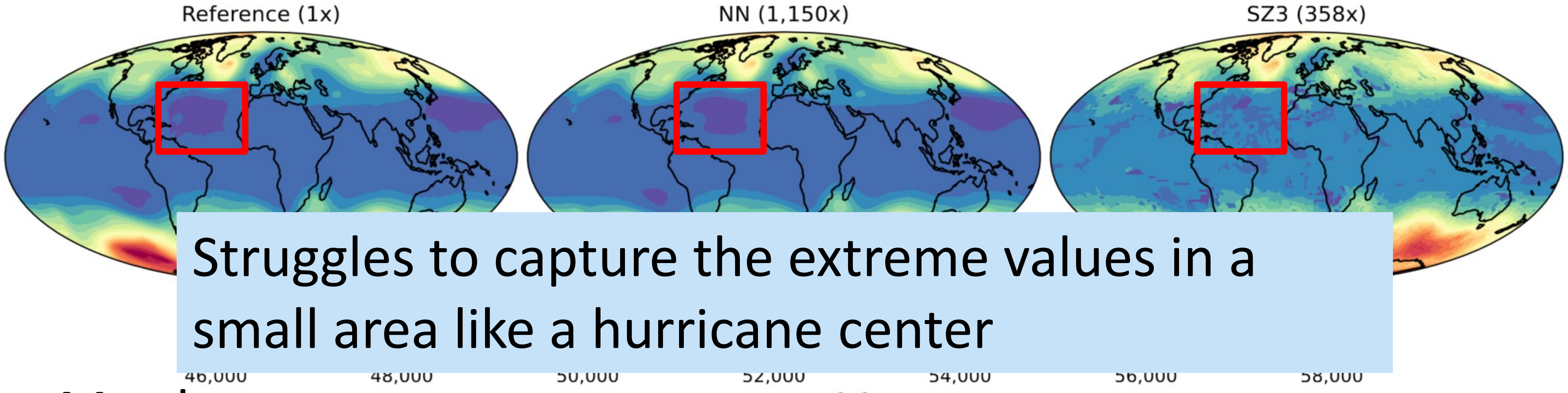
Hurricane Matthew

16.5bn damage
603 fatalities



Evaluation: Case Study

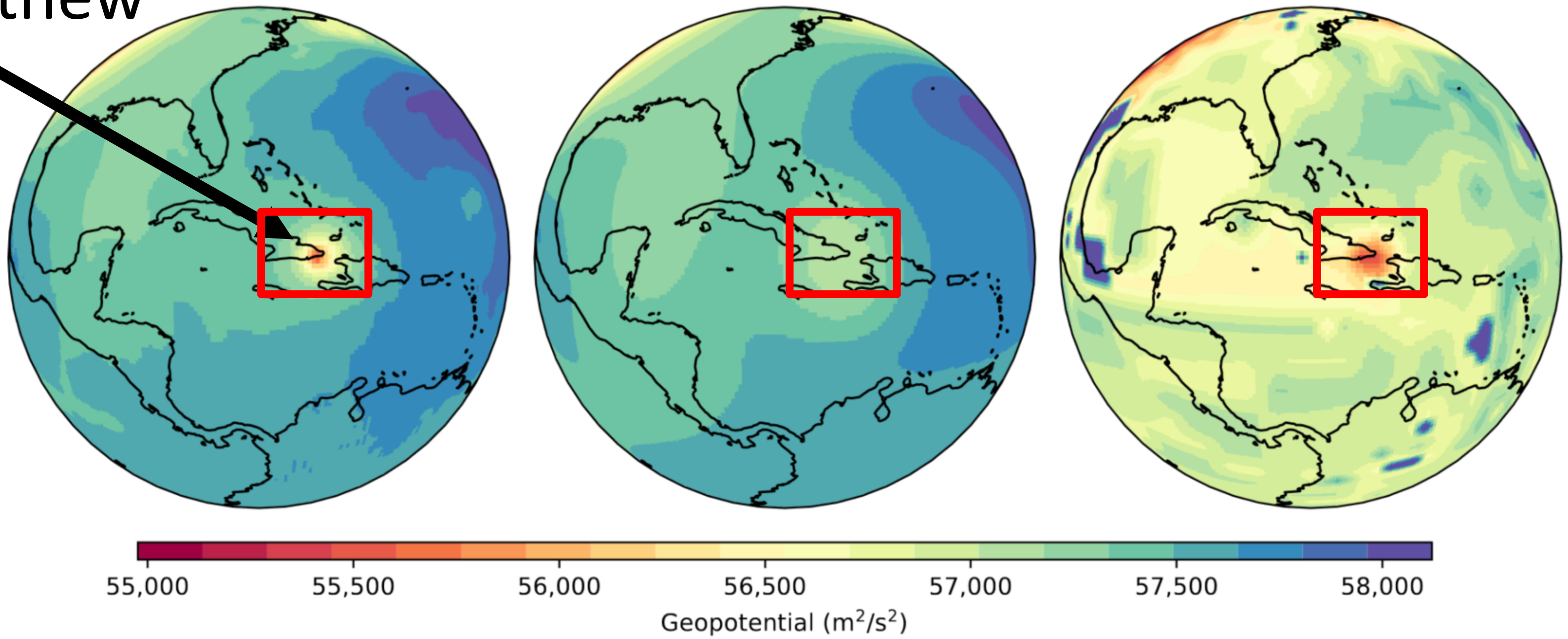
Geopotential at 500hPa, 2016 Oct 5th



Struggles to capture the extreme values in a small area like a hurricane center

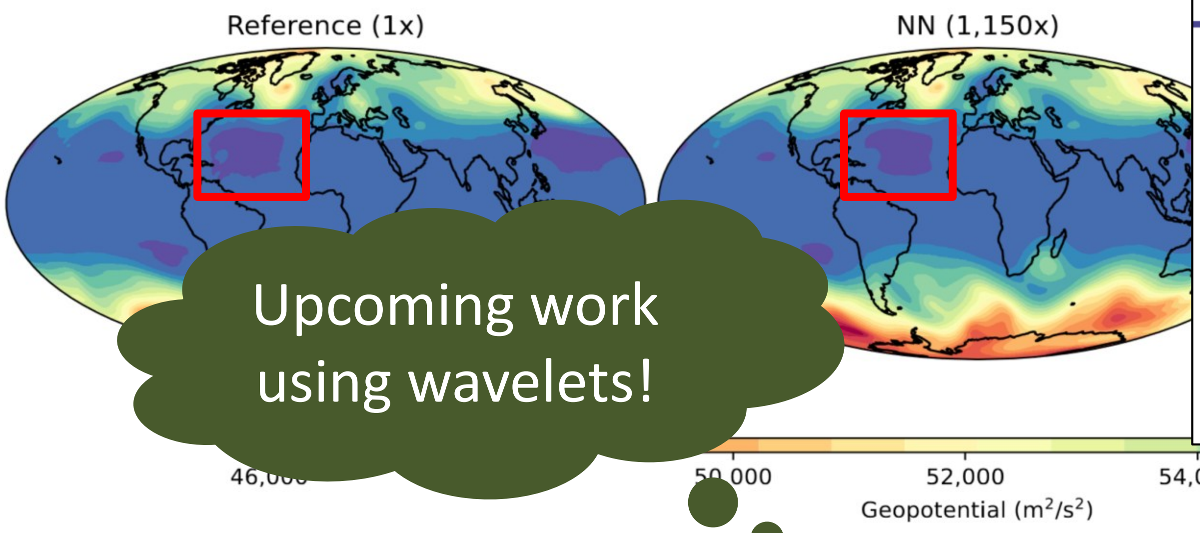
Hurricane Matthew

16.5bn damage
603 fatalities



Evaluation: Case Study

Geopotential at 500hPa, 2016



Upcoming work using wavelets!

nature computational science

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature computational science > articles > article

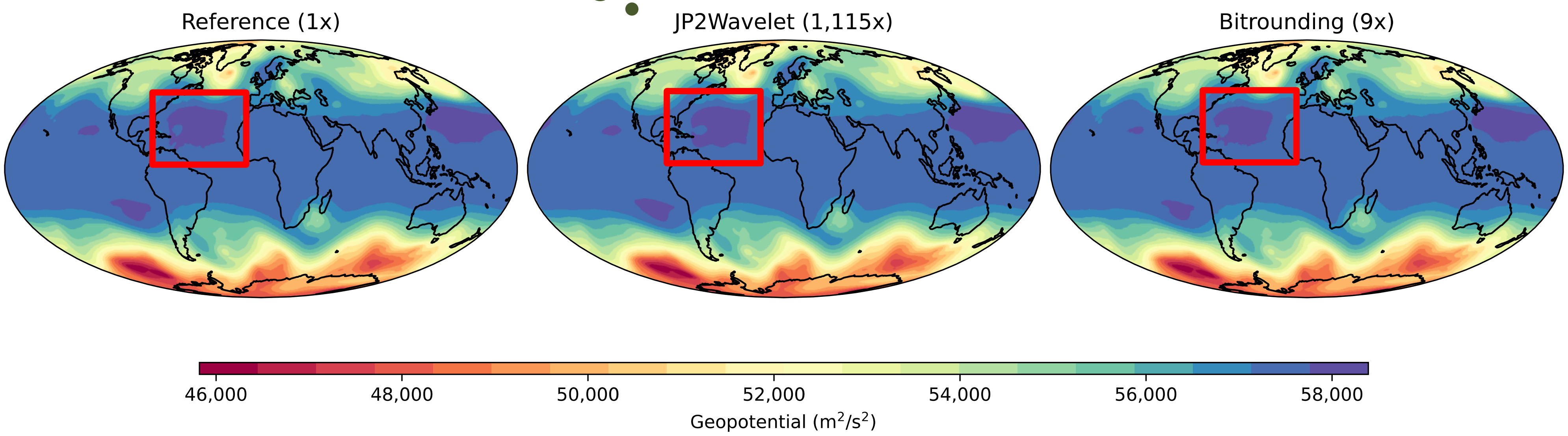
Article | Open access | Published: 25 November 2021

Compressing atmospheric data into its real information content

Milan Klöwer, Miha Razinger, Juan J. Dominguez, Peter D. Düben & Tim N. Palmer

Nature Computational Science **1**, 713–724 (2021) | Cite this article

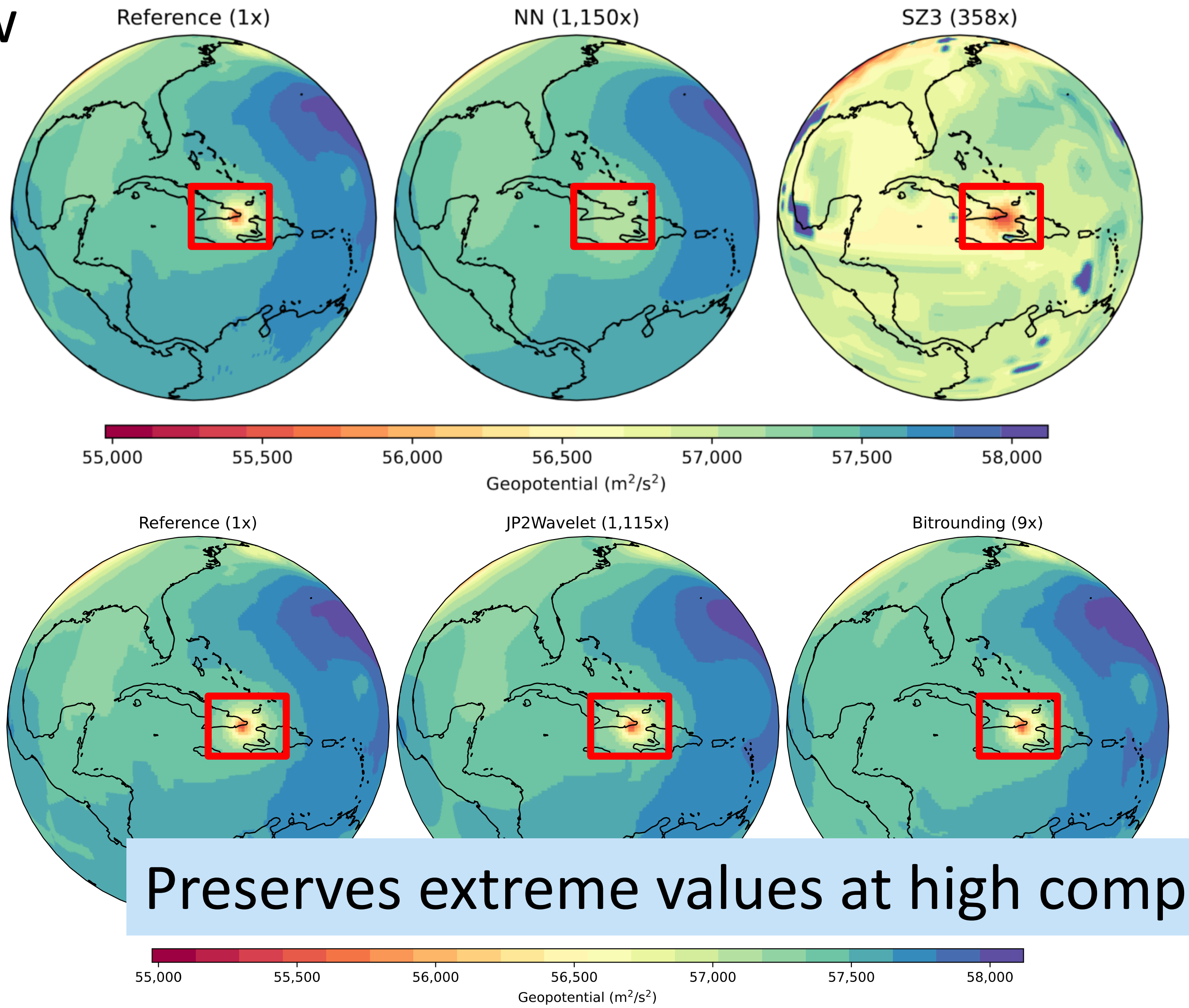
11k Accesses | **12** Citations | **64** Altmetric | Metrics



Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23

Evaluation: Case Study

Hurricane Matthew



Preserves extreme values at high compression ratio

Langwen Huang, Torsten Hoefler: Compressing Multidimensional Weather and Climate Data into Neural Networks, arXiv:2210.12538 + ICLR'23