Google DeepMind

Global weather forecasting with GraphCast

GraphCast team

Presenters:

Ferran Alet

Alvaro Sanchez-Gonzalez

Large-scale deep learning for the Earth system workshop

4 September 2023

Bonn Germany

GraphCast: Learning skillful medium-range global weather forecasting

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, Peter Battaglia


Remi Lam


Alvaro Sanchez


Matthew Willson


Peter Wirnsberger


Meire Fortunato


Ferran Alet


Suman Ravuri


Timo Ewalds


Zach Eaton-Rosen


Alex Merose


Stephan Hoyer


George Holland


Oriol Vinyals


Jacklynn Stott

arXiv

https://arxiv.org/abs/2212.12794


Alexander Pritzel

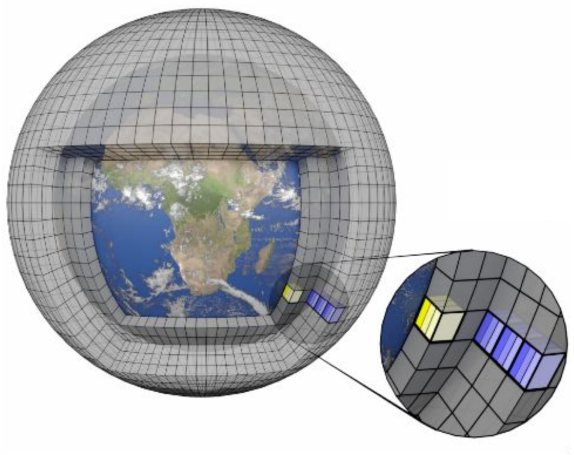

Shakir Mohamed


Peter Battaglia

https://github.com/deepmind/graphcast

Improving global weather forecasting with ML

We predict Earth's surface & atmospheric (3D) weather, **10 days ahead**, at **0.25°** latitude/longitude resolution.



ML can learn 'approximate physics' directly from data.

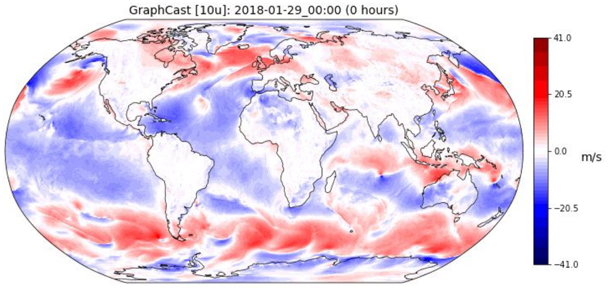
We learn more accurate, and more efficient, models than SOTA NWPs.

GraphCast: the best performing mid-range ML model,

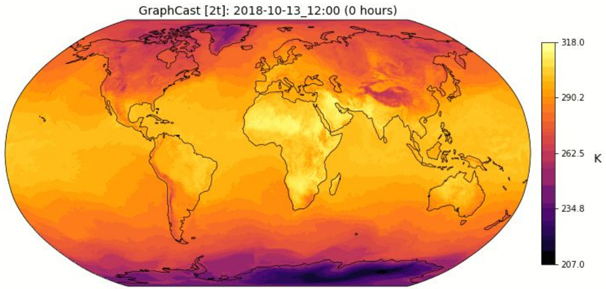
A careful comparison with SOTA NWP model (HRES),

Better prediction and planning for extreme events.

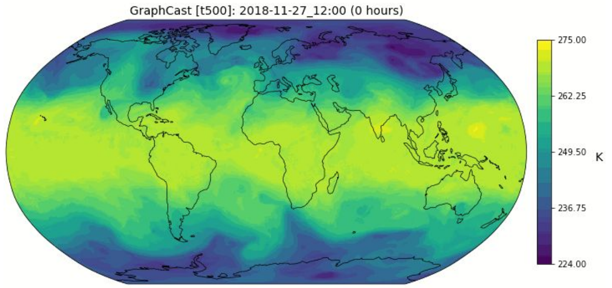
Surface E-W wind



Surface temperature



Temperature @ 500 hPa



3 of 227 weather variables modeled

Why now? 3 key factors

Data

ERA5 reanalysis

Assimilation is still NWP-based!

Massive dataset (40+ years)

High-quality data

Compute

32 TPUs

Increase in scale in ML models

Parallel compute

1 TPU-minute vs 10k CPU

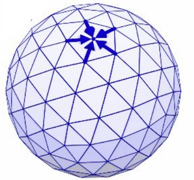
Deep Learning algorithms

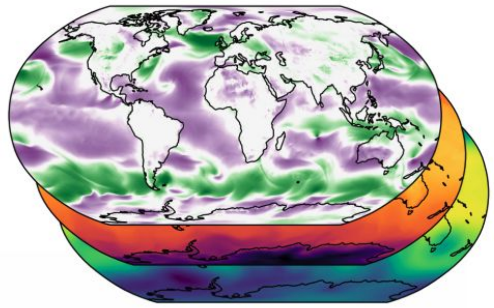
Graph Neural Networks

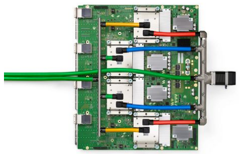
Encoded inductive biases

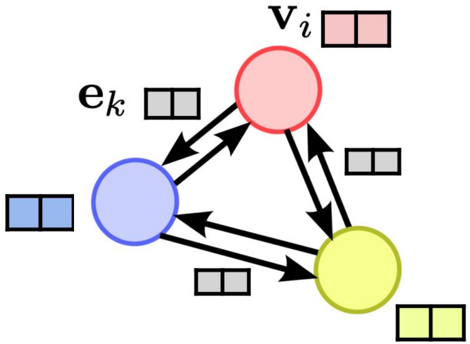
Locality

Equivariance





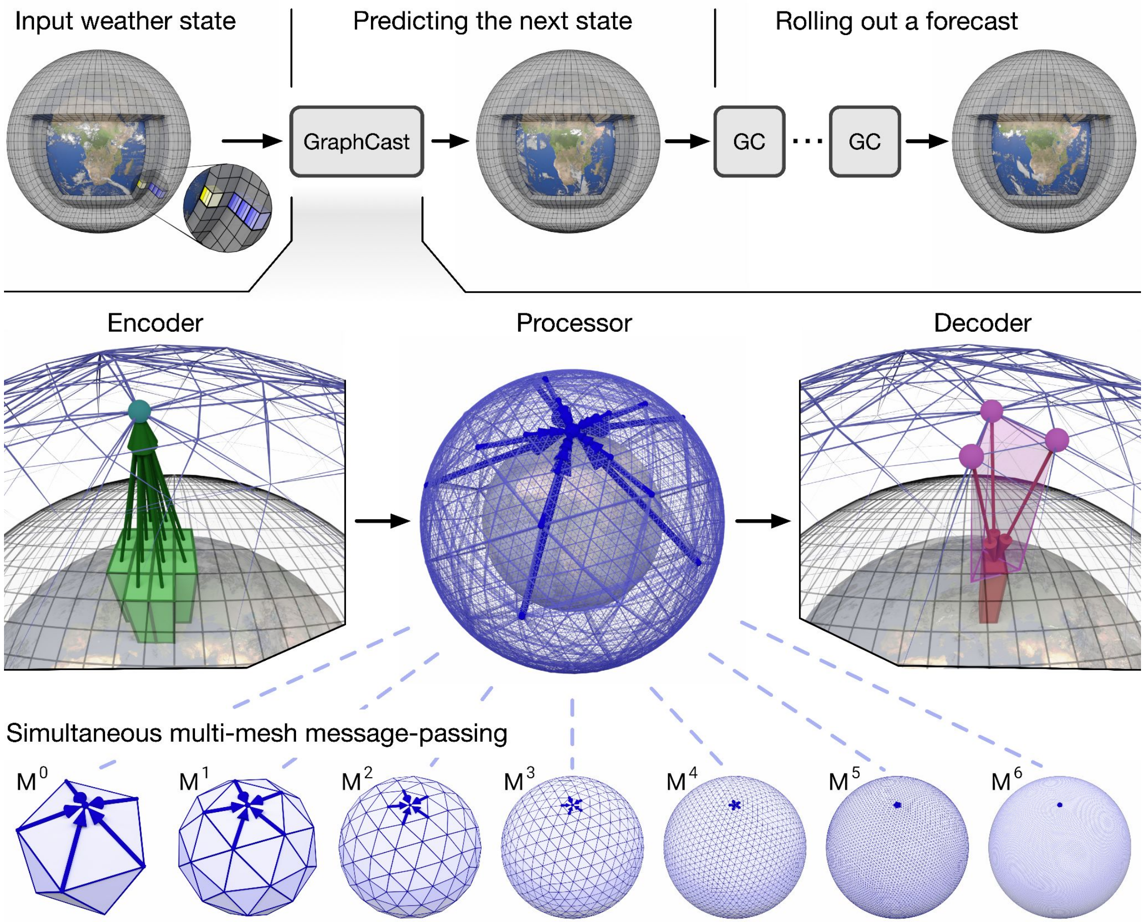




A variety of deep learning approaches have had similar goals:

CNNs: Weyn et al., Rasp&Thuerey GNNs: Keisler FNOs: FourCastNet Transformers: PangWu, ClimaX, FengWu ...

GraphCast: a learned simulator based on GNNs



Autoregressive: network predicts 6h steps and is applied repeatedly

3 main components:

- **Encoder** maps inputs to a “multi-mesh”
- **Processor** message-passing over mesh
- **Decoder** maps back to the state space

Multi-mesh: iteratively refined icosahedron

41k nodes, 328k edges

Training loss

$$\mathcal{L}_{\text{MSE}} = \underbrace{\frac{1}{|D_{\text{batch}}|}\sum_{d_0\in D_{\text{batch}}}}_{\text{forecast date-time}} \underbrace{\frac{1}{T_{\text{train}}}\sum_{\tau\in 1:T_{\text{train}}}}_{\text{lead time}} \underbrace{\frac{1}{|G_{0.25^\circ}|}\sum_{i\in G_{0.25^\circ}}}_{\text{spatial location}} \underbrace{\sum_{j\in J}}_{\text{variable-level}} \qquad \underbrace{(\hat{x}^{d_0+\tau}_{i,j} - x^{d_0+\tau}_{i,j})^2}_{\text{squared error}}$$

Area of latitude-longitude grid cell

Variable-level weight

Proportional to pressure-level

Inverse residual variance:

Predicting current state ⟶ loss=1

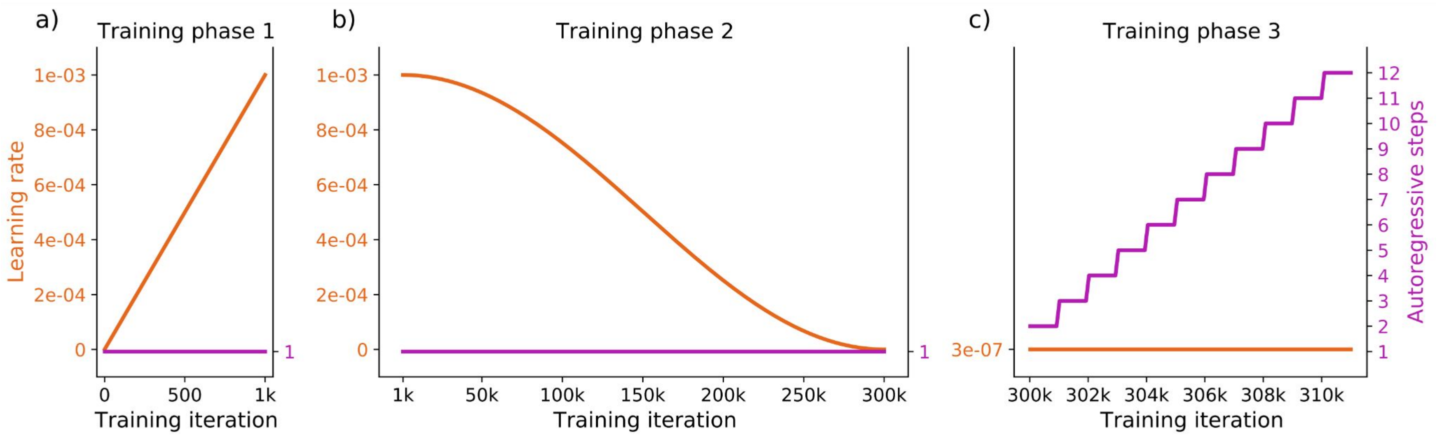
Autoregressive training

Most training is done at 1 autoregressive step (single forward pass)

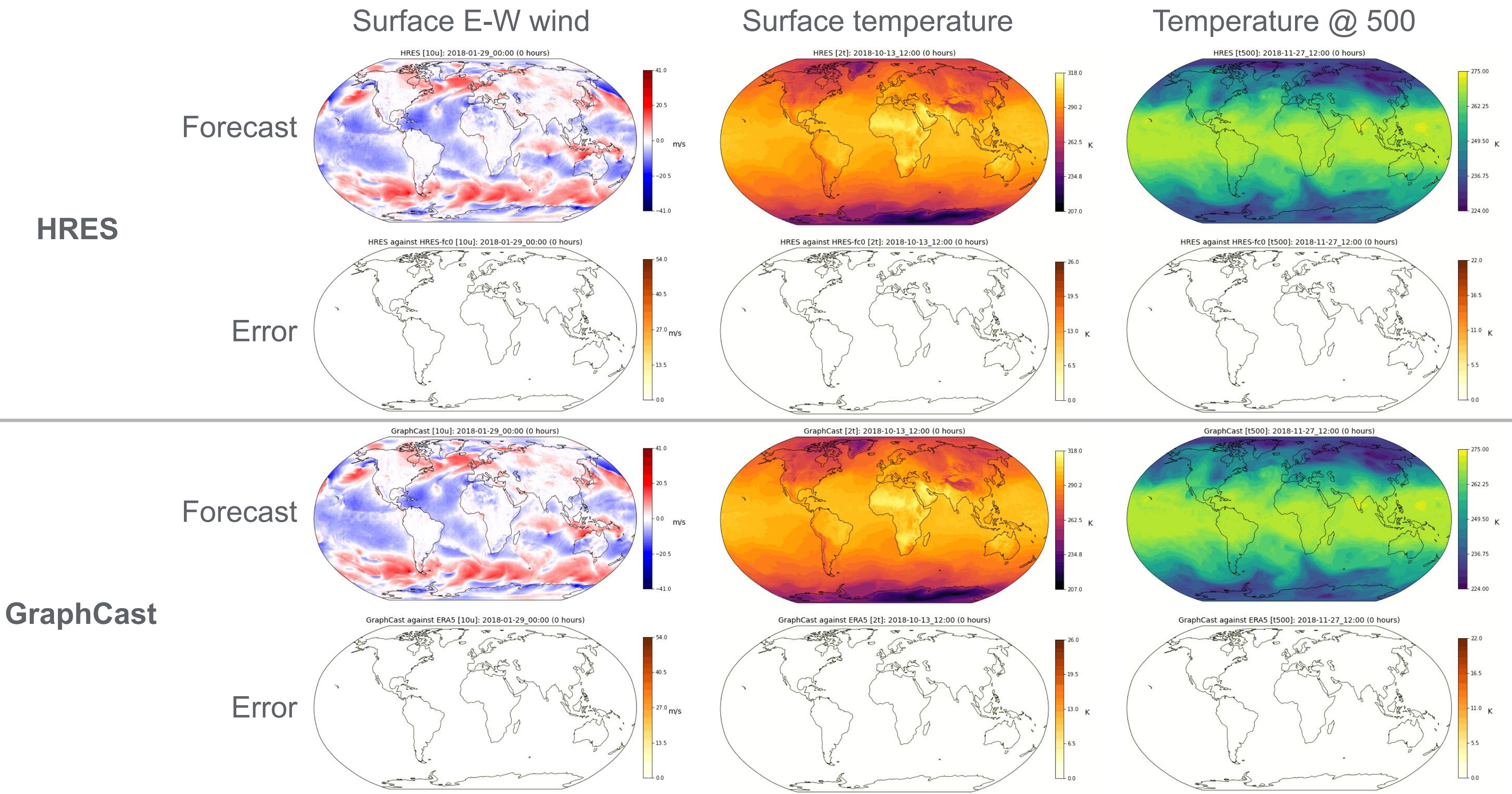
- Faster
- Compounding bad models leads to instabilities
- 3 weeks

Fine-tuning stage up to 3 days

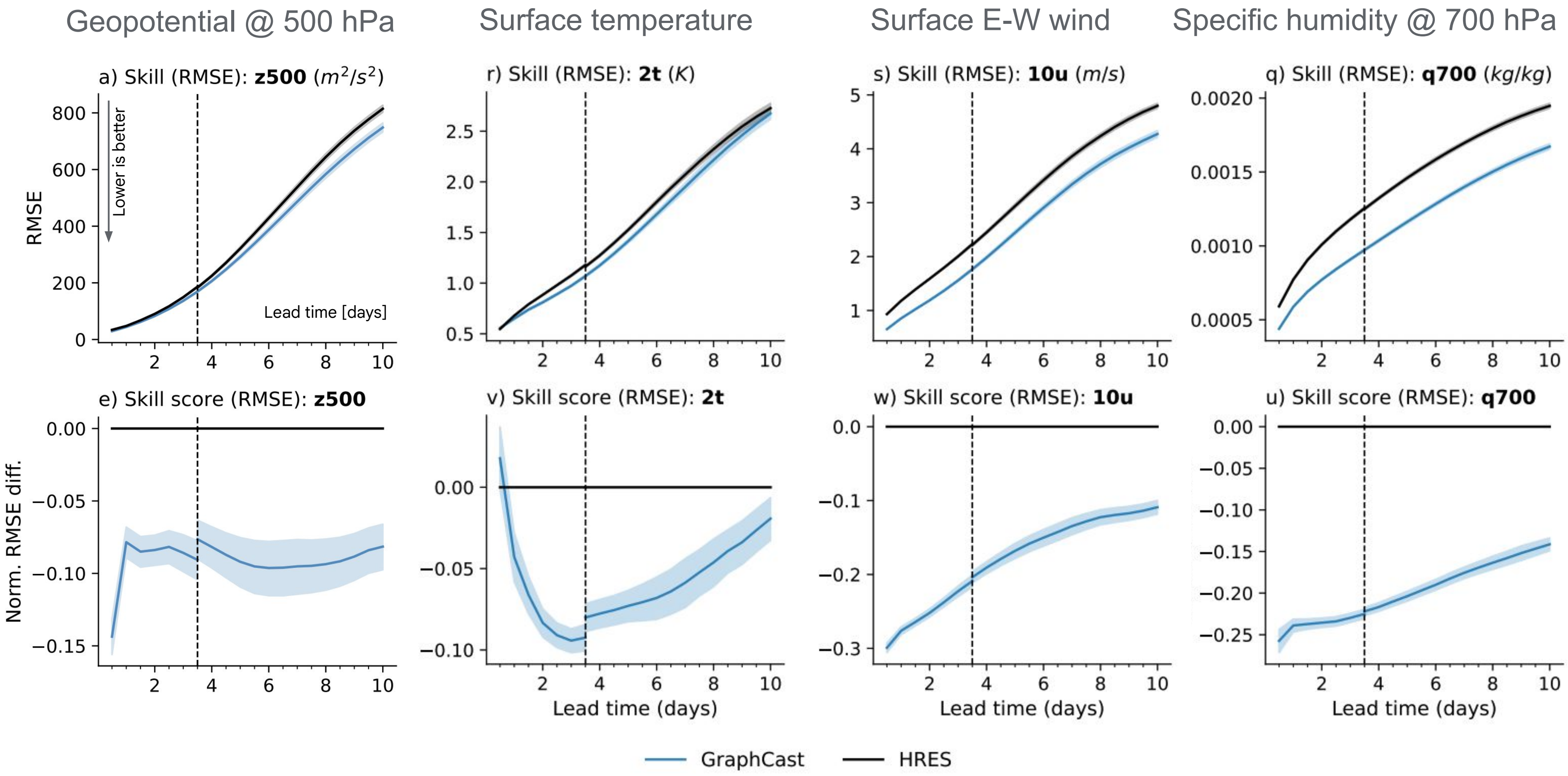
- 1 week



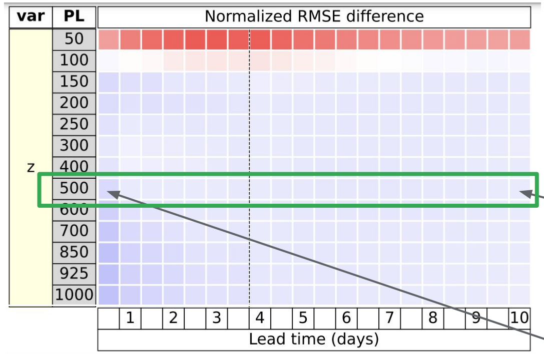
Representative HRES & GraphCast forecasts (median error in 2018)



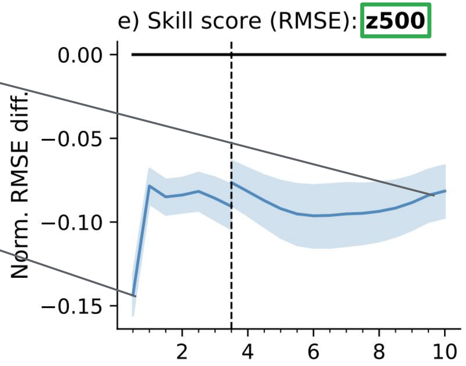
GraphCast outperforms HRES (top operational system)



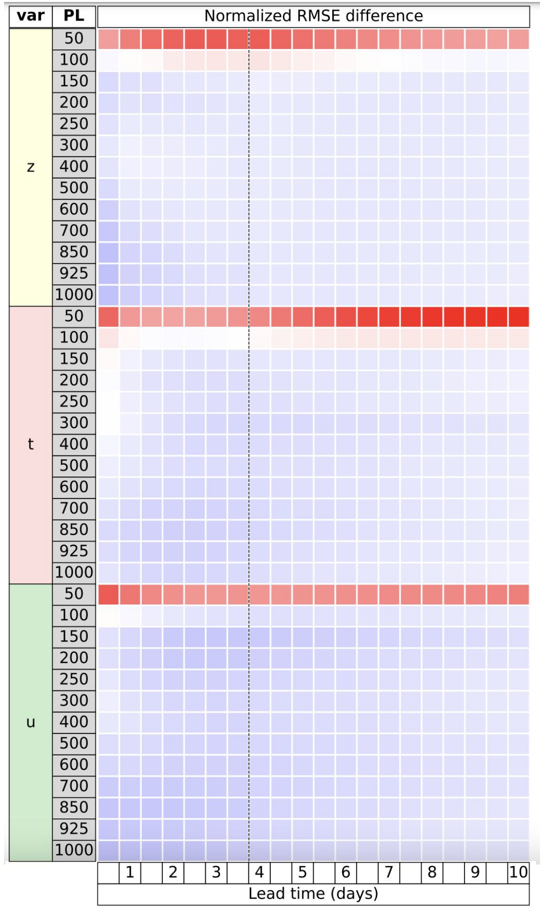
ECMWF-style “scorecard” for comparing GraphCast to HRES



Blue = GraphCast is better, Red = HRES is better.

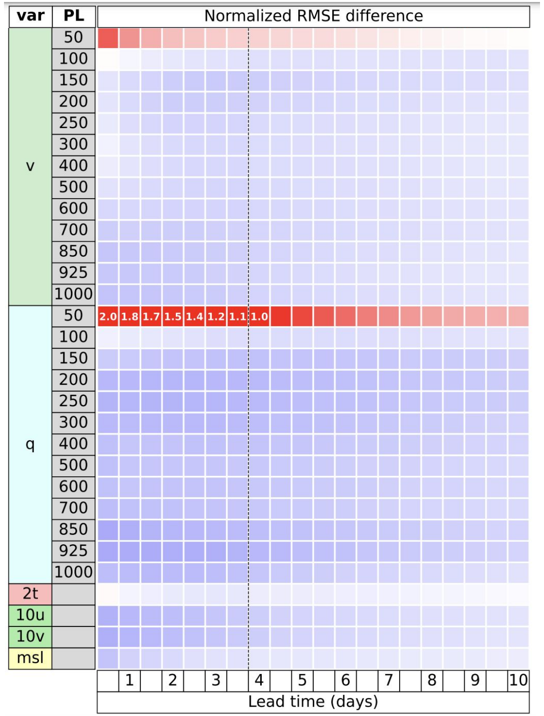


ECMWF-style "scorecard" for comparing GraphCast to HRES



GraphCast has better RMSE on 90.0% on 1380 targets

Blue = GraphCast is better, Red = HRES is better.



Severe weather applications

- Tropical cyclones
- Atmospheric rivers
- Extreme heat

Severe weather: tropical cyclones

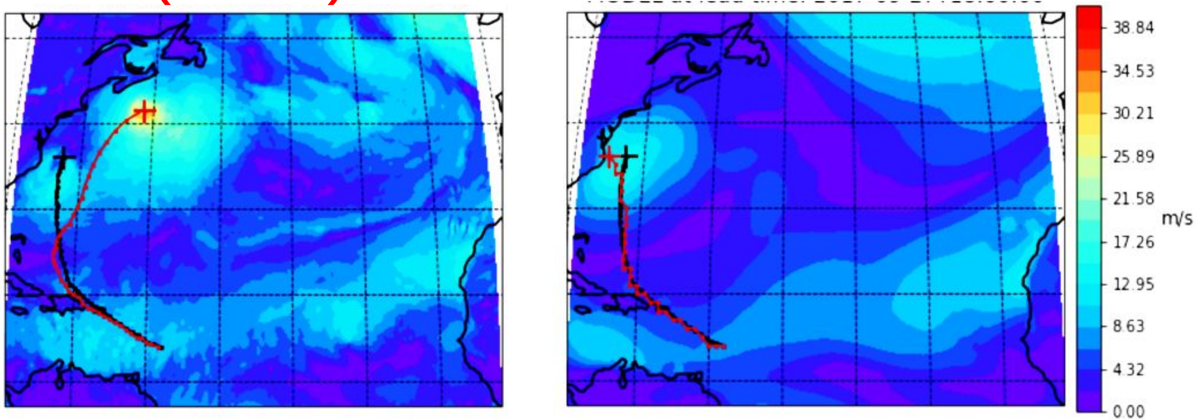
Evaluated cyclone tracks extracted from GraphCast's forecasts, against the IBTrACS dataset.

HRES

(IBTrACS)

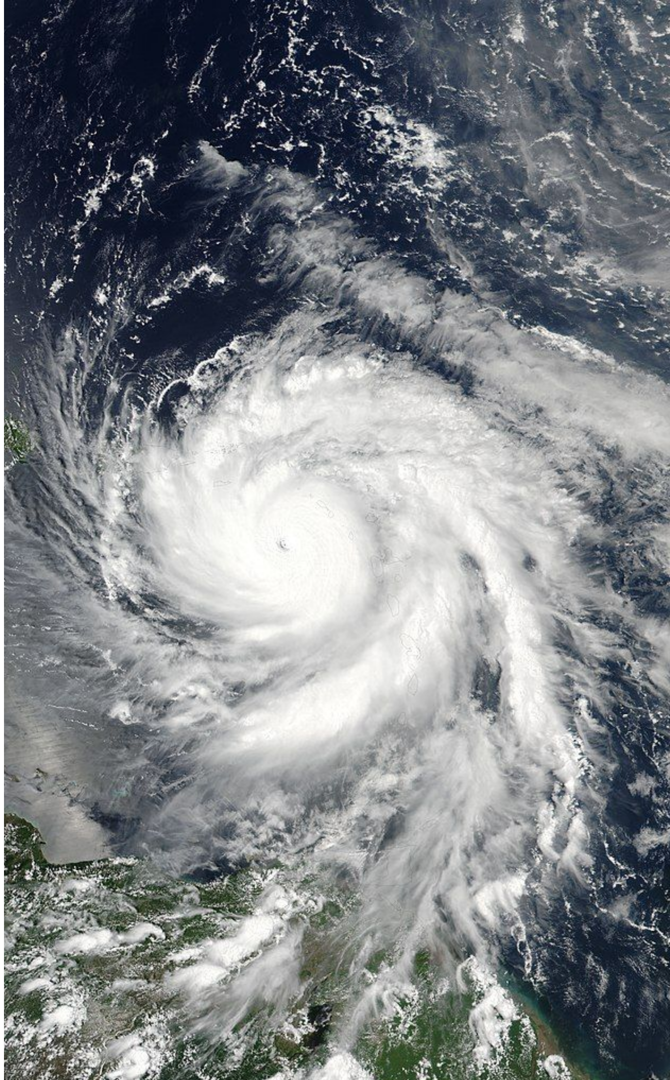
Ground truth

GraphCast



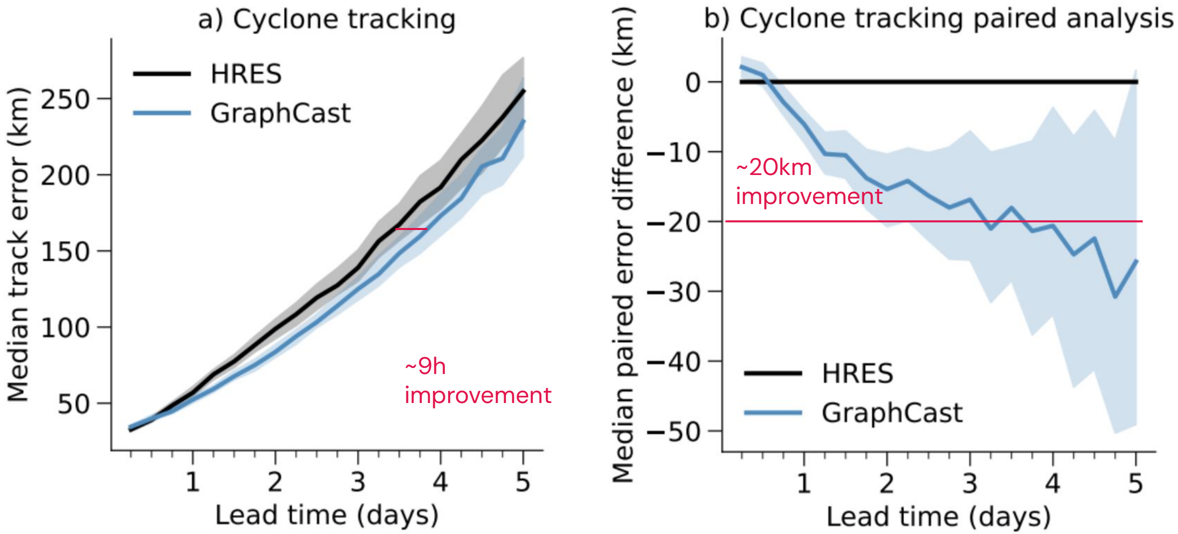
Example: Hurricane Maria (2017):

- Worst storm to ever hit Dominica, Saint Croix, Puerto Rico.
- ~$100B in damage. 3rd costliest storm on record.

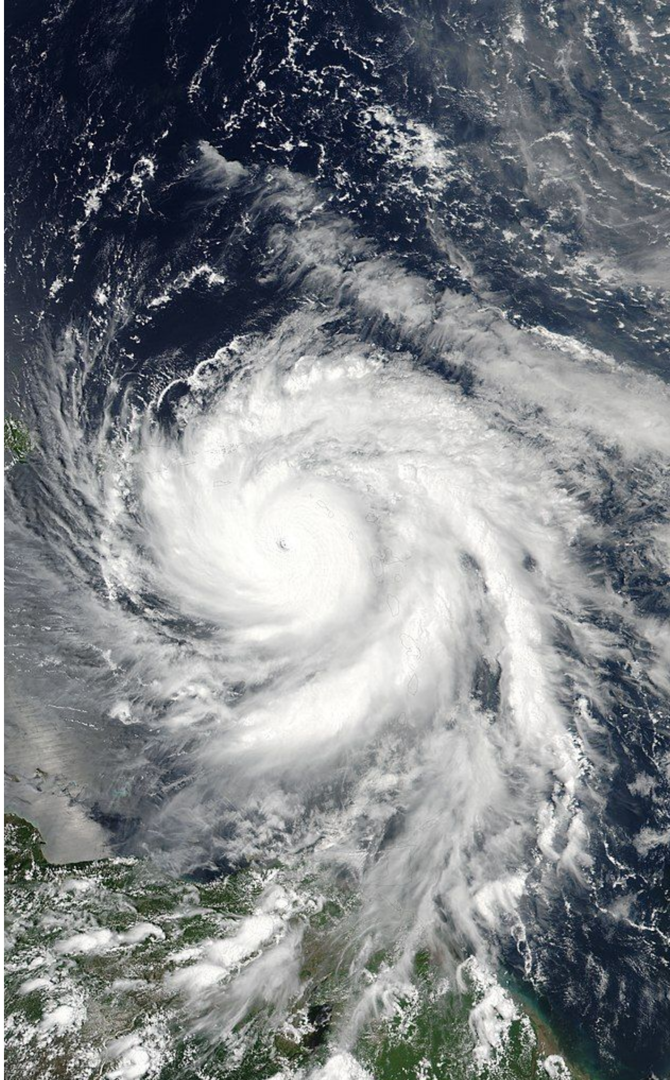


Severe weather: tropical cyclones

Evaluated cyclone tracks extracted from GraphCast's forecasts, against the IBTrACS dataset.



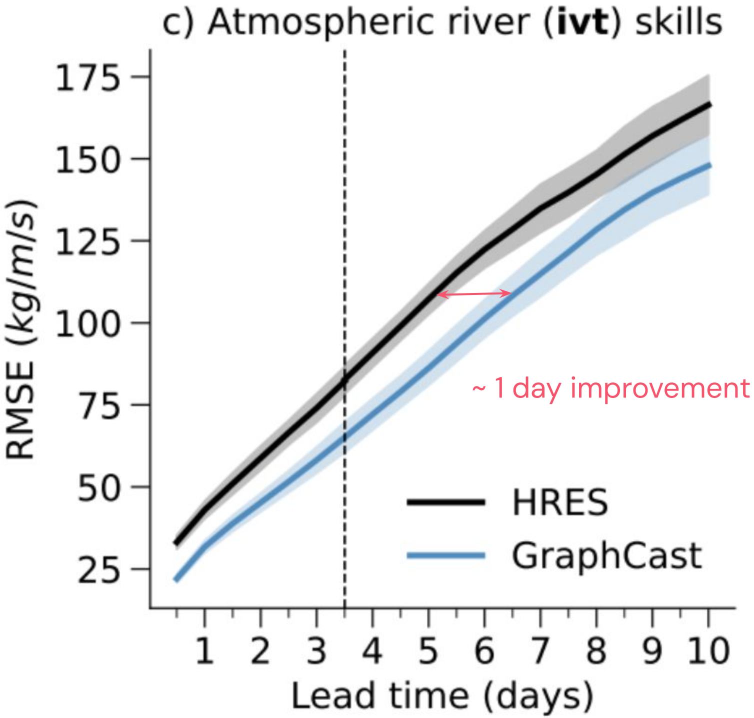
GraphCast gains ~9 hours in accuracy over HRES' published tracks



Severe weather: atmospheric rivers

‘Rivers in the sky’ which transport water vapor away from the tropics, delivering heavy rain.
Strength is characterized by **Integrated Vapour Transport** (IVT).

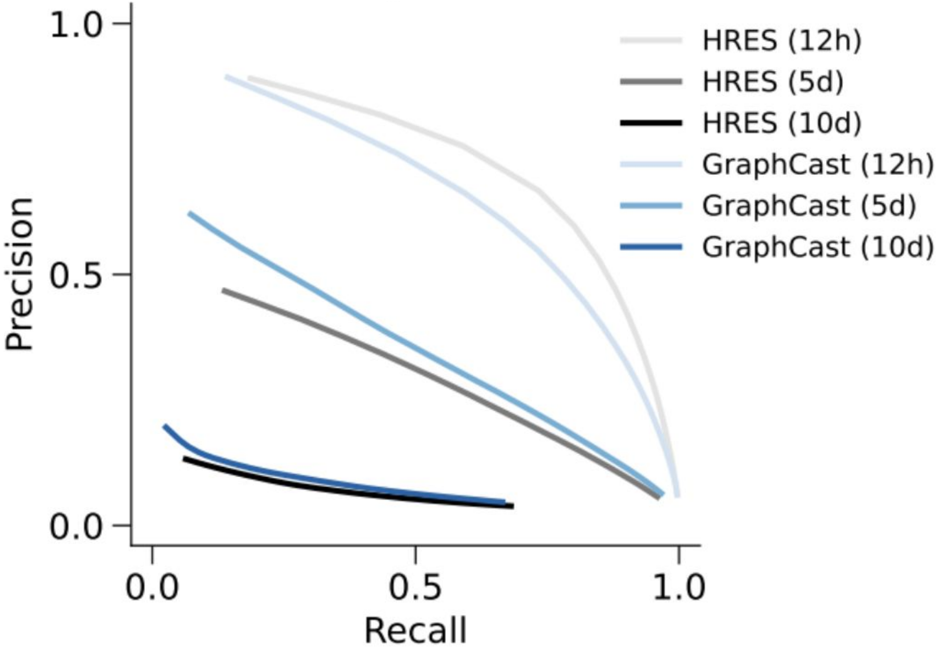
Credit: Mark Ross, Scientific American



GraphCast gains ~1 day of accuracy over HRES when forecasting IVT.

Severe weather: extreme heat

Classify top 2% extremes for surface temperature over land in summer



Predict when **surface temperature** will reach **top 2% extremes** over land in summer.

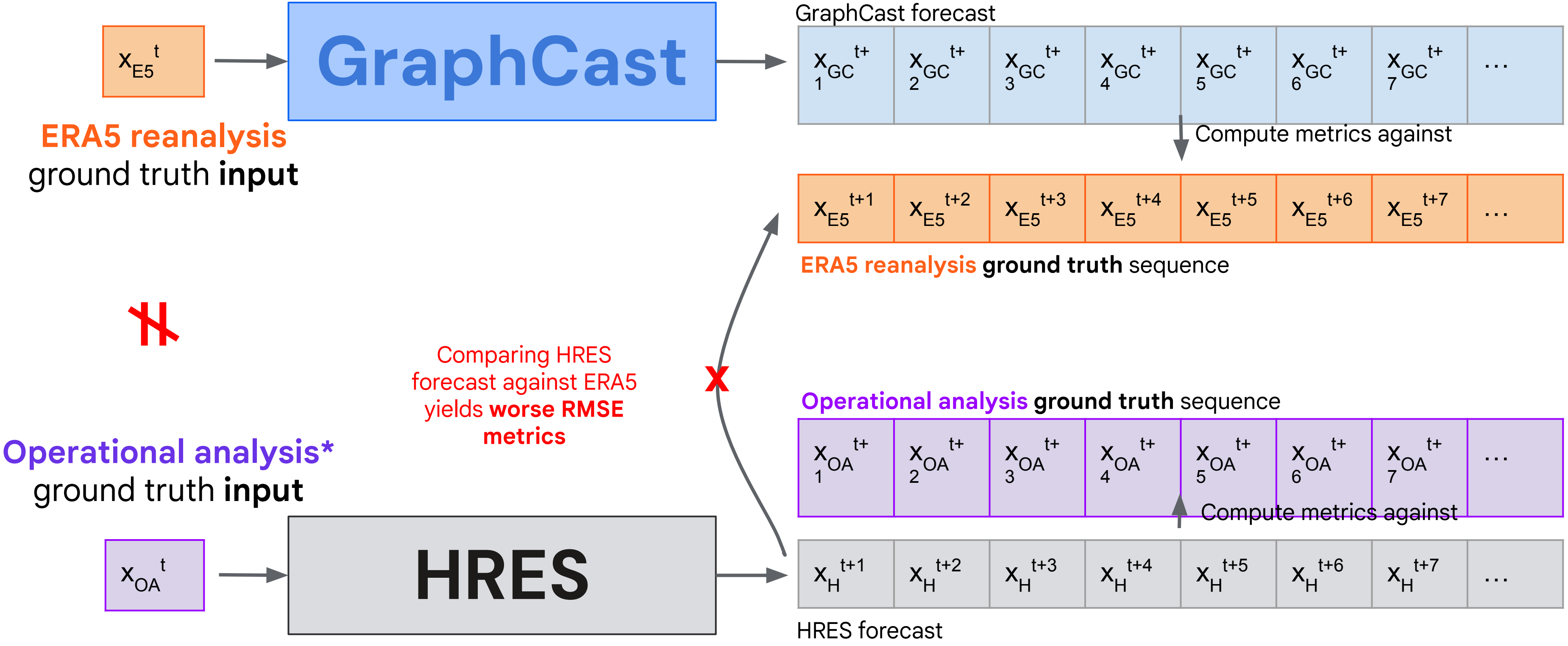
- GraphCast dominates at long lead times.
- HRES still dominates at very short lead times (12h).

Other variables also related to extreme heat (**t850, t500, z500**)

- GraphCast dominates at most lead times.

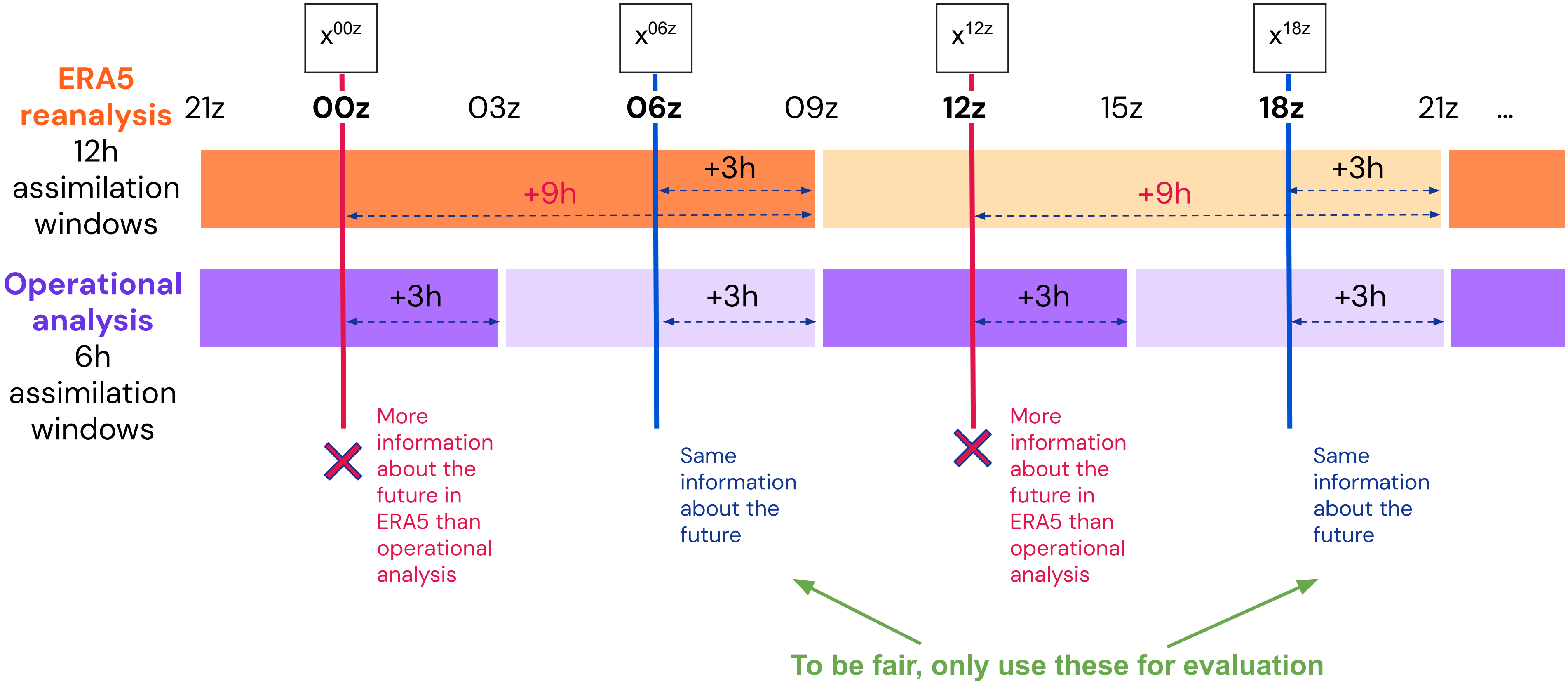
A fair comparison to HRES

A fair comparison to HRES: What to use as **ground truth**?



*Actually ECMWF prodives multiple operational analysis product. The one we used is the most favourable to HRES metrics, **which is not “HRES Analysis”**

A fair comparison to HRES: Assimilation window lookahead



A fair comparison to HRES: Assimilation window lookahead

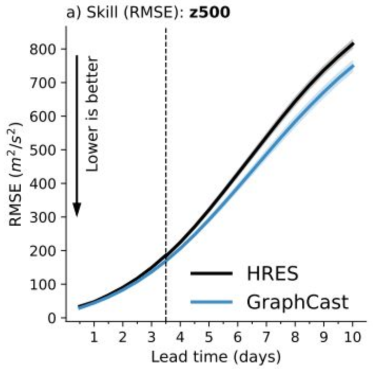
Problem: *ERA5 forecasts from 00z/12z initializations have an +9h look ahead*

Solution: only evaluate on initializations from 06z/18z*

Problem: *Easier to predict targets within the same assimilation window*

Problem: *Targets with 9h assimilation into the future are harder to predict*

Solution: **evaluate only at multiples of 12h lead time** always crossing the assimilation window, and always on data with +3h look ahead

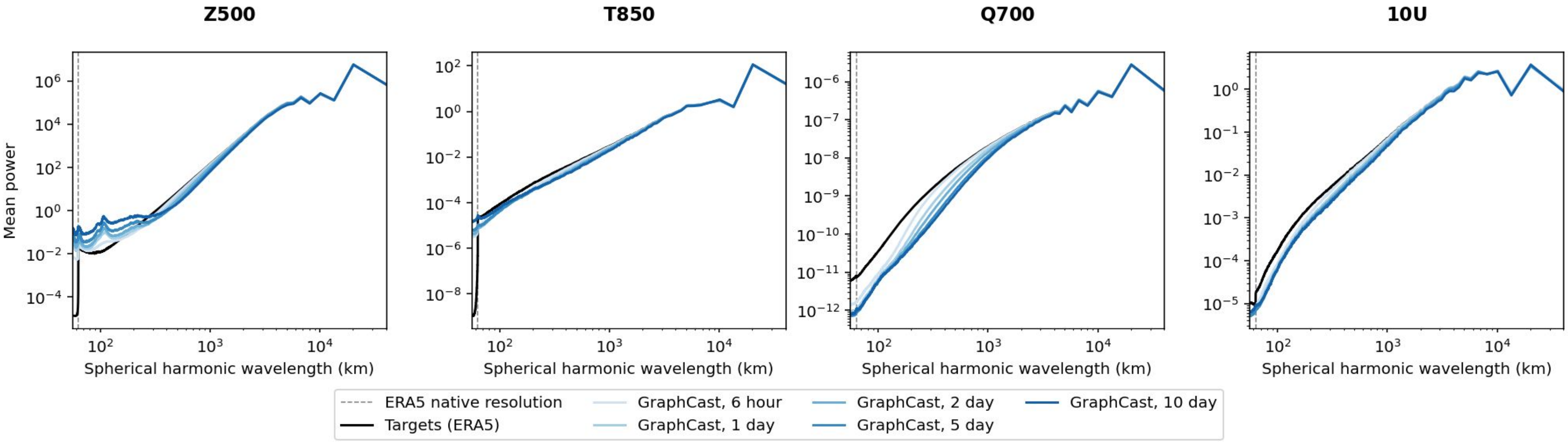


* Caveat: HRES 06z/18z initializations are available only up to 3.75 days lead times. For 4-10 days lead times we compare GraphCast 06z/18z inits with HRES 00z/12z inits.

A fair comparison to HRES: Is GraphCast blurring a lot?

The RMSE metric rewards models for averaging over uncertainty by **blurring**.

- ML models trained **to minimize RMSE will learn to blur,** which may reduce their RMSE significantly



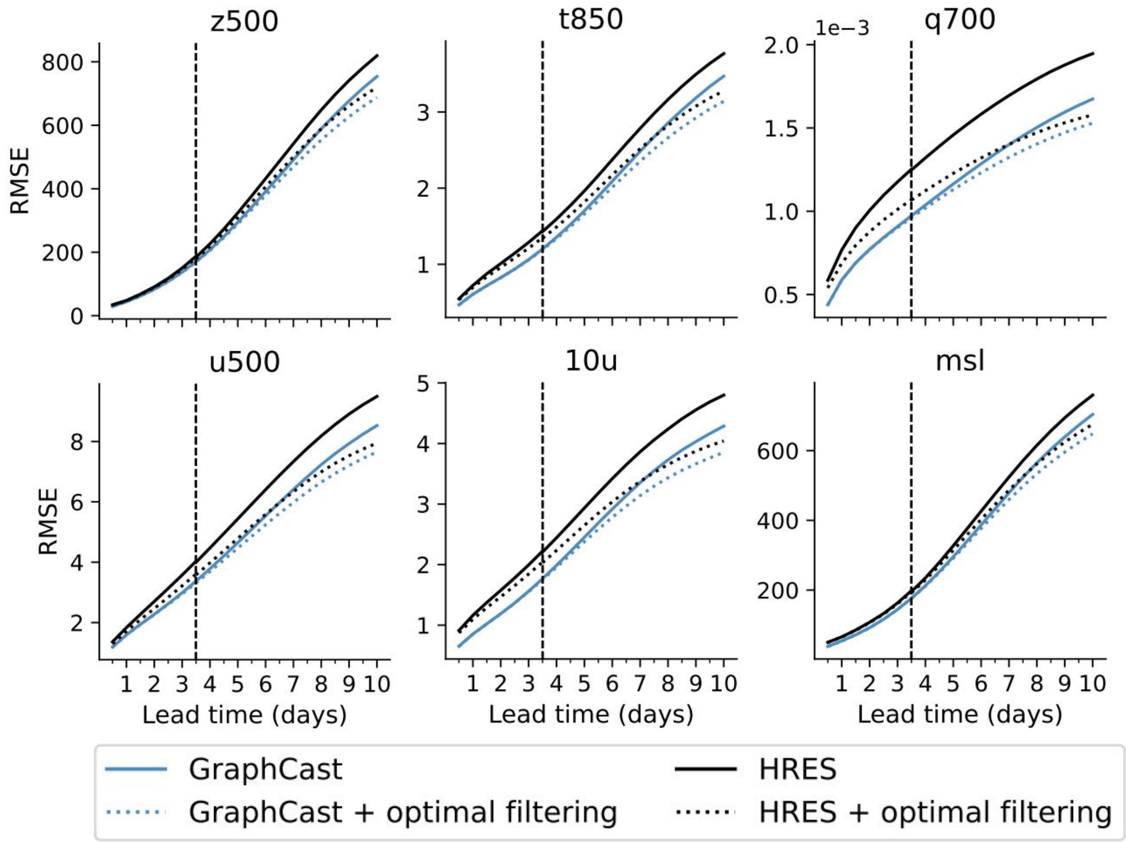
Is GraphCast improving over HRES on RMSE metrics simply because HRES can't blur?

A fair comparison to HRES: Optimal filtering to control for blurring

Optimal filtering (for each model):

- Fit isotropic spectral filter that minimizes RMSE for that model

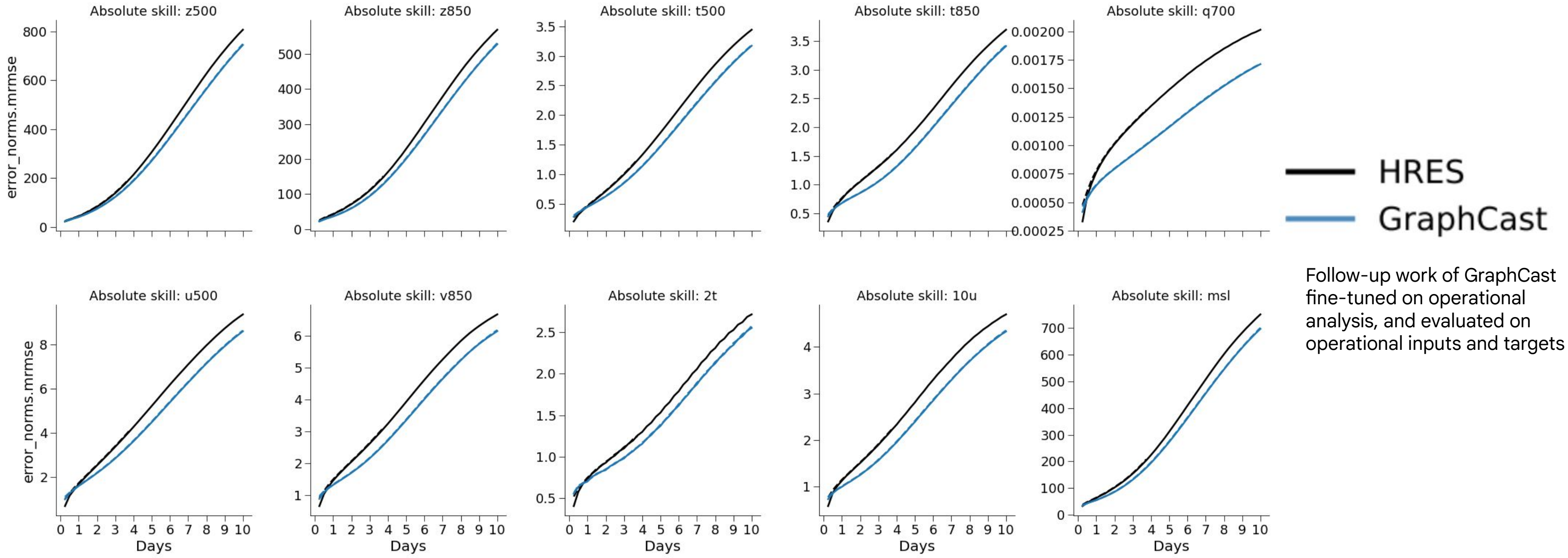
After both models' predictions have been filtered in this way, **GraphCast still outperforms HRES** on RMSE.



A fair comparison to between HRES and ML models is subtle

Potential solutions:

- **Always use operational analysis**

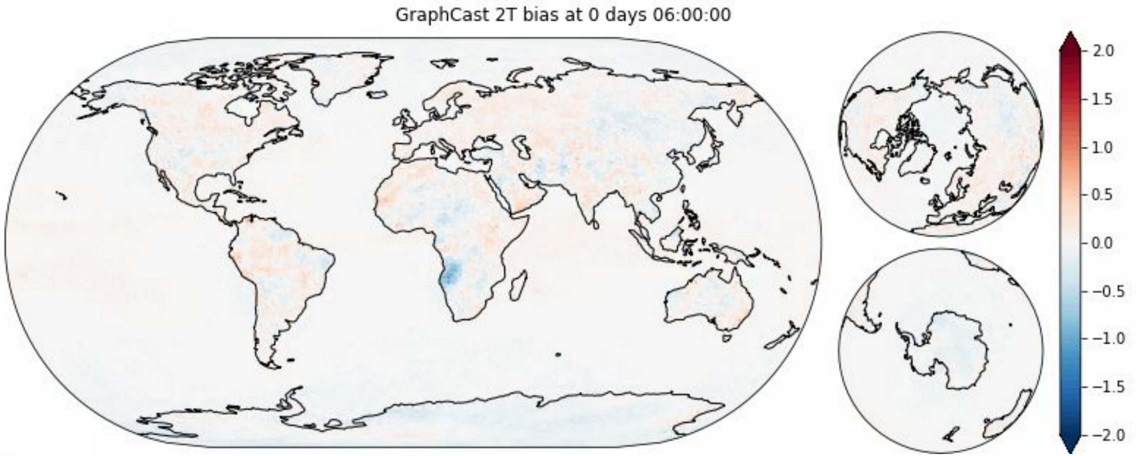


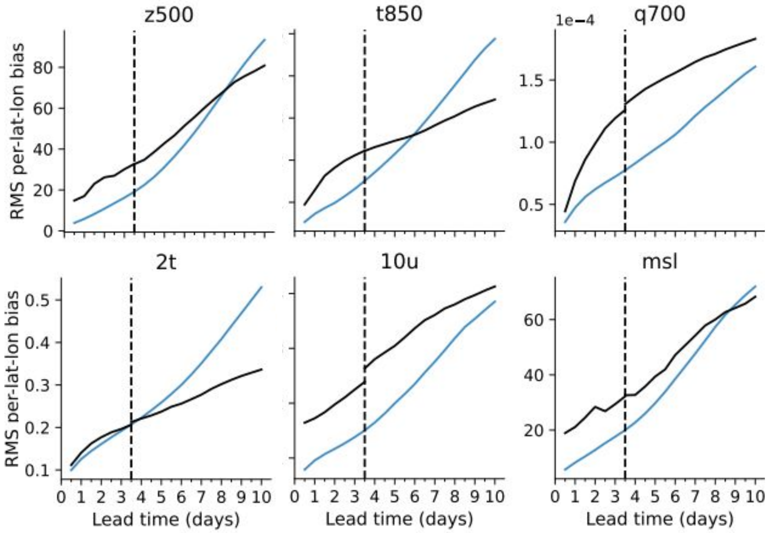
Follow-up work of GraphCast fine-tuned on operational analysis, and evaluated on operational inputs and targets

- **Improve next ERA5** with resolution and assimilation windows closer to operational setting?
- **Evaluation against observations** (good benchmarks dataset needed)

Advanced analyses and model variations

Advanced analysis: Geographic biases



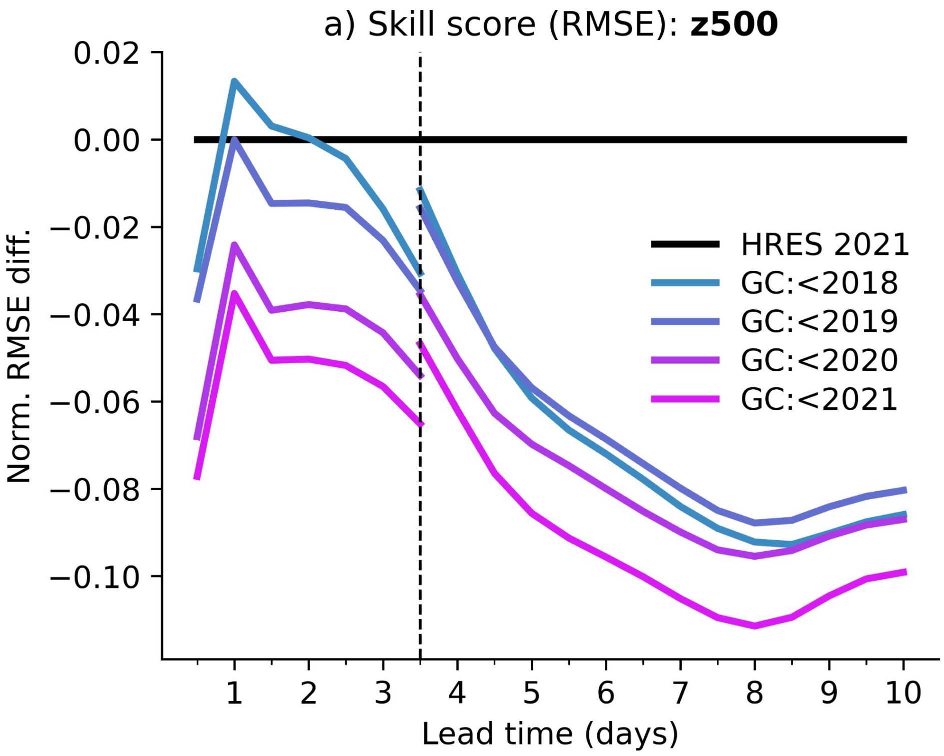


HRES vs **GraphCast**

Average magnitude of geographic biases:

- Similar in magnitude
- Both grow with lead time
- Some correlation (R=0.4-0.6) at long lead times

Model variations: Future years and training on recent data



Main results:

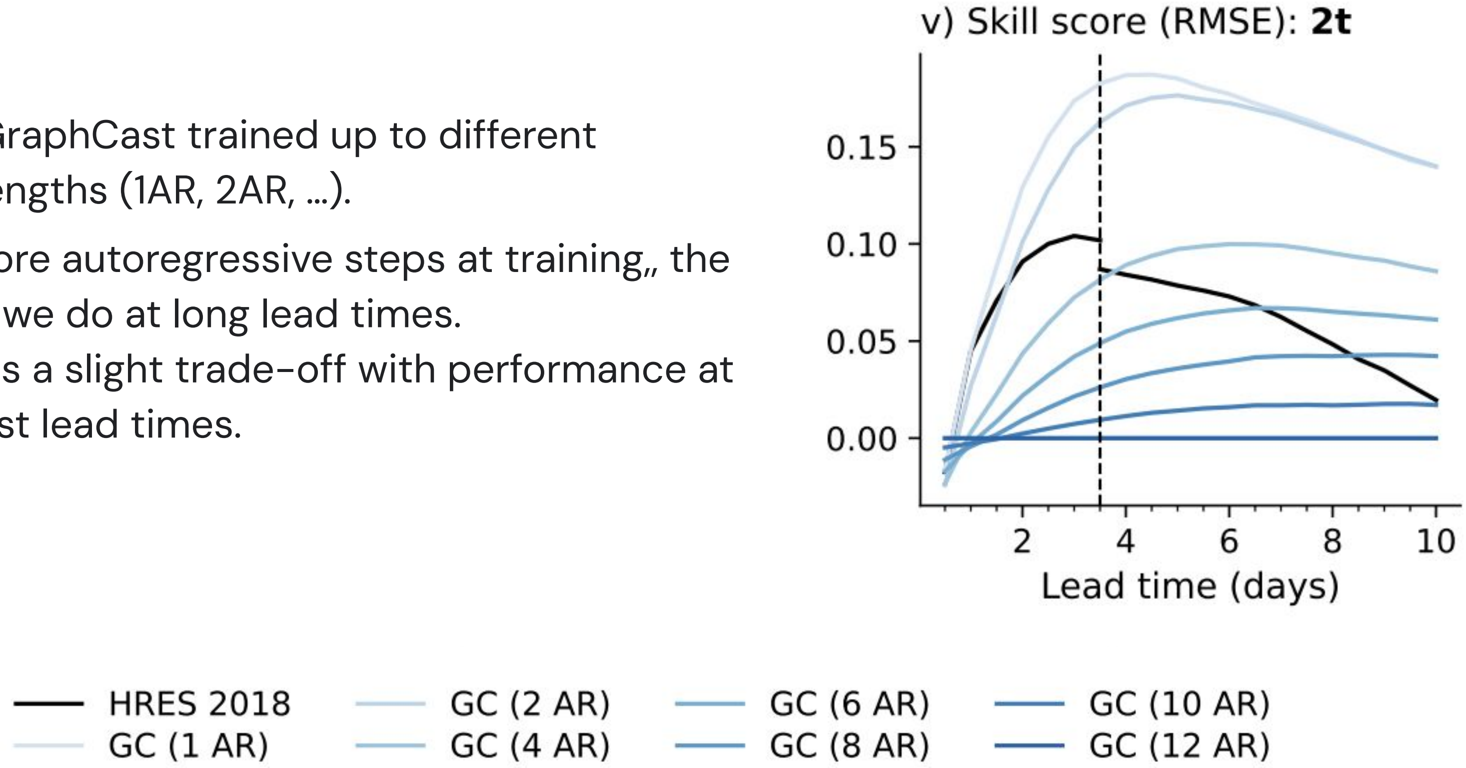
- Train on data <2018, eval on 2018
- **2021 eval**?

The closer to 2021 that we train up to, the better we do on 2021.

Model variations: Autoregressive training

RMSEs for GraphCast trained up to different sequence lengths (1AR, 2AR, ...).

- The more autoregressive steps at training,, the better we do at long lead times.
- There is a slight trade-off with performance at shortest lead times.



Conclusions

- **GraphCast outperforms the best existing operational model** in many ways

  Better in most **scorecard metrics**

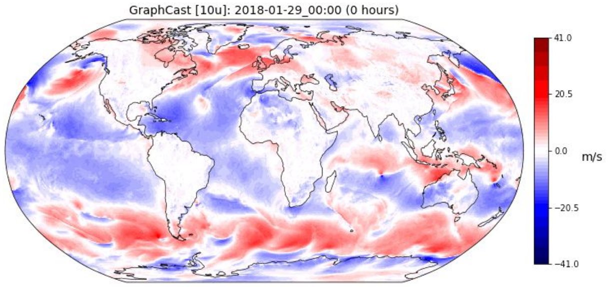
  Useful for real world applications (e.g. cyclone tracking)

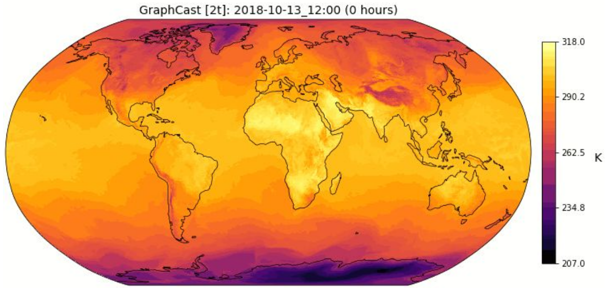
  Faster inference

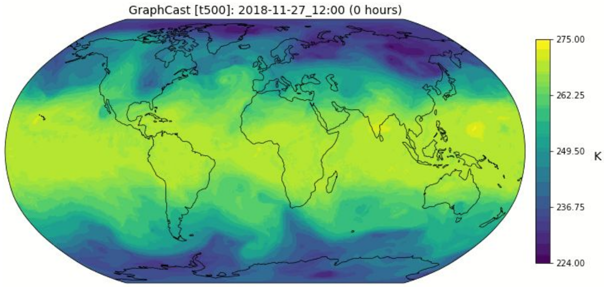
  Comparisons should **not be summarized to just RMSE**

- **Deep-learning** weather models are **here to stay**

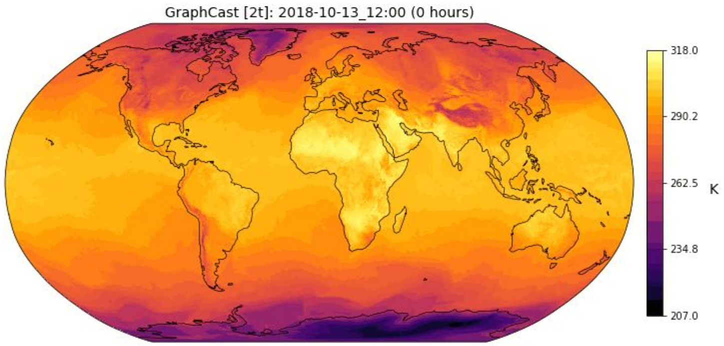
  Probably also for **any Earth scale modelling** with abundant data







Thank you!
Any questions?



GraphCast: Learning skillful medium-range global weather forecasting

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, Peter Battaglia

arXiv https://arxiv.org/abs/2212.12794

https://github.com/deepmind/graphcast

Google DeepMind