Lessons Learned From Building And Operationalizing ML Weather Prediction Models

(and a few opinions too)

vivek@excarta.io

Feb 1, 2024

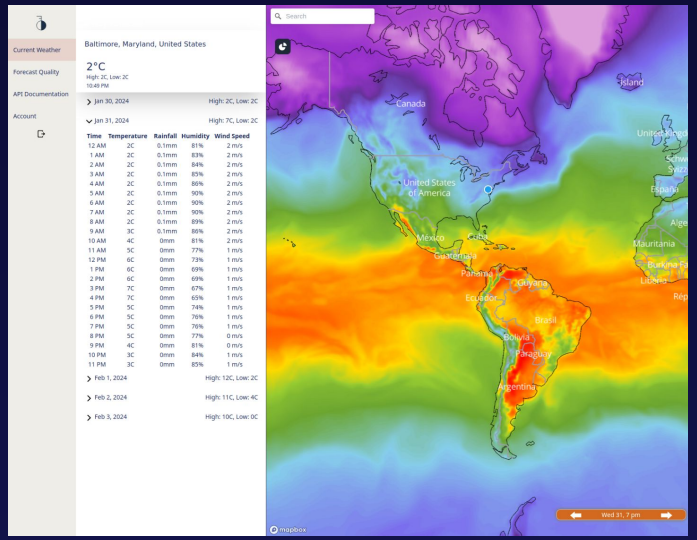
A bit about us

Excarta: building+operationalizing AI models for better forecasts + intelligence

- End users consume weather for operational decision making, e.g., wind and solar power prediction, load forecasting.

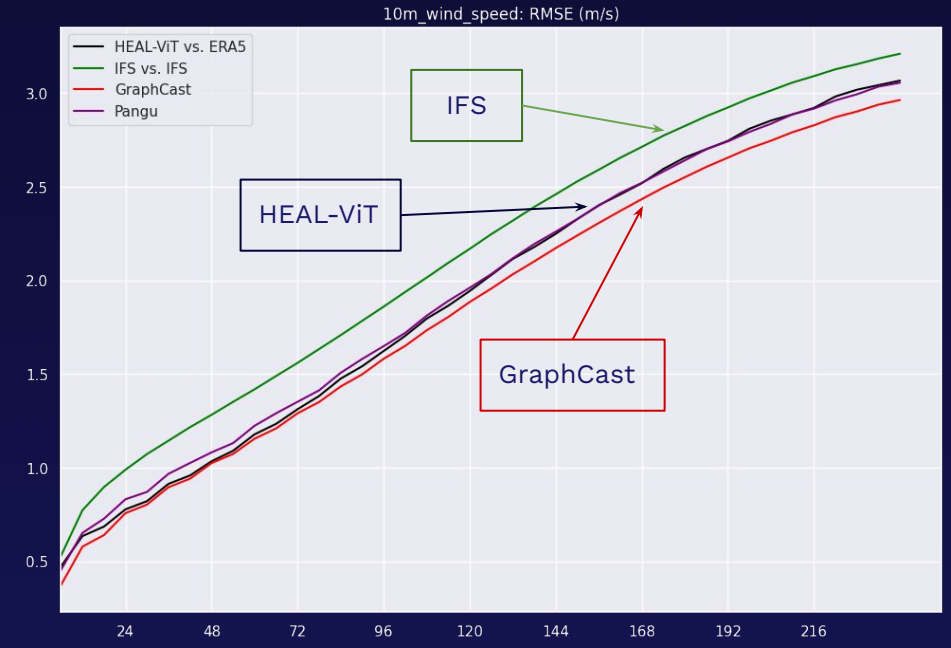
Operational model:

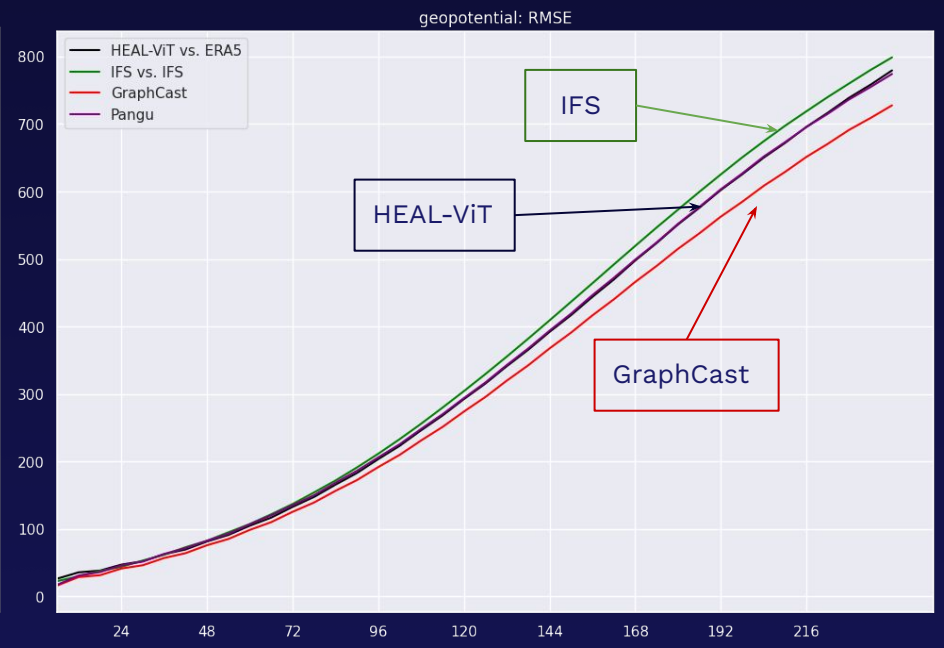
- 14-day forecasts, hourly step
- Initialized with ECMWF analyses, 4x a day
- Forecast “core” variables (t2m, sp, msl, u10, v10, ...)
- And other variables necessary for end users
  - d2m, u100, v100, SSRD, FDIR, tcc ...
- ~4 minutes from prediction start to being available in the API
- Daily rolling evaluations against observations



https://app.excarta.io

Our MLWP (“HEAL-ViT”) improves upon IFS for key variables.





Some quick observations

- ML weather models do seem to work in the “real world”
  - End users see tangible benefits from MLWPs, even in conjunction with NWPs
  - Skepticism still remains (but reducing…)
  - Time to start seriously thinking about how end-users can use MLWPs alongside NWPs

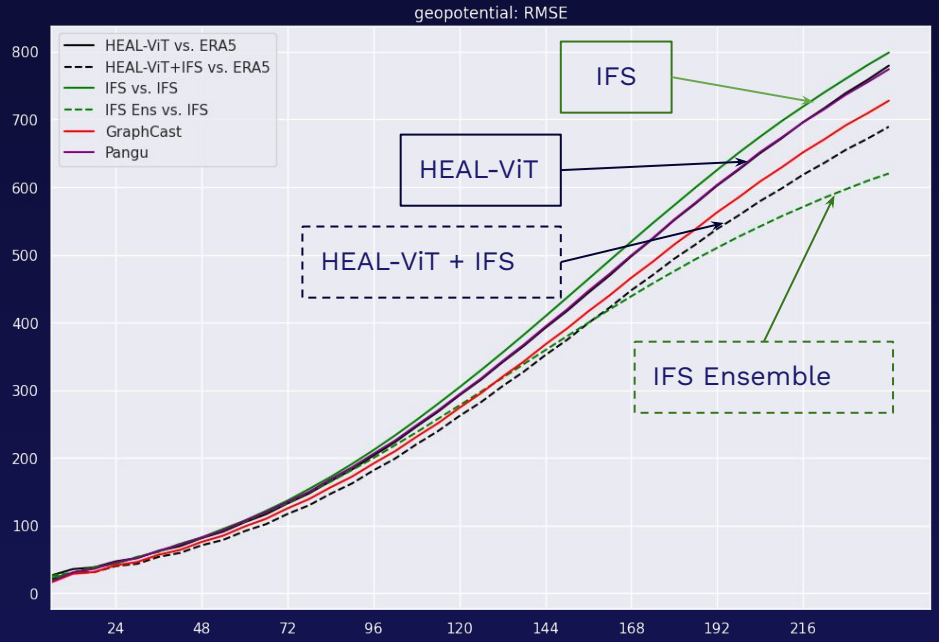
- “Best” approach is not obvious
  - Different MLWPs behave differently from each other, and from NWPs
  - Can also impact operational use
  - Domain experts can offer valuable guidance

- Many ML techniques work “out of the box”, good engineering is critical
  - Different requirements and constraints than NWPs present
  - Helpful to adopt best practices from other domains where ML is deployed
  - Robust train -> test -> deploy -> monitor pipeline necessary to keep up with rapid evolution

MLWPs provide novel information, not just copy NWPs.

Multi-model ensembles significantly benefit from MLWPs

- Naive ensemble of MLWP + IFS shows significant improvement over either one separately



End users: predict solar + wind farm power output using weather as input

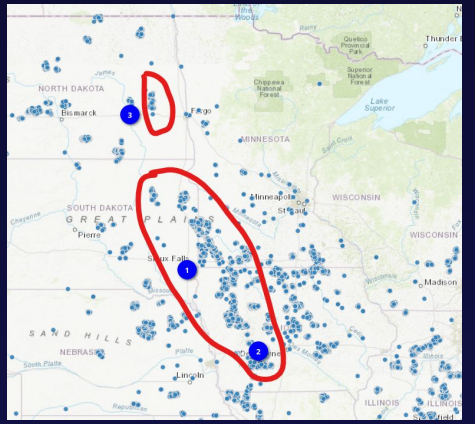
- Multiple NWPs already used as inputs, from global 0.25-deg NWPs to regional 3km models
- MLWP forecasts still provide significant error reduction when predicting power output

MLWPs complement even regional, high-res NWPs.

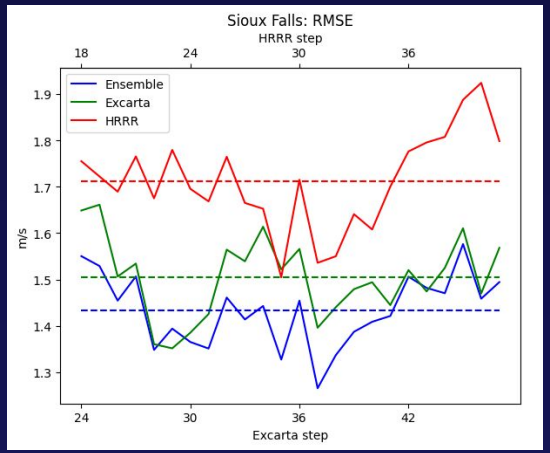
HEAL-ViT vs. HRRR: Day-ahead wind forecasts

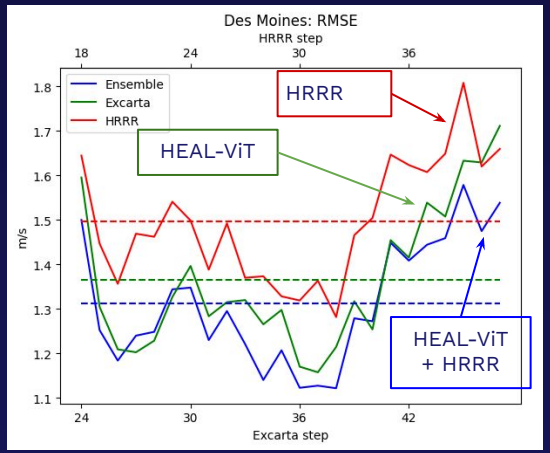
- 1 year of hourly METAR observations (from NOAA's MADIS)
- HEAL-ViT
  - Forecast issued 00 UTC each day, evaluate step 24-48
- HRRR
  - Forecast issued 06 UTC each day, evaluate step 18-42
- HEAL-ViT + HRRR: Naive ensemble of HEAL-ViT + HRRR

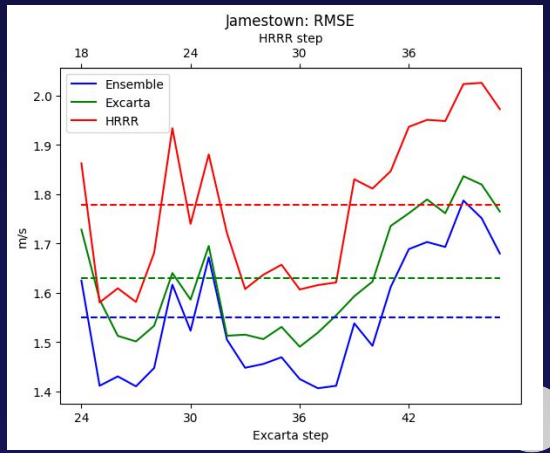
HEAL-ViT outperforms HRRR, but HEAL-ViT+HRRR outperforms both.



Great Plains

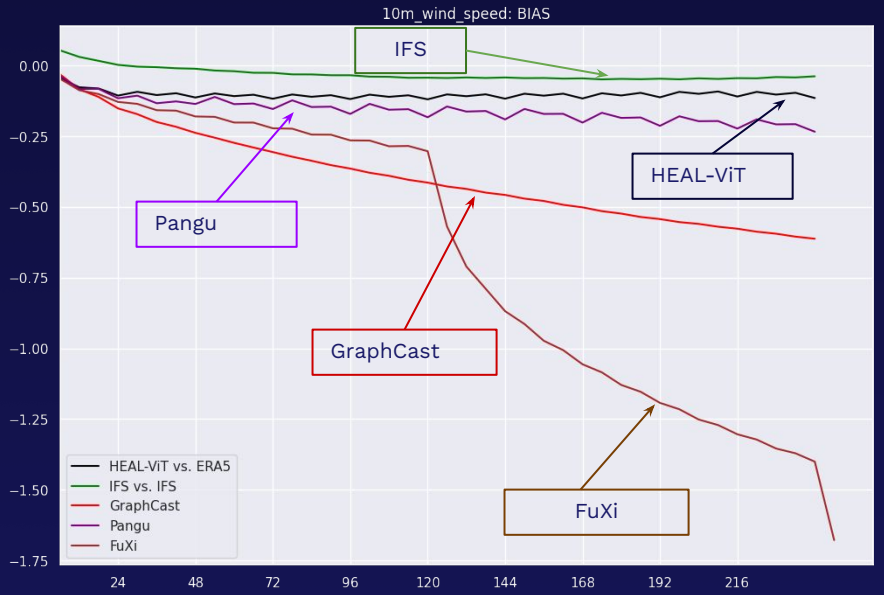






“Best” approach is not obvious.

- Many approaches work: Fourier operators, vision transformers, graph-based approaches…
- Currently: RMSE is the metric used to rank models
- But other metrics cannot be ignored
  - MLWPs struggle with predicting cyclone intensity, especially for weaker storms
  - Seems to be related to the accumulation of negative bias



More thorough benchmarks are needed.

In the absence of “standard” benchmarks, ML community defines its own

- FourCastNet: Improve RMSE
- Pangu-Weather: Cyclone track prediction using IBTrACS (doesn’t include all storms)
- GraphCast: Much more thorough, (e.g., precision/recall tests for extreme heat/cold events)
- WeatherBench2: very helpful in standardizing many evals

All of the above are good, but not complete

- Trade off between bias and RMSE? Spectra?
- Physical consistency, model stability, long-tail events, sensitivity to initial conditions…
- Some metrics less relevant for NWPs, but more for MLWPs

These questions are critical for R2O, not just “picking the winner”

- Non-trivial effort to build + operationalize + maintain MLWPs: which approach?
- What idiosyncrasies of the MLWP should end-users/forecasters know about?
- What kinds of continuous testing (unique to MLWPs) is necessary?
- What else does the MLWP need to produce to be trustworthy?

Domain experts have an opportunity!

It looks like many ML techniques work “out of the box”...

- ... and the low-hanging fruit has been picked
- Larger models, new architectures, more fine-tuning, likely to keep squeezing more RMSE
- But squeezing RMSE won’t lead to meaningfully better models
    - “When a measure becomes a target, it ceases to be a good measure”

Can we draw inspiration from other domains, e.g., protein folding?

- Protein Structure Prediction Center runs CASP (Critical Assessment of protein Structure Prediction) every 2 years
- Challenge: predict recently discovered protein structures
- Well understood problem, accepted by practitioners as a *meaningful* challenge
- Excelling on CASP => AlphaFold/AlphaFold2 could indeed “solve the problem”

What’s the CASP for weather prediction?

- “Grand challenge” can provide direction, and focus the field on the right problems
- Can accelerate meaningful, beneficial progress
- Build trust, ease adoption amongst (justified) skeptics

TL;DR

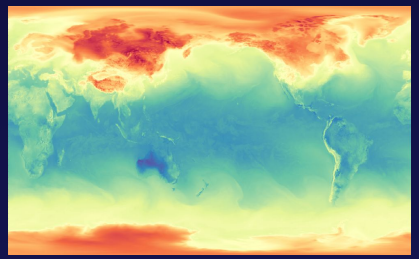
- ML Weather Models already offer tangible value, and complement NWPs
- Need to start seriously thinking about how MLWPs should be used alongside NWPs
- Many ML techniques work, but not obvious which one is the “best”
- Domain experts have an opportunity to provide leadership/guidance
- Operational and engineering constraints are unique to MLWPs
- Can/should learn from other fields where ML has been deployed successfully

HEAL-ViT: Vision transformers on a spherical mesh

HEAL-ViT

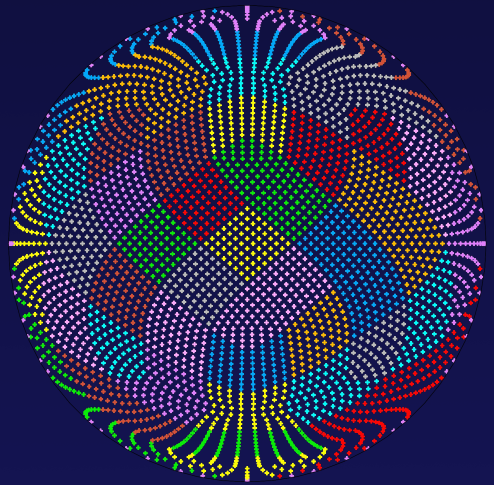
- Use HEALPix mesh to get a uniform spherical mesh
  - No distortion at poles, each pixel represents an equal area
- ViTs on spherical mesh, more memory efficient than graph models
  - Helpful when running other models needed for operational forecasts
  - “Regularity” of HEALPix mesh makes ViTs easier than on icosahedral meshes
- Other advantages (better bias, spectra)

(Arxiv preprint coming ~1 week!)



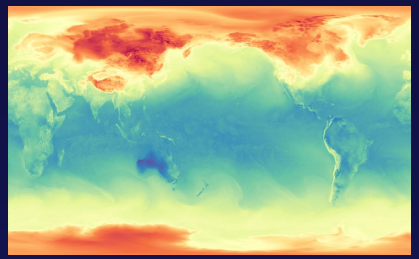
Lat-lon rectilinear grid

Graph Net



Vision Transformer on HEALPix mesh

Graph Net



Lat-lon rectilinear grid