

A review of the recent papers on fully data-driven NWP with AI

Thomas Rieutord

Postdoctoral researcher on AI at Met Éireann

ACCORD all-staff meeting – March 2023

Thanks to Clément Brochet, Matthew Chantry, Arnaud Mounier, Laure Raynaud
and many Met Éireann colleagues

Outline

Introduction

Individual presentations

Comparative presentation

Conclusion

Fully data-driven NWP

Numerical weather prediction based only on data, without any physical model. With the availability of large datasets (e.g. ERA5), specialized computing resources (e.g. GPUs) and new algorithms (e.g. transformers), such methods made great progress

- ▶ Over the last year, several papers presented fully data-driven NWP with AI
- ▶ Although they are limited to the resolution of ERA5 (0.25°), they show impressive performances, especially in saving computational costs
- ▶ In terms of forecast precision, they compare with the ECMWF model at medium-range, according to a limited set of metrics and variables
- ▶ These papers are not from meteorologists but from major tech companies
- ▶ The pace of publication is very high

Methodology and limitations for this review

Methodology: Compare and explain what has been done in each paper in terms of

- ▶ AI technique
- ▶ Hardware
- ▶ Computing power
- ▶ Forecast skill

Limitations

- ▶ Pre-prints: these papers have not been peer-reviewed
- ▶ My expertise is limited: meteorologist background with math specialization

Big-Tech papers

- ▶ **Pathak and 12 co-authors (22 Feb. 2022)**, Nvidia. **FourCastNet**: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.
- ▶ **Bi and 5 co-authors (3 Nov. 2022)**, Huawei. **Pangu-Weather**: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast.
- ▶ **Lam and 17 co-authors (24 Dec. 2022)**, DeepMind. **GraphCast**: Learning skillful medium-range global weather forecasting

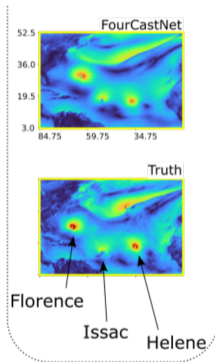
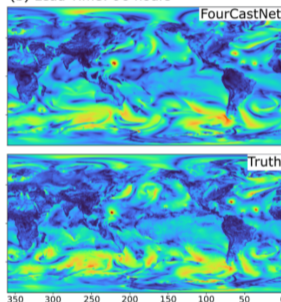
Not included in the review but worth mentioning:

- ▶ **Keisler (15 Feb. 2022)**, personal work. Forecasting Global Weather with Graph Neural Networks
- ▶ **Nguyen and 4 co-authors (24 Jan. 2023)**, Microsoft. **ClimaX**: A foundation model for weather and climate

Pathak et al. (22 Feb. 2022), Nvidia. **FourCastNet**: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.

- ▶ First to produce forecast at NWP resolution 0.25° (previous were $\geq 1^\circ$), quantitatively evaluate extreme, use transformers.
- ▶ Adaptive Fourier Neural Operator (AFNO)
- ▶ "The FourCastNet model can compute a 100-member 24-hour forecast in 7 seconds" using four A100 GPUs

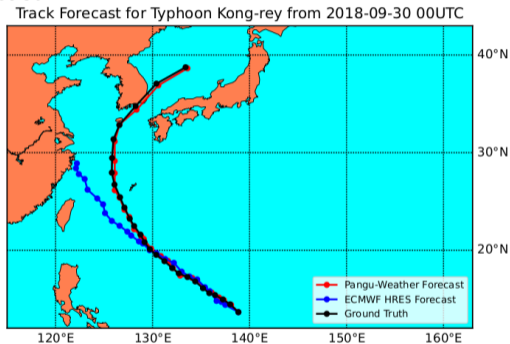
(b) Lead Time: 96 hours



Example of forecast: 10m wind at 96h lead time

Bi et al. (3 Nov. 2022), Huawei. **Pangu-Weather**: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast.

- ▶ First to outperform IFS (RMSE and ACC at 0.25°)
- ▶ 3D Earth-specific transformer (3DEST)
- ▶ Hierarchical temporal aggregation: train models for 1h, 3h, 6h, 24h time steps.

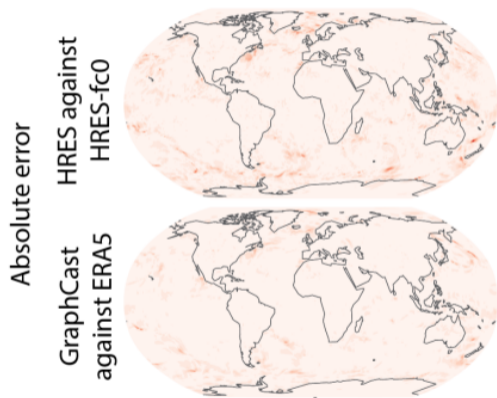


Example of cyclone forecast: track of typhoon Kong-rey

GraphCast



Lam et al. (24 Dec. 2022), DeepMind. **GraphCast**: Learning skillful medium-range global weather forecasting

- ▶ First to produce a score card, outperform Pangu-Weather
- ▶ Main technical innovation: graph neural network (GNN)
- ▶ Auto-regression from the two last state $\hat{X}_{t+1} = f(X_t, X_{t-1})$
- ▶ Predict much variables than the previous model, thus enable more comprehensive assessment



Forecast error at 48h lead time for u_{10} . IFS (HRES) evaluated against its own analysis

Comparative presentation

	FourCastNet	Pangu-Weather	GraphCast
AI technique	AFNO (trans-former)	3DEST (trans-former)	Graph neural network
Hardware – train (inference)	64 A100 (1 A100)	192 V100 (1 V100)	32 TPU v4 (1 TPU v4)
Speed – train (inference¹)	16 hours (2.8 s)	16 days (14 s)	3 weeks (60 s)
Forecast scores²	Comparable to IFS	Better than IFS	Better than IFS
# of variables	20	69	227
Open-source	Yes 	Yes 	No

- ▶ Much faster than conventional NWP
- Common points:
- ▶ Trained on ERA5 (0.25^o resolution)
 - ▶ Do not provide ensemble scores

¹for a deterministic 10-day forecast with the hardware for inference

²only RMSE and ACC

AI technique

	FourCastNet	Pangu-Weather	GraphCast
AI technique	AFNO (trans-former)	3DEST (trans-former)	Graph neural network

Transformers

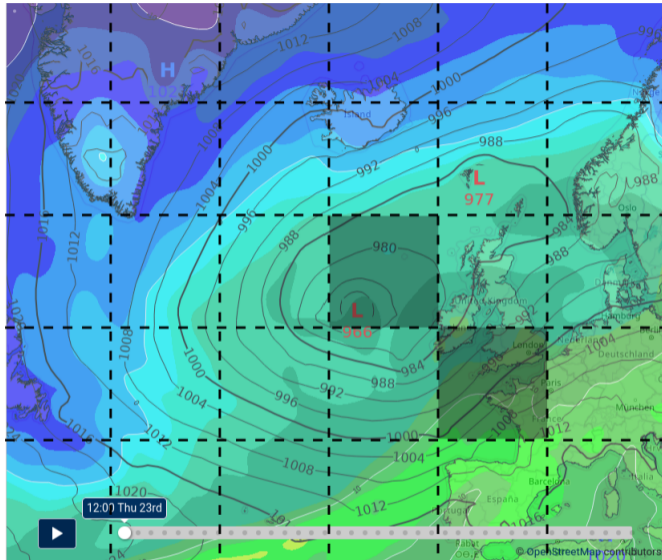
- ▶ Initially designed for natural language processing. Now widely used (DallE, ChatGPT...)
- ▶ Solution to connect words regardless of distance between them
- ▶ The core of transformers is the attention layer:

$$Attention(Q, K, V) = softmax\left(\frac{Q \cdot K}{\sqrt{d}}\right) V$$

Graph network

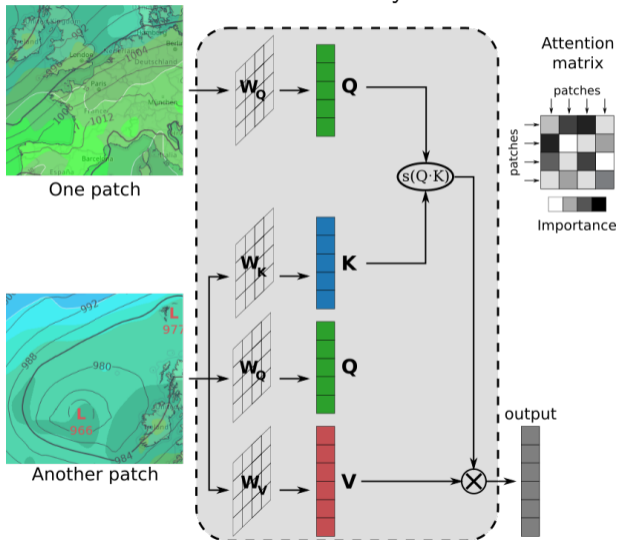
- ▶ Invented by Peter Battaglia in 2018
- ▶ Neurons connections are represented by a graph (much more flexible than convolution or recurrent networks)
- ▶ Can be seen a generalization of transformers

Focus on transformers



Focus on transformers

Illustration of a self-attention layer:



- ▶ Learnable parameters: W_K , W_Q and W_V .
- ▶ In Pangu-Weather, the attention layer is modified to account for position-related bias B (learnable):

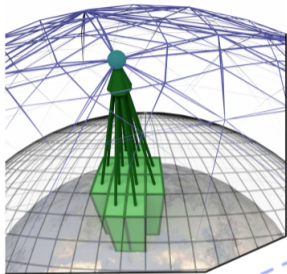
$$\text{softmax} \left(\frac{Q \cdot K}{\sqrt{d}} + B \right) V$$

- ▶ In FourCastNet, the attention is re-written as a convolution, that is then computed in Fourier space:

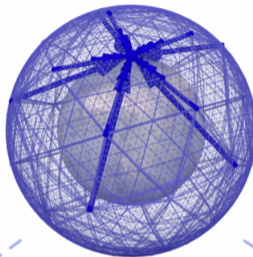
$$\mathcal{F}^{-1}(\mathcal{F}(\kappa)\mathcal{F}(X))$$

Focus on graph networks

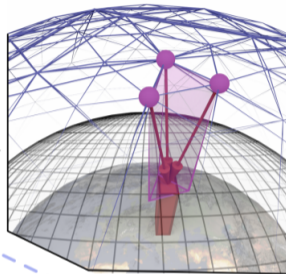
d) Encoder



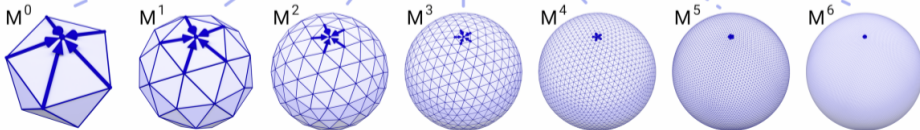
e) Processor



f) Decoder



g) Simultaneous multi-mesh message-passing



Screenshot of Fig. 1 in the GraphCast paper

Hardware

	FourCastNet	Pangu-Weather	GraphCast
Hardware – train (inference)	64 A100 (1 A100)	192 V100 (1 V100)	32 TPU v4 (1 TPU v4)

Different types of GPUs:

- ▶ Google TPU v4: 275 TFLOPS, 32GB
- ▶ Nvidia A100: 312 TFLOPS, 40/80GB
- ▶ Nvidia V100: 112 TFLOPS, 16/32GB
- ▶ Nvidia 1080Ti: 10 TFLOPS^a, 11GB

^anot an official spec, figure from my own calculations

Example of existing infra:

- ▶ Met Eireann: 1 A100 80GB
- ▶ Meteo-France: 13 V100 32GB, 2 1080Ti
- ▶ ECMWF: 72 A100 40GB
- ▶ European Weather Cloud (soon): 36 A100 80GB

Computing time

	FourCastNet	Pangu-Weather	GraphCast
Speed – train (inference)	16 hours (2.8 s)	16 days (14 s)	3 weeks (60 s)
Speedup from IFS³	44727	24919	2368

Inference figures for a deterministic 10-day forecast

Note that these figures are only for GPUs. Expect ~100 times slower on CPUs

Warning: tricky comparison

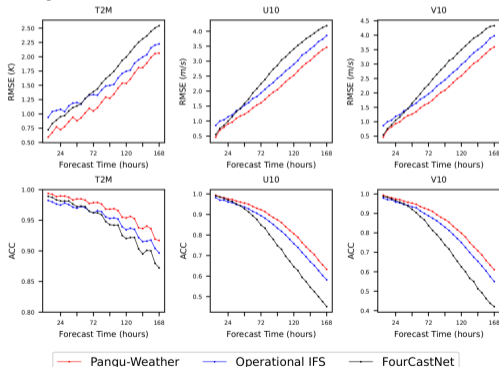
- ▶ Heterogeneous hardware. Not the same number of variables. Not the same output frequency. Lower resolution than IFS.
- ▶ Although, figures are several orders of magnitude above conventional NWP, which is significant

³obtained from FourCastNet figures proportioned by inference time and hardware speed. Ex:
 $44727 \frac{2.8}{14} \frac{312}{112} = 24919.$

Forecast scores

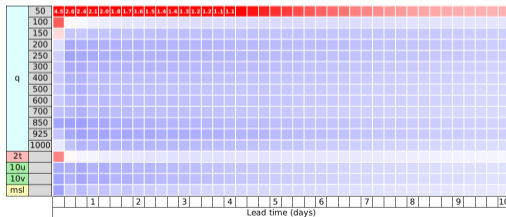
	FourCastNet	Pangu-Weather	GraphCast
Forecast scores	Comparable to IFS	Better than IFS	Better than IFS

Only evaluated with RMSE and ACC. No probabilistic score.



← Pangu-Weather, FourCastNet and IFS scores with ERA5 as reference

↓ GraphCast score card against IFS with ERA5 as reference (blue is better)



Conclusions

- ▶ Results are very impressive, beyond what was thought as realistic 3 years from now
- ▶ Weather and climate appear to be the new competition field of big tech companies
- ▶ Resolution of AI-generated forecast is not yet at the level of operational models. Evaluation still has some flaws (set of variables, probabilistic scores, real-life feedback...)

Threats

- ▶ Obsolescence of current NWP
- ▶ Devaluation of physics knowledge
- ▶ Unreachable levels of hardware and AI skills (except for European scale)

Opportunities

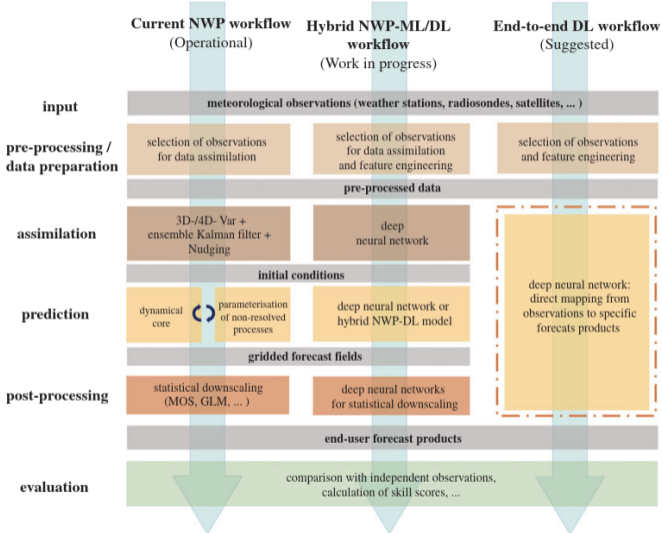
- ▶ Improve performance where the physics is poorly known
- ▶ Improve computing and energy efficiency
- ▶ Take advantage of hardware heterogeneity

Closing thoughts

- ▶ How would this change the way we work in national met services?
- ▶ **If** physics knowledge were no longer required to run NWP, what would it be useful for?
 - ▶ Auditing AI forecast (to ensure security of people and goods)?
 - ▶ Creating training set for the next AI updates?
- ▶ What changes does it imply in producing forecast and meteorological information?
 - ▶ Get the hardware and the knowledge to run these AI?
 - ▶ Stronger focus on post-processing and impact-based forecast?
- ▶ Regional scale: no competing AI at the moment.
 - ▶ Provide open regional reanalysis?
 - ▶ Investigate ourselves more data-driven NWP?
 - ▶ Additional challenges at higher resolution?

Thank you for your attention!

Appendix: multiple levels of AI integration



The weather prediction workflow as described in Schultz et al. (2021) [↗](#)