

Do AI models produce better atmospheric river forecasts than physics-based models? A quantitative evaluation

Isaac Davis¹, Aneesh Subramanian¹, Timothy Higgins¹, Luca Delle Monache², Agniv Sengupta²



Project Background







- Machine learning (ML) numerical weather prediction (NWP) models have demonstrated remarkable accuracy and efficiency compared to traditional NWP models
- They perform as well or better than the European Centre for Medium-range Weather Forecasts (ECMWF) system on root mean square error (RMSE) for most variables, with orders of magnitude less compute
- However, no comprehensive study has looked at their ability to forecast Atmospheric Rivers (ARs)
- Its crucial to understand their forecast capability as these models become more and more commonplace

The Models and the Methods

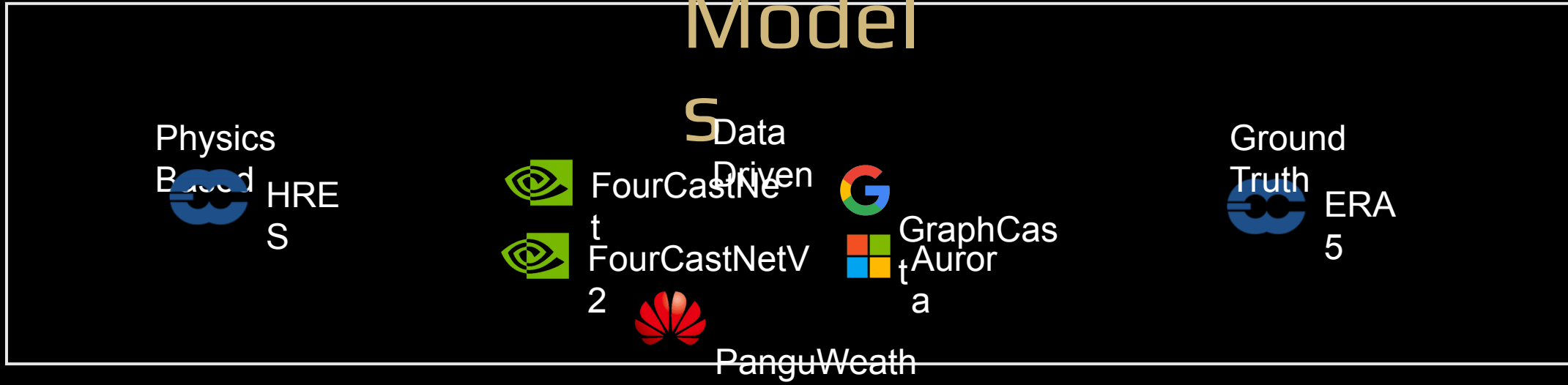


CW3E

Model Descriptions

Model	Architecture	Resolution	Variables
Graphcast	 Graph neural net in “encode-process-decode”	0.25° x 0.25°, 37 Pressure Levels, 6 hourly	12
FourCastNet	 Vision transformer with Adaptive Fourier Neural Operator	0.25° x 0.25°, 13 Pressure Levels, 6 hourly	14
FourCastNet-V2	 Spherical Fourier Neural Operators	0.25° x 0.25°, 13 Pressure Levels, 6 hourly	13
Aurora	 3D Perceiver encoder with Multi-scale 3D Swin Transformer U-Net backbone	0.25° x 0.25°, 13 Pressure Levels, 6 hourly	9
PanguWeather	 3D Earth-specific transformer with encoder–decoder	0.25° x 0.25°, 13 Pressure Levels, 6 hourly	9
HRES	 Conventional numerical weather prediction model	0.1° x 0.1°, 137 Pressure Levels, 6 hourly	100s

Model



Methods

- 10-day Forecasts are initialized at 0 UTC each day from November 1, 2023, to March 31, 2024
 - Creates 152 10-day forecasts per model
- Compare forecast skill of the different models on the U.S. West Coast (USWC) across a variety of metrics, ERA5 is ground truth
 - USWC bounding box: 15°-65° Lat, 170°-250° Lon
- Atmospheric River (AR) masks are created with CG-Climate¹
- Forecast ARs attributed to reanalysis ARs by ATRISK² algorithm at 1000km threshold

¹ Higgins, T. B., Subramanian, A. C., Graubner, A., Kapp-Schwoerer, L., Watson, P. A. G., Sparrow, S., et al. (2023). Using deep learning for an analysis of atmospheric rivers in a high-resolution large ensemble climate data set. *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003495. <https://doi.org/10.1029/2022MS003495>

² DeFlorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. M. Ralph, and F. Vitart, 2018: Global Assessment of Atmospheric River Prediction Skill. *J. Hydrometeor.*, 19, 409–426, <https://doi.org/10.1175/JHM-D-17-0135.1>.

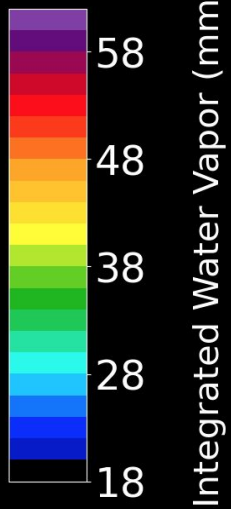
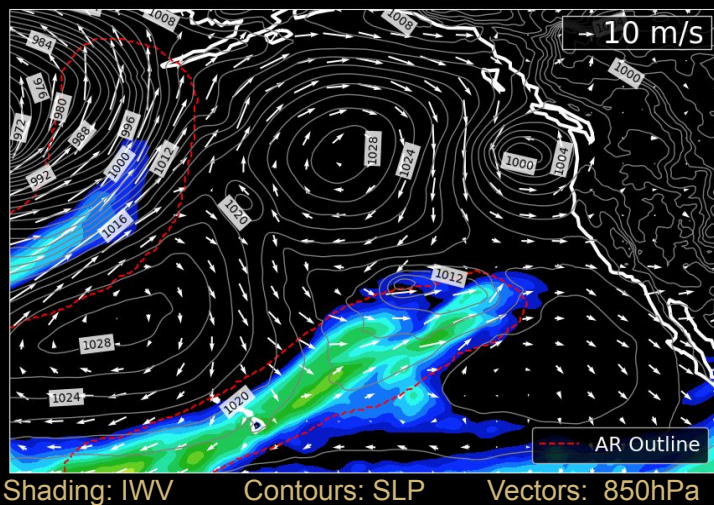
Results



CW3E

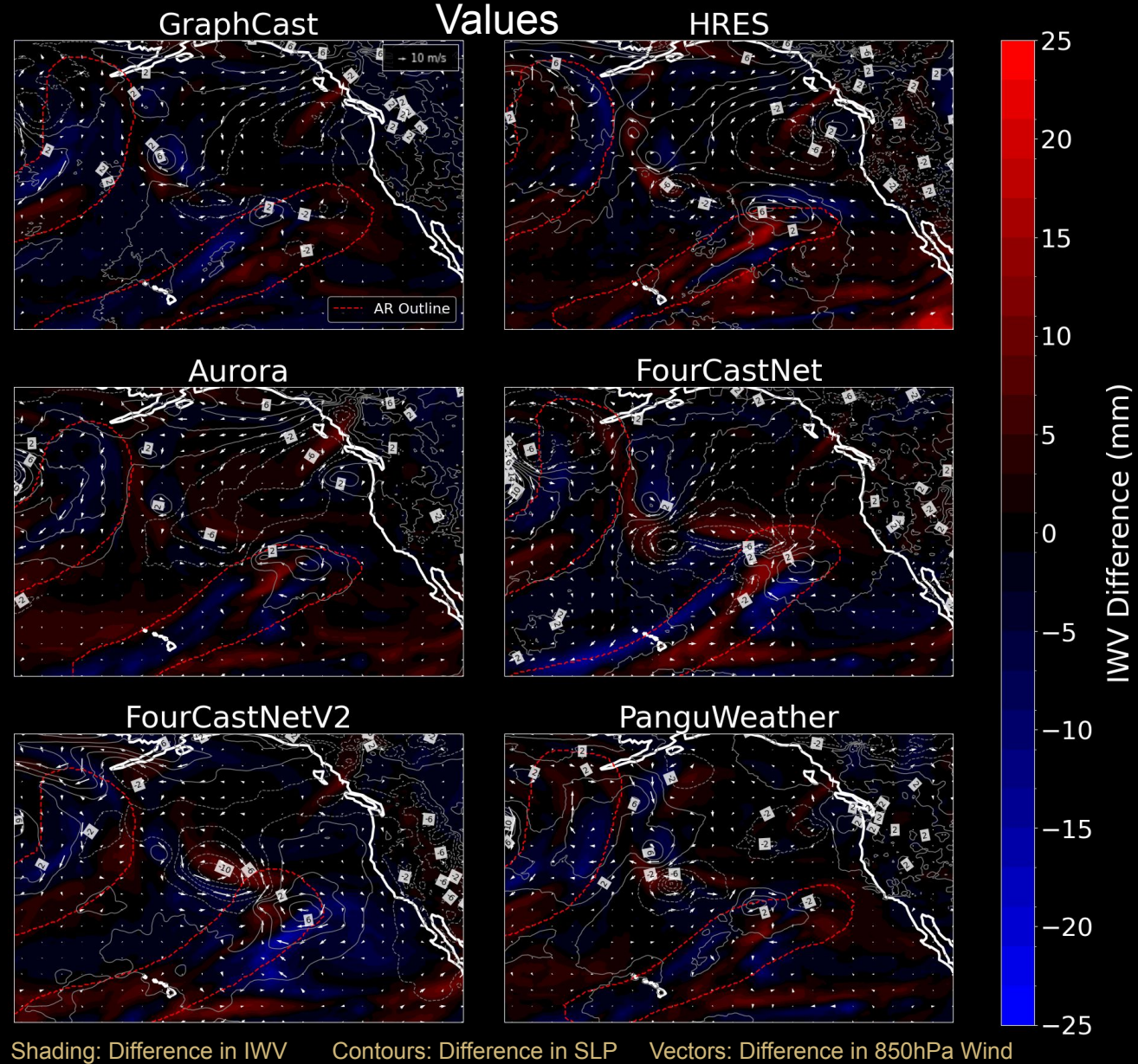
Case Study: Feb 3, 2024, 00 UTC AR Event

ERA5 Reanalysis:
Valid 02/02/24 00 UTC



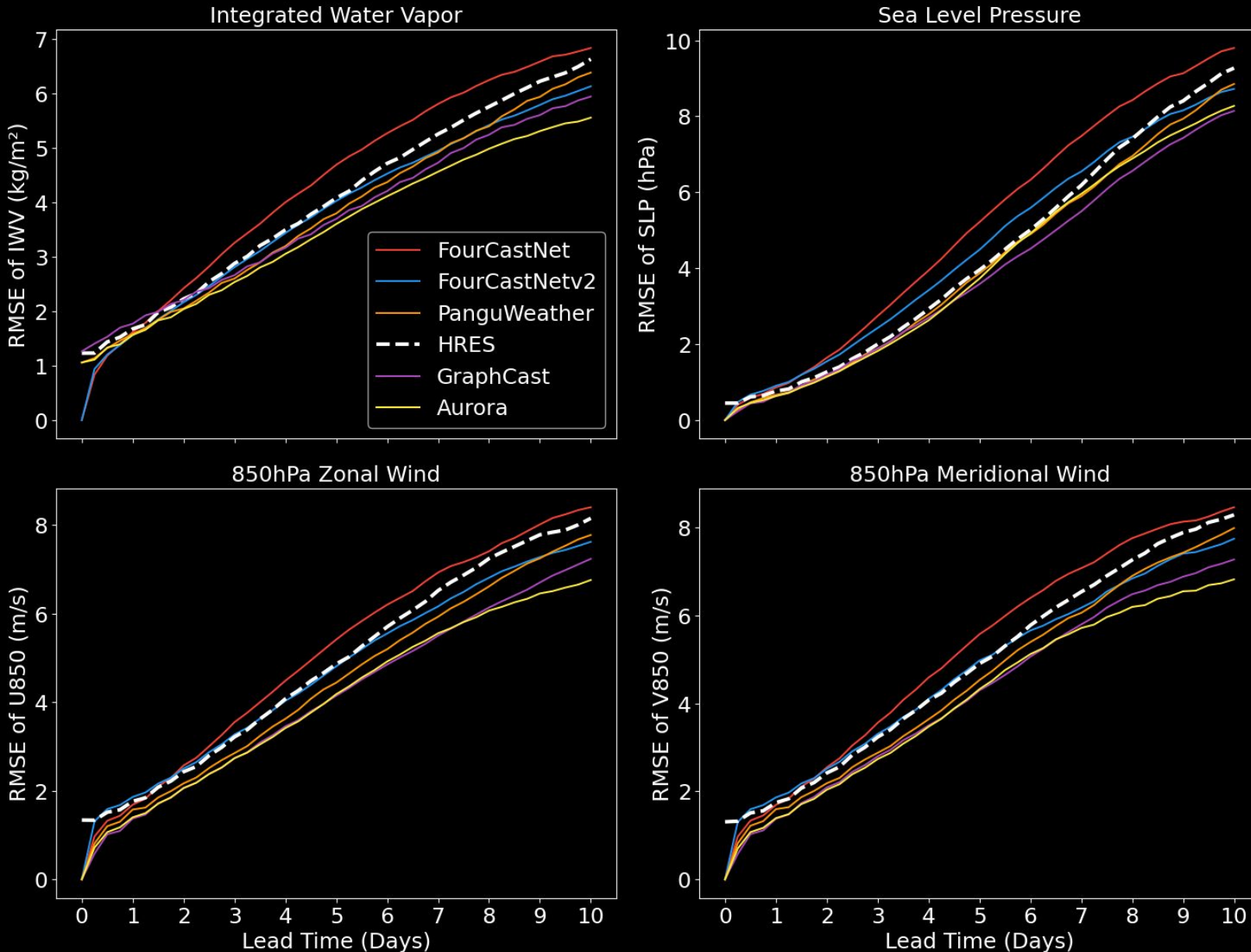
Integrated Water Vapor (IWV): The total amount of water vapor in a column of air from the surface to the top of the atmosphere, measured here in mm of water, also equal to kg/m^2

Difference Plots:
Four Day Leadtime Forecasts minus ERA5



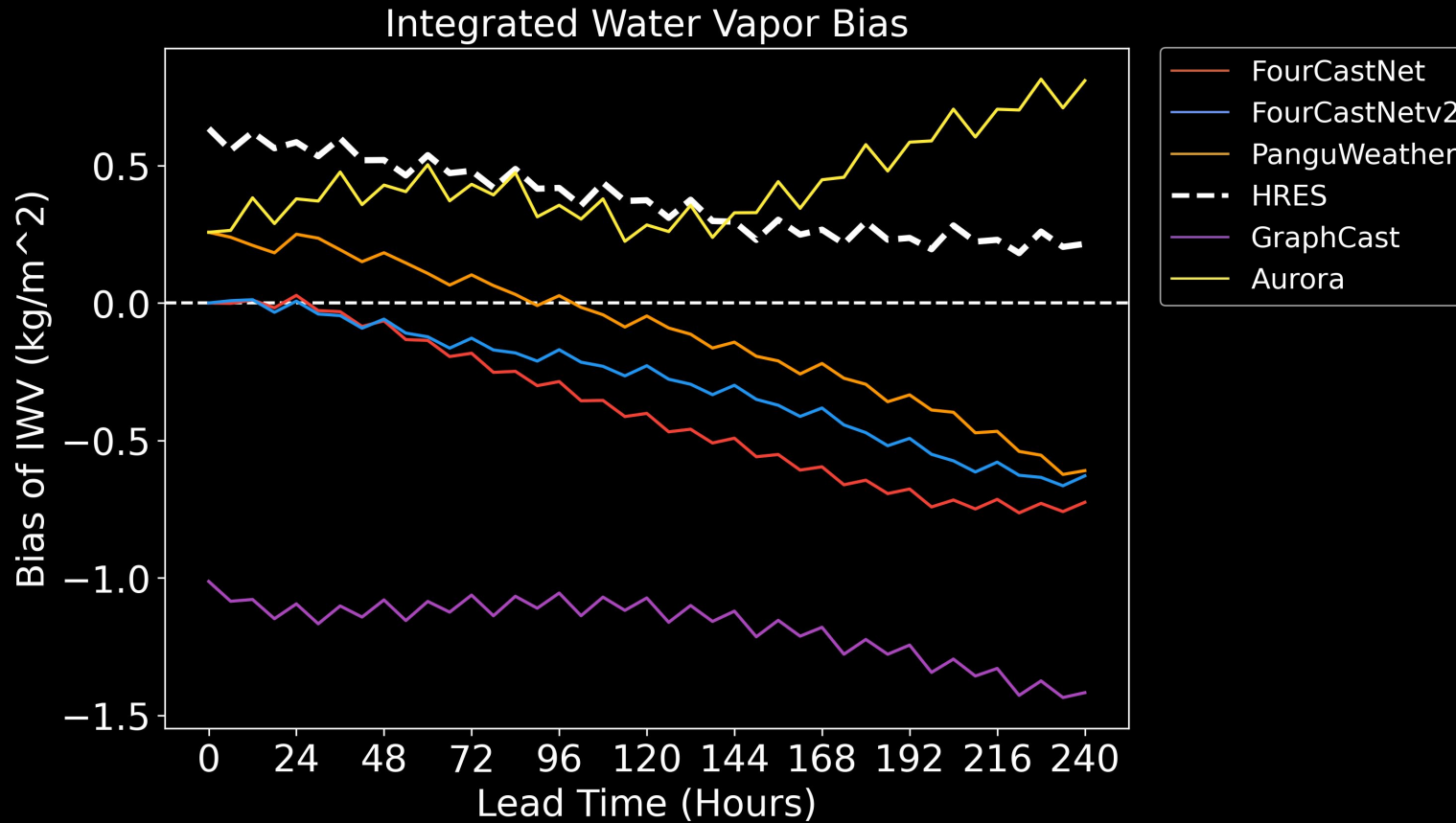
RMSE – U.S. West Coast

RMSE on U.S. West Coast vs Lead Time



- **Integrated Water Vapor (IWV):** The total amount of water vapor in a column of air from the surface to the top of the atmosphere, measured here in kg/m^2
- High performing **physics based HRES is outperformed by most of the AI models**, especially at longer lead times
- **Aurora and Graphcast are the highest performing models**, followed by PanguWeather, FourCastNetV2, then FourCastNet

Integrated Water Vapor Bias

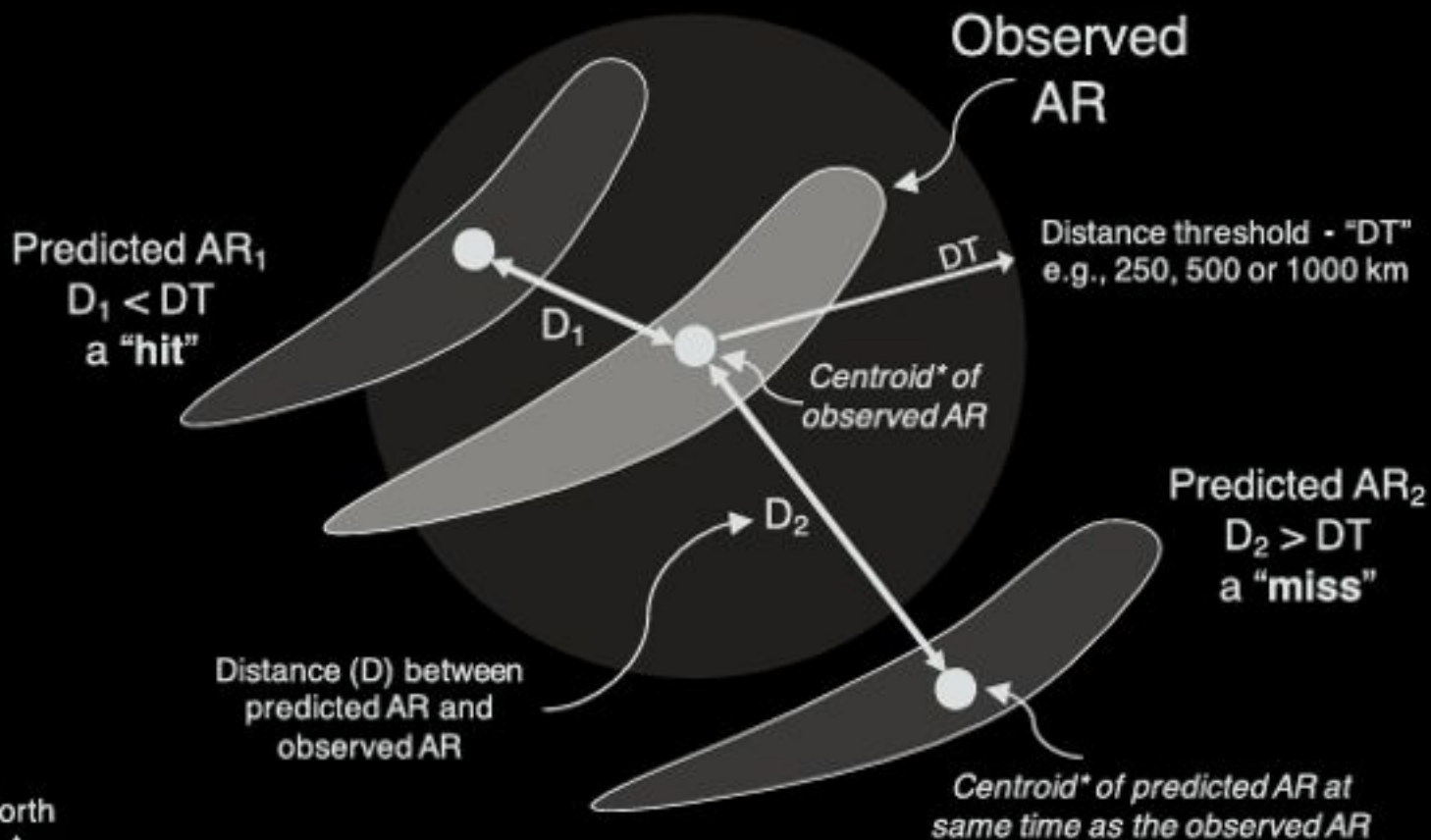


Graphcast has a consistent dry bias around 1-1.5 kg/m^2 , most models lose moisture with lead time.

Atmospheric River Skill (ATRISK)

Algorithm

Method of determining if a predicted atmospheric river (AR) is a “hit” or a “miss” relative to an observed AR

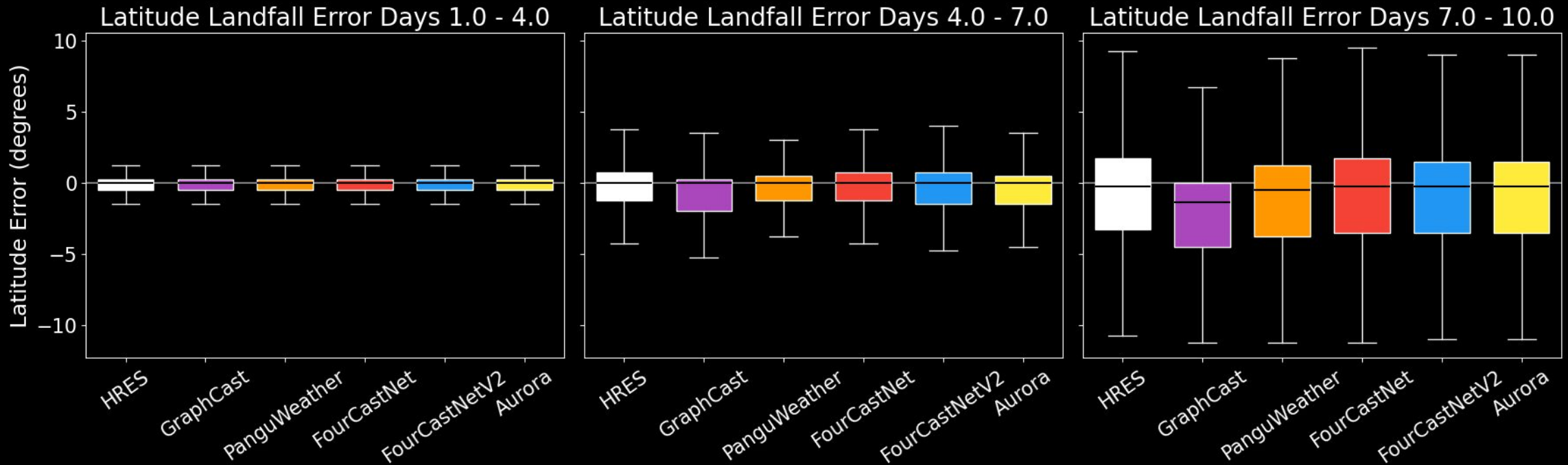


*Centroid is the IVT-weighted center of the AR, based on Guan and Waliser (2015)

- Algorithm developed and described in (DeFlorio et al., 2018)
- AR centers defined by the IVT-weighted centroid; I use IWV
- A forecast is considered a “hit” if the forecast AR centroid is within a user specified distance from an observed AR centroid
- Considered a miss if not within the threshold

Meridional Landfall Distributions

Meridional Landfall Error at Various Lead Times



- On a first order, performance is similar between HRES and the AI models
- Results are indistinguishable on days 1-4
- On days 4-7 GraphCast picks up a southern bias, other than that performance is still similar
- On days 7-10 GraphCast continues its southern bias, the other AI models perform very similarly to HRES

CSI, POD, FAR

- **Critical Success Index (CSI):** Measures the proportion of correctly predicted events (hits) out of the total actual and predicted events, accounting for both false alarms and missed events.

- Formula:

$$CSI = \frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False Alarms}}$$

- **Probability of Detection (POD):** Indicates how often an atmospheric river event is correctly forecasted when it occurs.

- Formula:

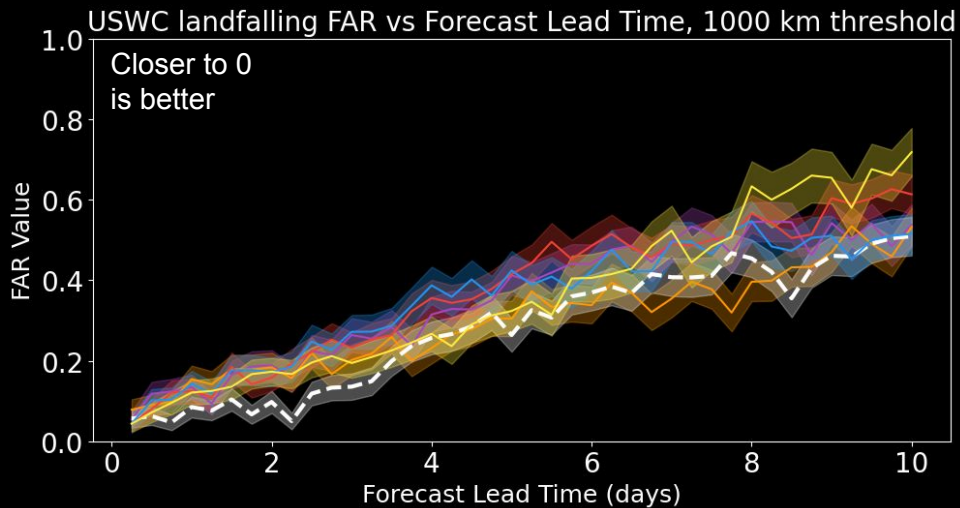
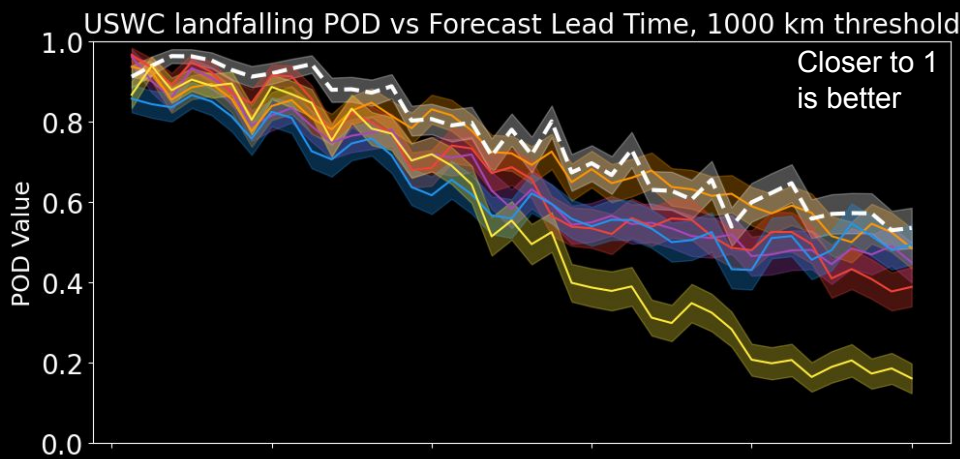
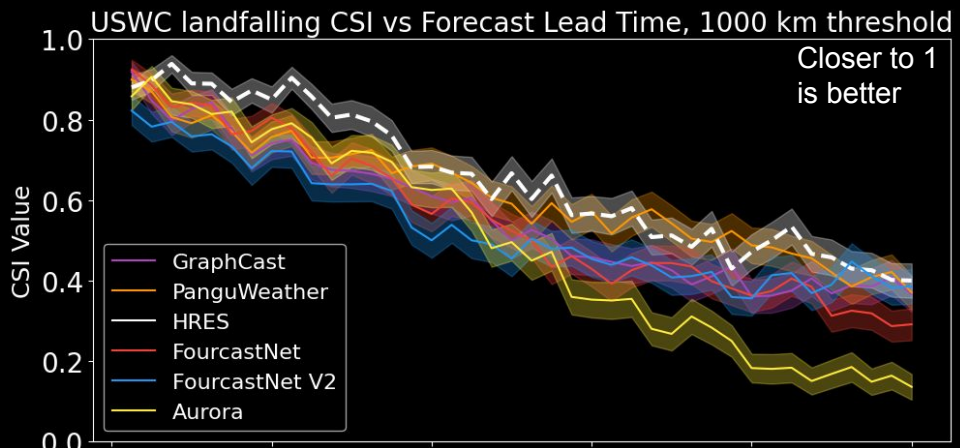
$$POD = \frac{\text{Hits}}{\text{Hits} + \text{Misses}}$$

- **False Alarm Ratio (FAR):** Shows the proportion of predicted events that did not occur

- Formula:

$$FAR = \frac{\text{False Alarms}}{\text{Hits} + \text{False Alarms}}$$

Critical Success index (CSI) Probability of Detection (POD) False Alarm Ratio (FAR)



- HRES is a top performer in all metrics
- Followed closely by PanguWeather after ~5 days lead time
- Top performer in RMSE, Aurora, is consistently bottom of the pack
- All other models perform very closely
- RMSE performance doesn't necessarily translate to these metrics
- Similar results are seen for a 500km detection threshold (not shown)

Plot shading shows
Standard Error:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Summary

- ML techniques show a significant improvement over HRES in reducing the RMSE for forecasting AR-related variables along the US West Coast.
 - Aurora and GraphCast are top performers
- Most of the ML models showed competitive performance to HRES in forecasts of USWC meridional landfall locations
- However, these models don't necessarily perform best for metrics such as CSI, POD and FAR
 - Only PanguWeather showed comparable skill to HRES, and after ~5 day lead times
 - Notably, leader in RMSE Aurora performed worst at this task

Conclusion

Despite the AI models decreased average RMSE, **physics based HRES is still the better option for forecasting actual AR events**. However, most of these AI models still proved to be worthy competitors.

Why might this be?

- **Lower resolution:** HRES is $0.1^\circ \times 0.1^\circ$ with 137 levels, vs $0.25^\circ \times 0.25^\circ$ with 13 levels (37 for GraphCast) for the AI models
- **Smoother Forecasts:** AI models produce smoother forecasts than physics-based ones, especially at longer lead times. This is a known issue and an artifact of the training targets
- **Training data:** AI models may not have seen enough examples of USWC ARs in the training data

Final Thoughts

- **RMSE is not everything:** AI NWP models are not ready to replace tried and true physics-based systems, even if they outperform in RMSE
- However, **these models are still first-generation systems** for the most part, and will improve in the future
- The **AI models still prove useful where computation power is a significant bottleneck**, and especially if RMSE performance is desired

Thank You!



Email: Isaac.Davis@Colorado.edu



CW3E