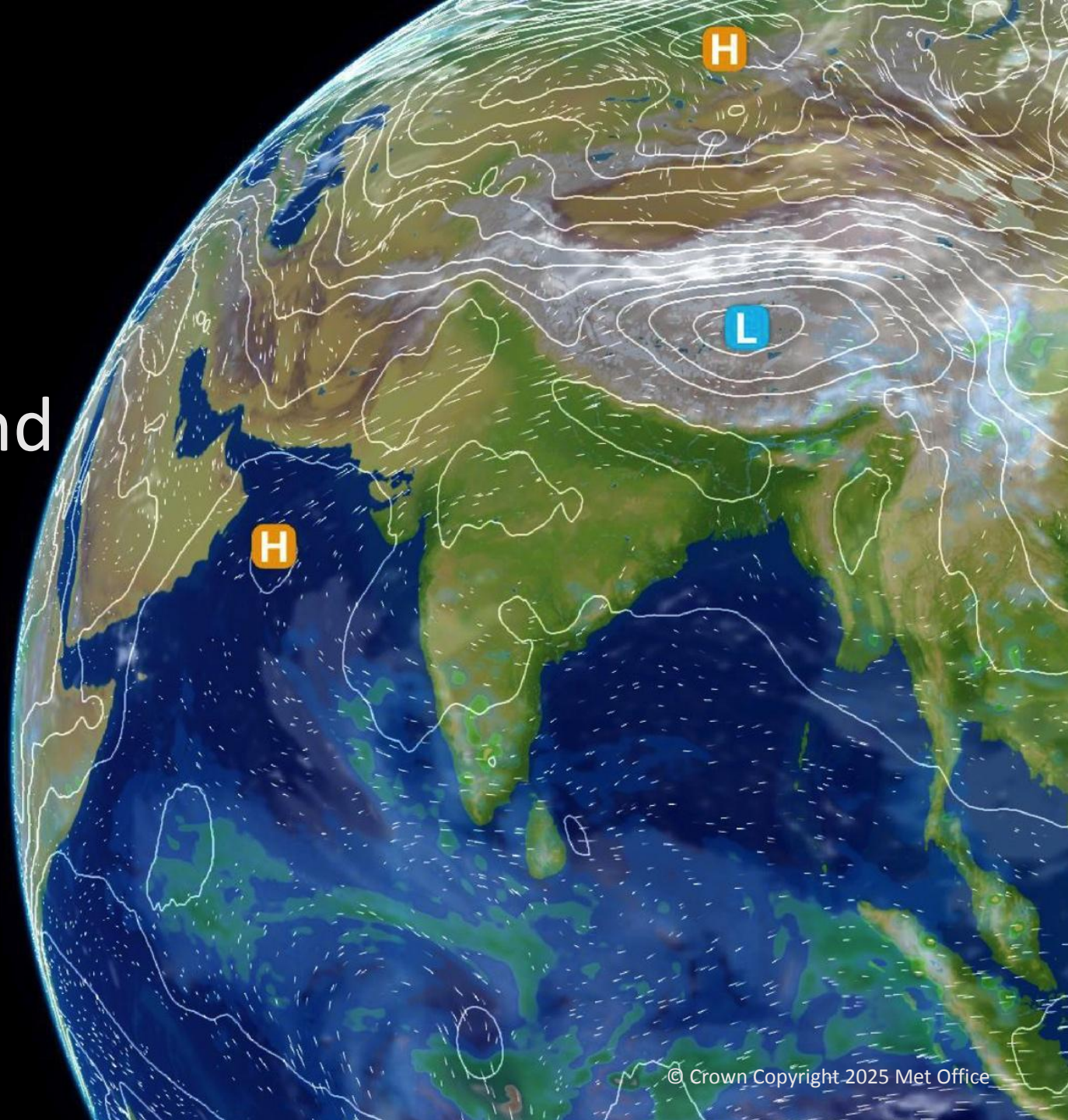
Error growth and predictability scales across models: a comparative analysis between middle and tropical latitudes

Richard Keane
Met Office, Exeter, UK
CEMAC, School of Earth and Environment, University of Leeds, UK

Thanks to: Doug Parker, Etienne Dunn-Sigouin, Erik Kolstad, John Marsham, Martin Willett



Motivation

- Short-range weather forecasting is more challenging at tropical latitudes than mid-latitudes.
  - But scope for more skilful longer-range forecasts (e.g., sub-seasonal, seasonal).
- This is generally accepted within the meteorological community.
  - Some theoretical understanding as to why this is the case.
    - Error growth at different rates for convective and baroclinic regimes.
    - Fewer observations at tropical latitudes?
    - Surface forcing more important at tropical latitudes?
- It would be useful to quantify this.
  - Determine at what time/space scale forecast performance becomes better at tropical latitudes.
  - Also important to demonstrate it systematically, to convince forecast users outside the meteorological community.

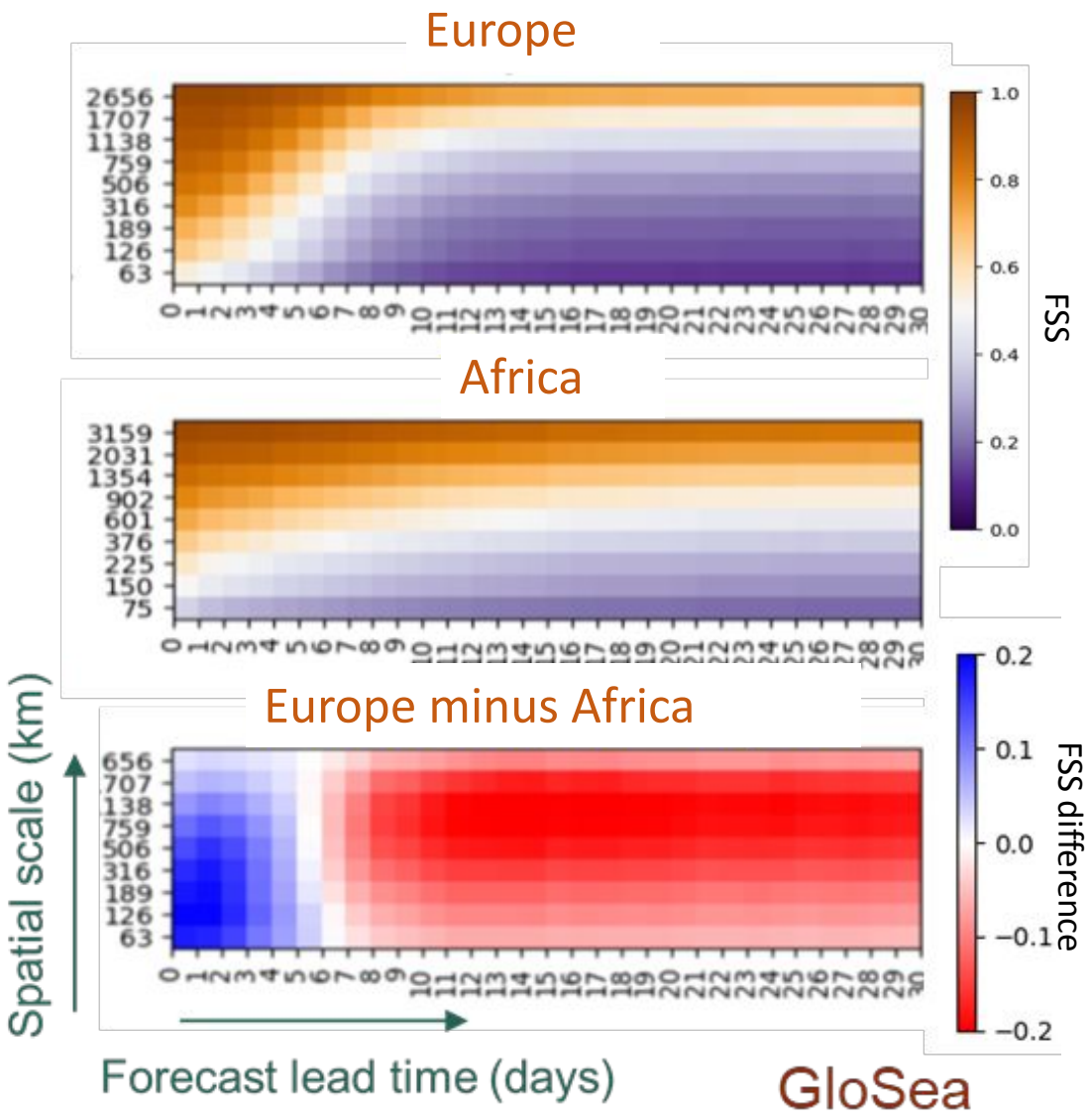
Method

- Evaluate the following models:
  - Met Office seasonal hindcasts (GloSea6, 2001—2016).
  - Met Office (2020) and ECMWF (IFS: 2022, 2024) operational weather forecasts.
  - ECMWF AIFS machine-learning-based forecasts (2024).
- Verification is carried out separately over tropical latitudes (15S to 15N) and northern hemisphere middle latitudes (30N to 60N).
  - First look at Europe and Africa (both 30W to 60E).
    - Motivated by particularly poor weather forecast performance over Africa.
  - Then look at whole latitude bands.
  - Look also at global maps.

Fractions skill score

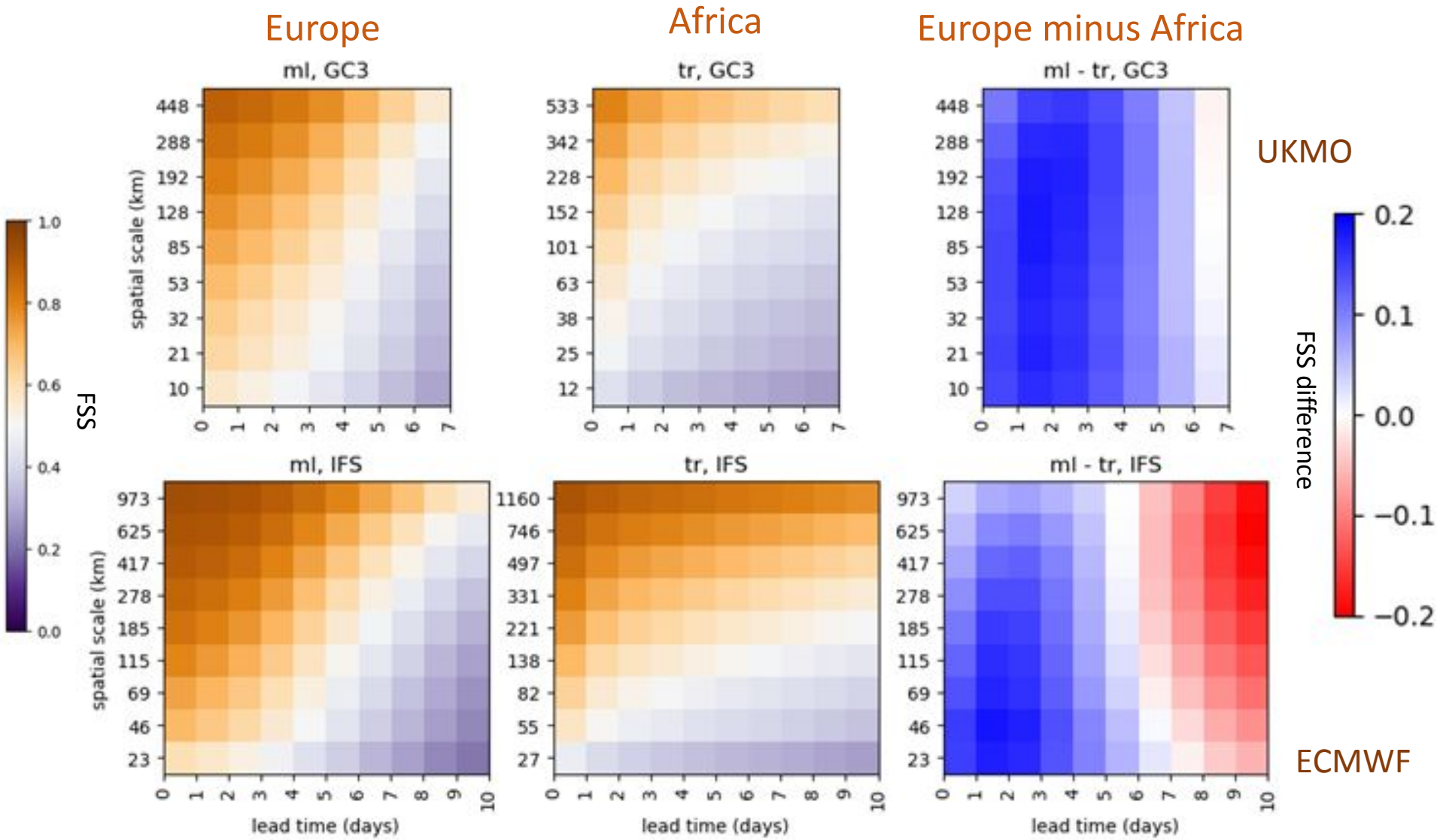
- This provides a way of evaluating forecasts at different spatial scales.
  - Threshold-based binary precipitation fields are averaged over different numbers of grid points.
    - The threshold is chosen as a percentile over the climatology for that grid point.
    - Results shown here are for a 90% threshold.
      - Defined separately for forecasts and observations (IMERG data).
  - Scores range from 0 (no skill) to 1 (perfect forecast **at the relevant scale**).
    - Generally increase with increasing scale.
- The scores are calculated for different lead times.
  - Generally decrease with increasing lead time.

Results for GloSea



- Forecast quality is better (orange colours) for larger scales and shorter lead times.
  - Quality degrades more slowly with lead time in Africa.
- Overall, forecast quality is better in Europe on shorter scales (blue colours).
- Forecast quality is better in Africa on longer scales (red colours).
- There is a crossover time between these two regimes at 5—7 days.
  - The spatial scale has less of an effect, although the crossover time is slightly shorter for larger scales.

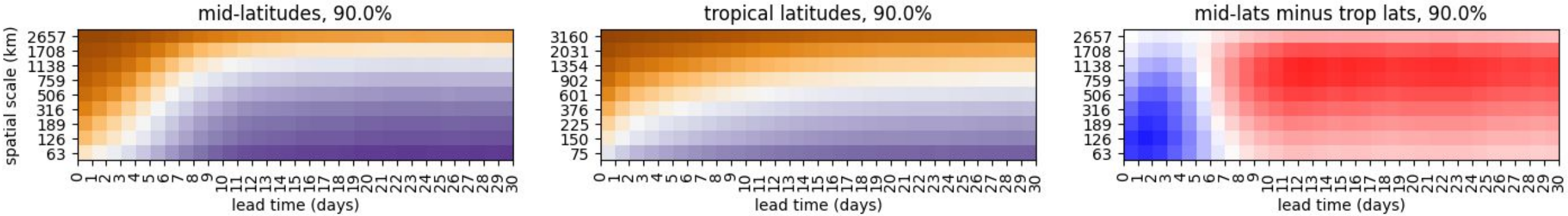
Results for other models



- The same result holds for weather models with different grid spacings.
  - Similar ‘crossover’ times of 5—7 days.
- Suggests it may be a fundamental property of atmospheric predictability.

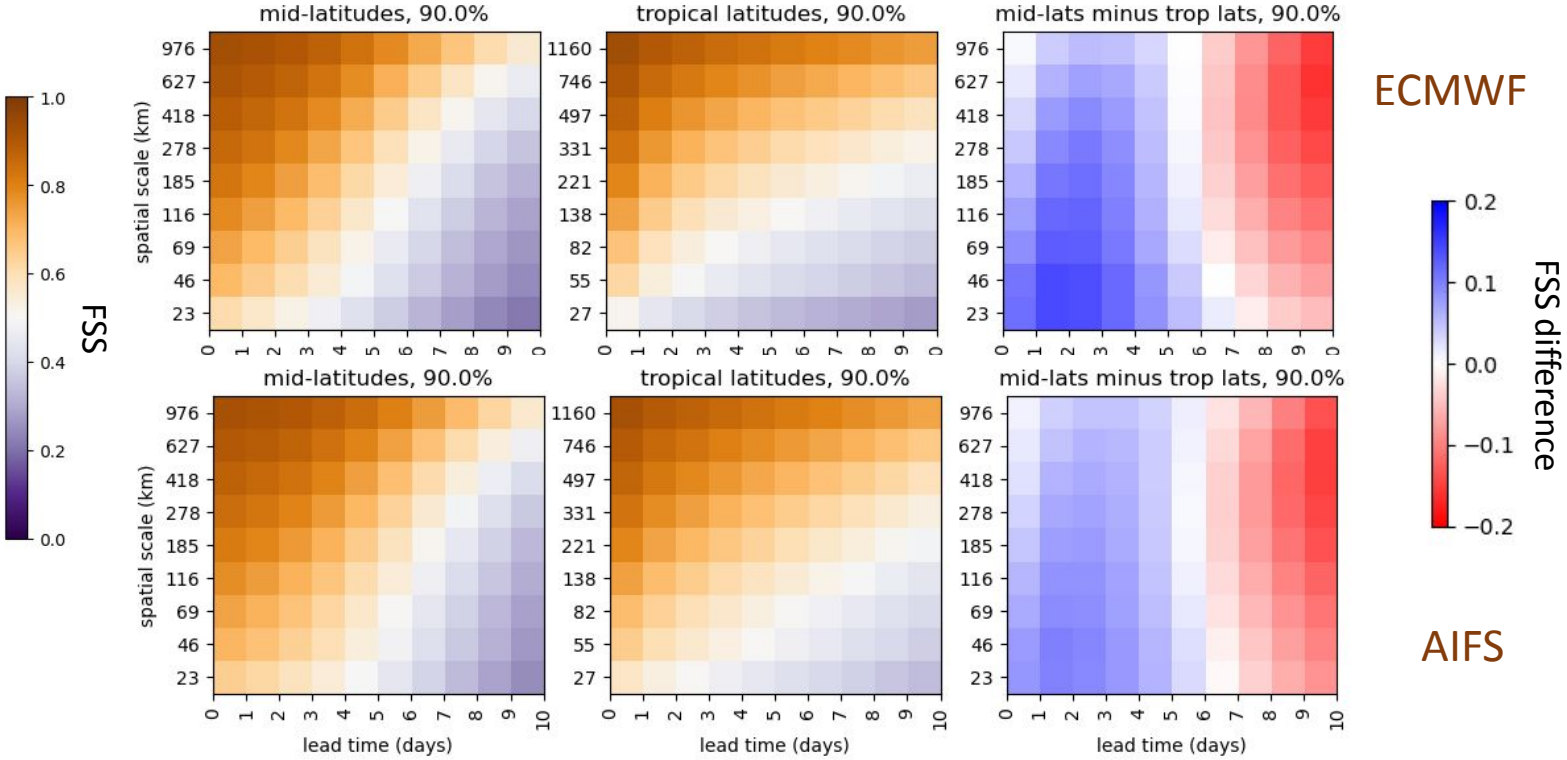
Application to full latitude bands

GloSea



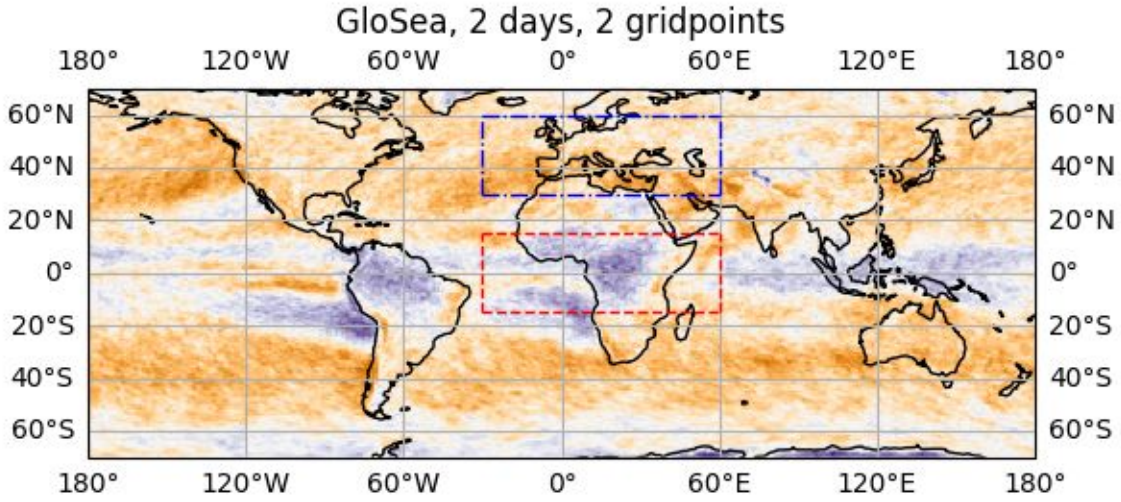
- Similar behaviour is seen, with a crossover of about 5—8 days.
- Also applies to a machine learning model.
  - Further evidence it is a fundamental property of atmospheric predictability.

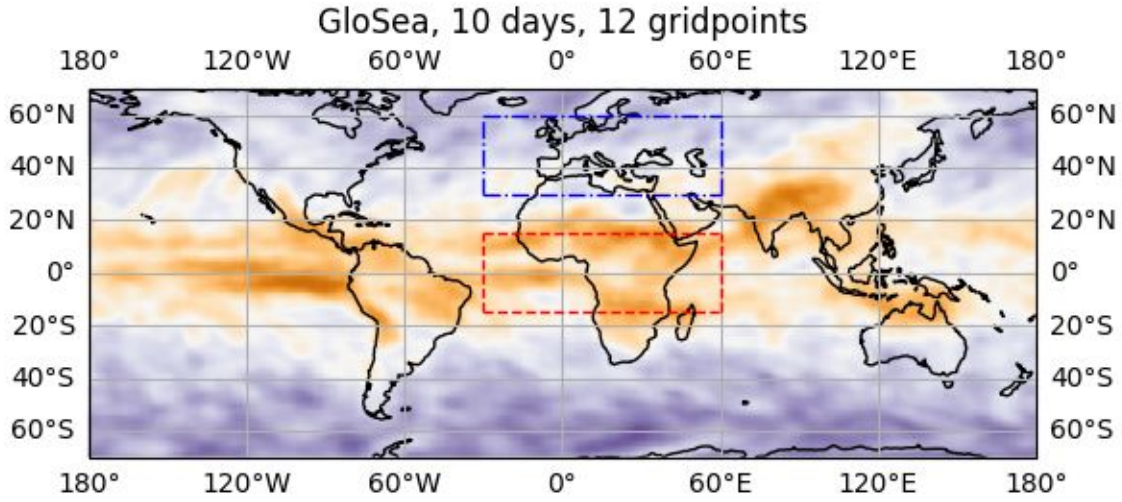
ECMWF

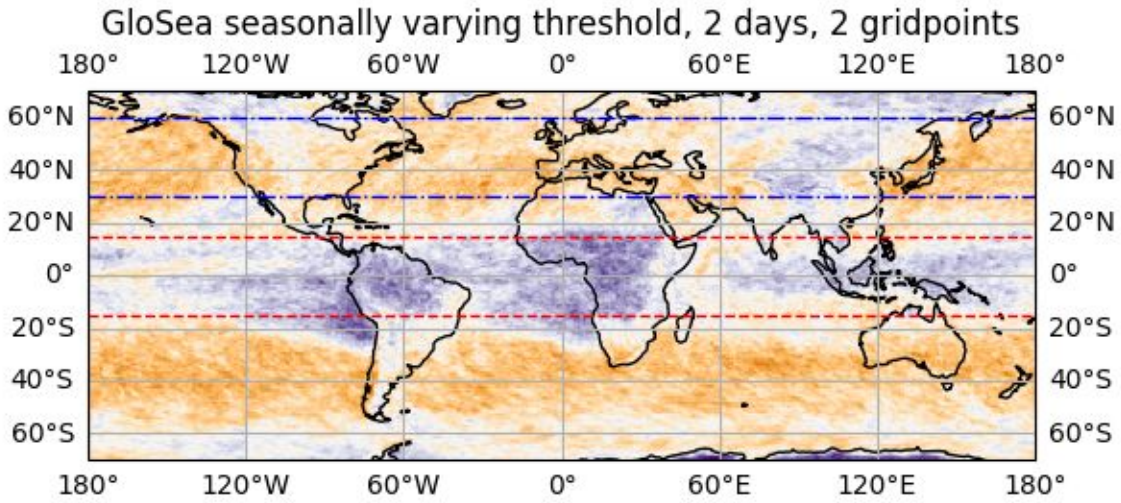


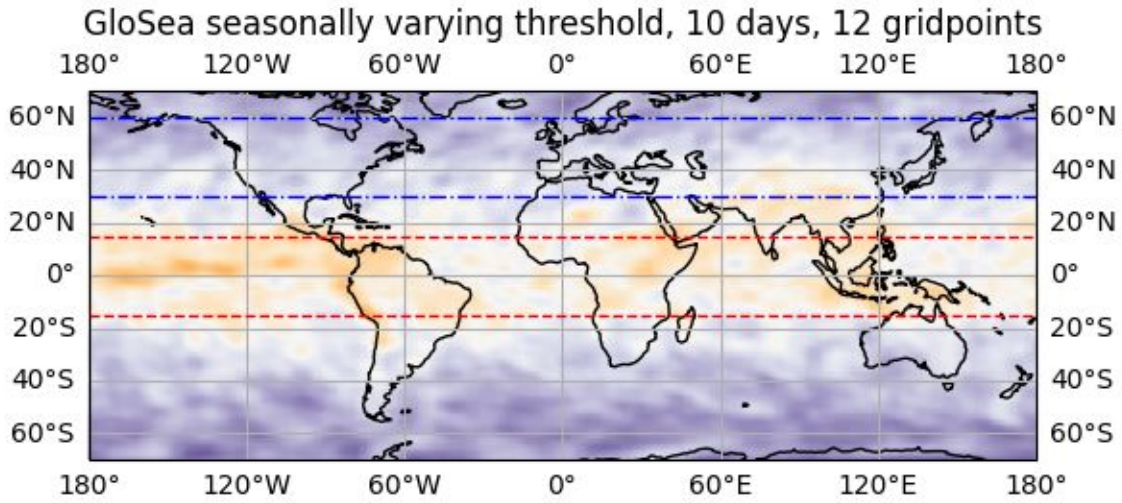
AIFS

Spatial


GloSea, 2 days, 2 gridpoints
180° 120°W 60°W 0° 60°E 120°E 180°
60°N 40°N 20°N 0° 20°S 40°S 60°S


GloSea, 10 days, 12 gridpoints
180° 120°W 60°W 0° 60°E 120°E 180°
60°N 40°N 20°N 0° 20°S 40°S 60°S

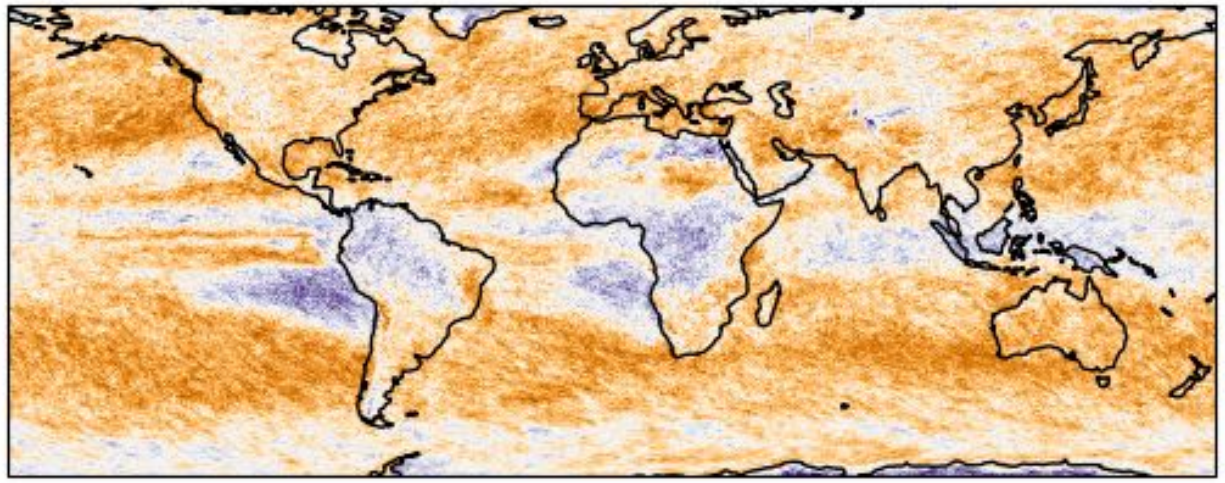

GloSea seasonally varying threshold, 2 days, 2 gridpoints
180° 120°W 60°W 0° 60°E 120°E 180°
60°N 40°N 20°N 0° 20°S 40°S 60°S


GloSea seasonally varying threshold, 10 days, 12 gridpoints
180° 120°W 60°W 0° 60°E 120°E 180°
60°N 40°N 20°N 0° 20°S 40°S 60°S

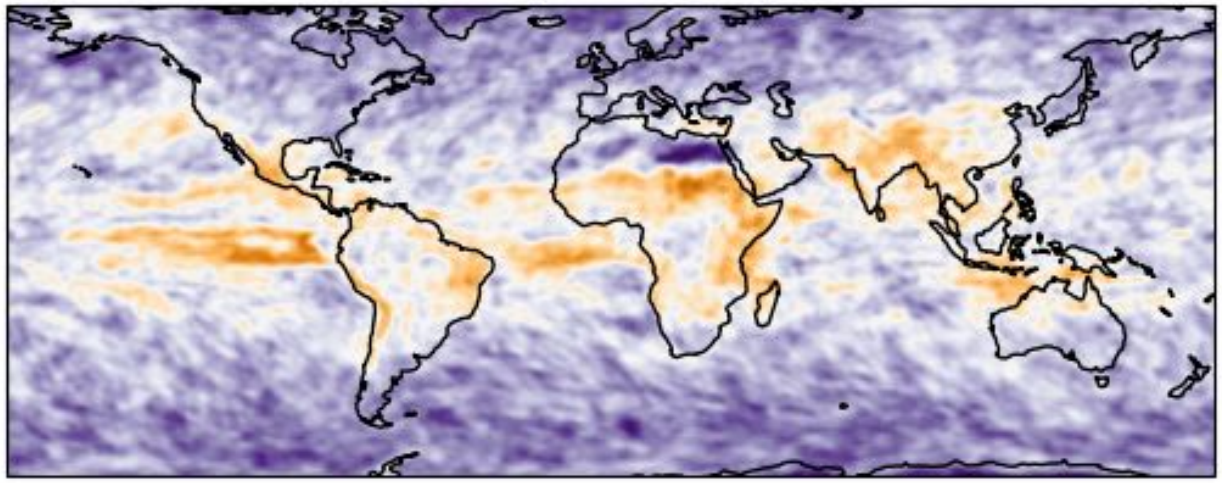

0.0 0.2 0.4 0.6 0.8 1.0

Comparing ML and physical

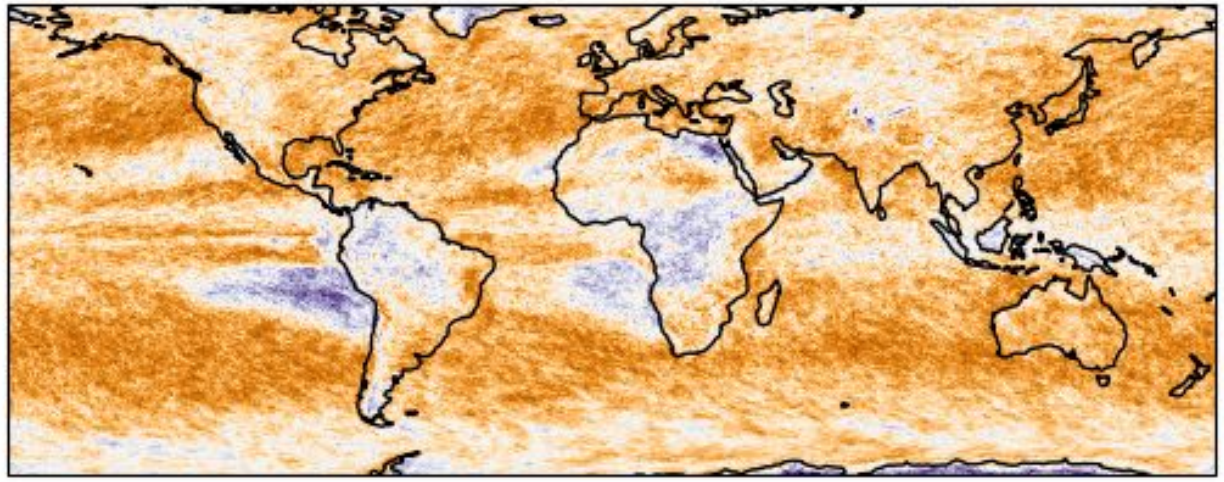
IFS, 2 days, 2 grid points



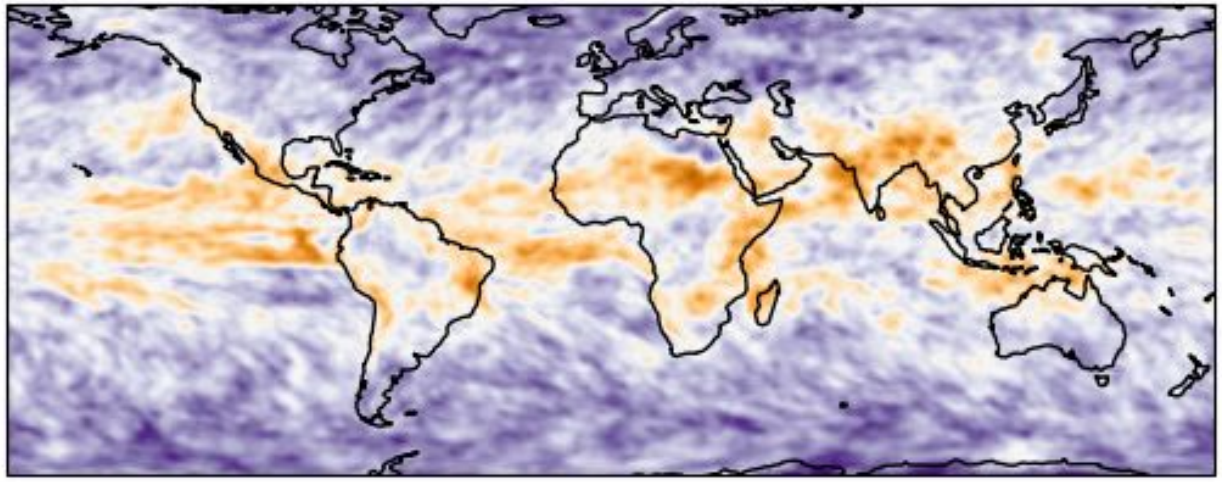
IFS, 10 days, 12 grid points



AIFS, 2 days, 2 grid points



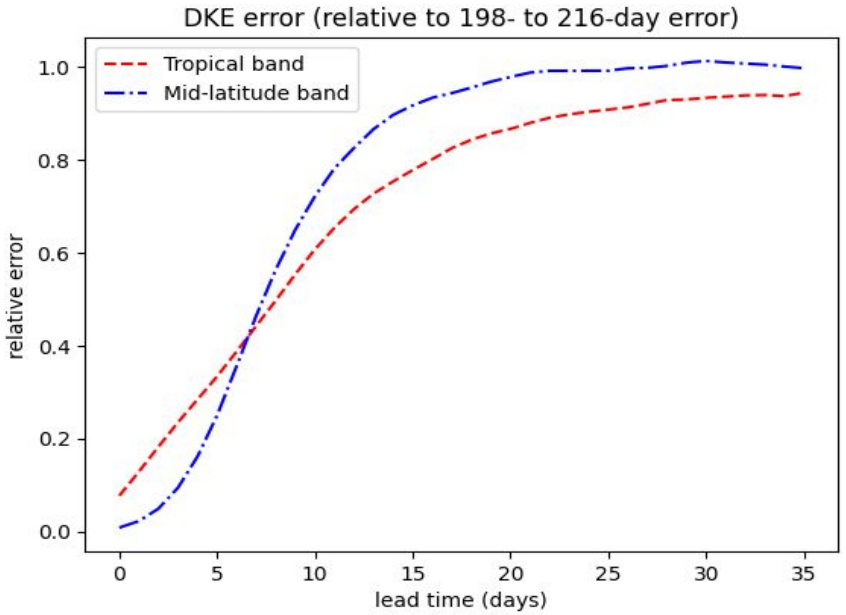
AIFS, 10 days, 12 grid points



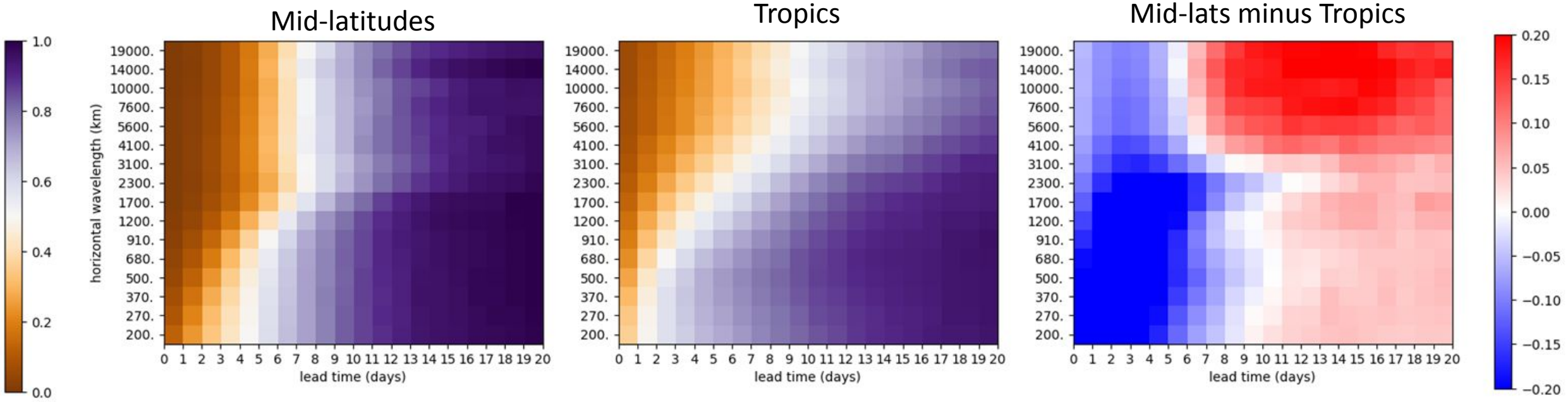
Difference kinetic energy

- Given by $\left(\left(u_f - u_a\right)^2 + \left(v_f - v_a\right)^2\right)/2$
  - Applied here to wind speeds at 500 hPa.
  - A quantity frequently used in theoretical studies on error growth.
- Calculate relative to saturation error.
  - For GloSea, use value at end of hindcast.
  - For shorter forecasts use
    - $\langle (f-a)^2 \rangle = \langle f-a \rangle^2 + \sigma_f^2 + \sigma_a^2 - 2\sigma_f \sigma_a \rho_{fa}$
    - So saturation error is
      - $\left(\mathrm{var}(u_f) + \mathrm{var}(u_f) + \mathrm{var}(u_f) + \mathrm{var}(u_f)\right)/2$
    - Bias is small compared to the other terms.

Results – difference kinetic energy

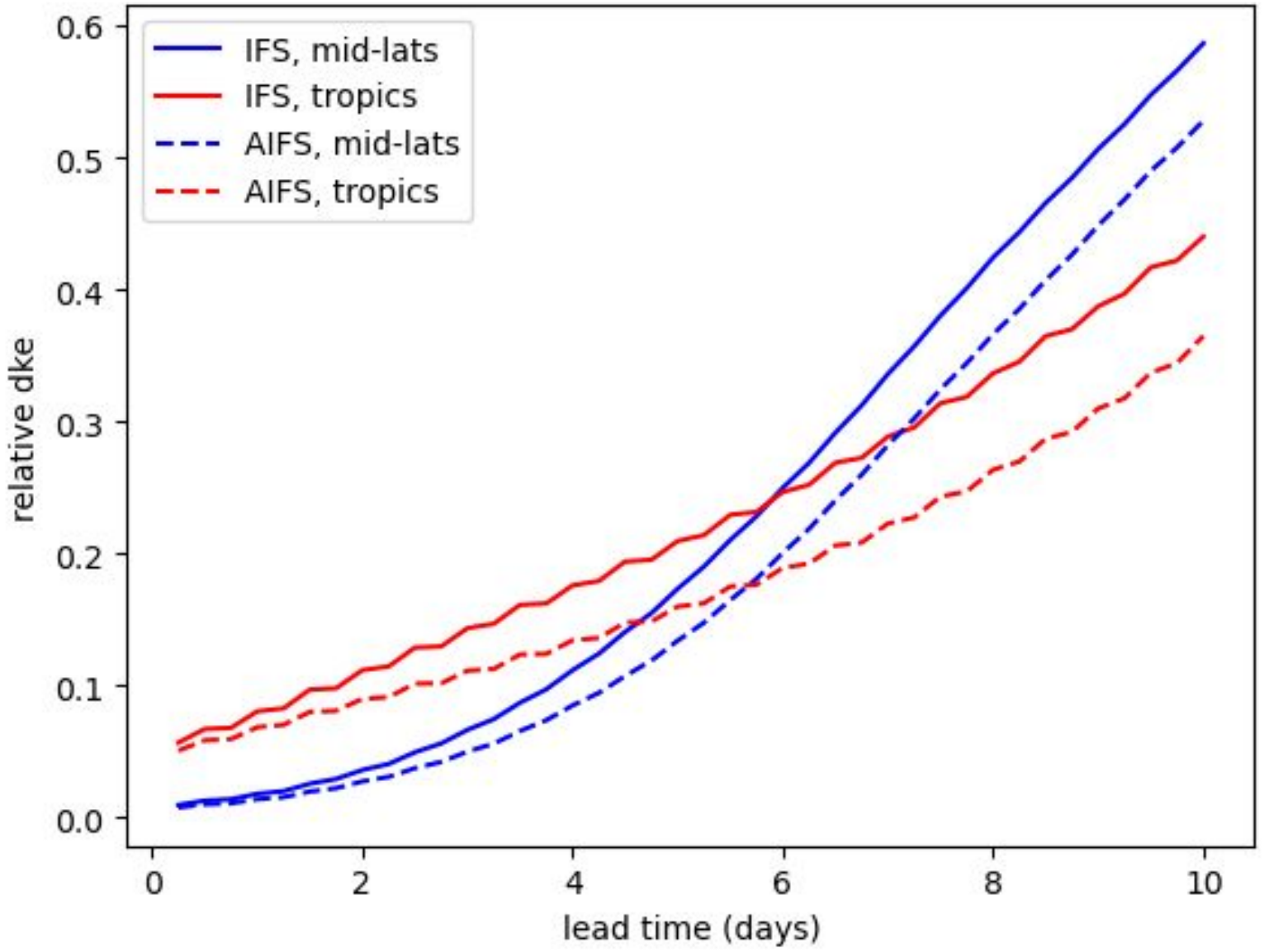


- Shown here for GloSea.
- Similar crossover time (about 7 days).
- Values are separated into spatial scales by taking a Fourier transform in longitude.
- The dependence on spatial scale is similar to that for FSS, although difficult to compare directly.



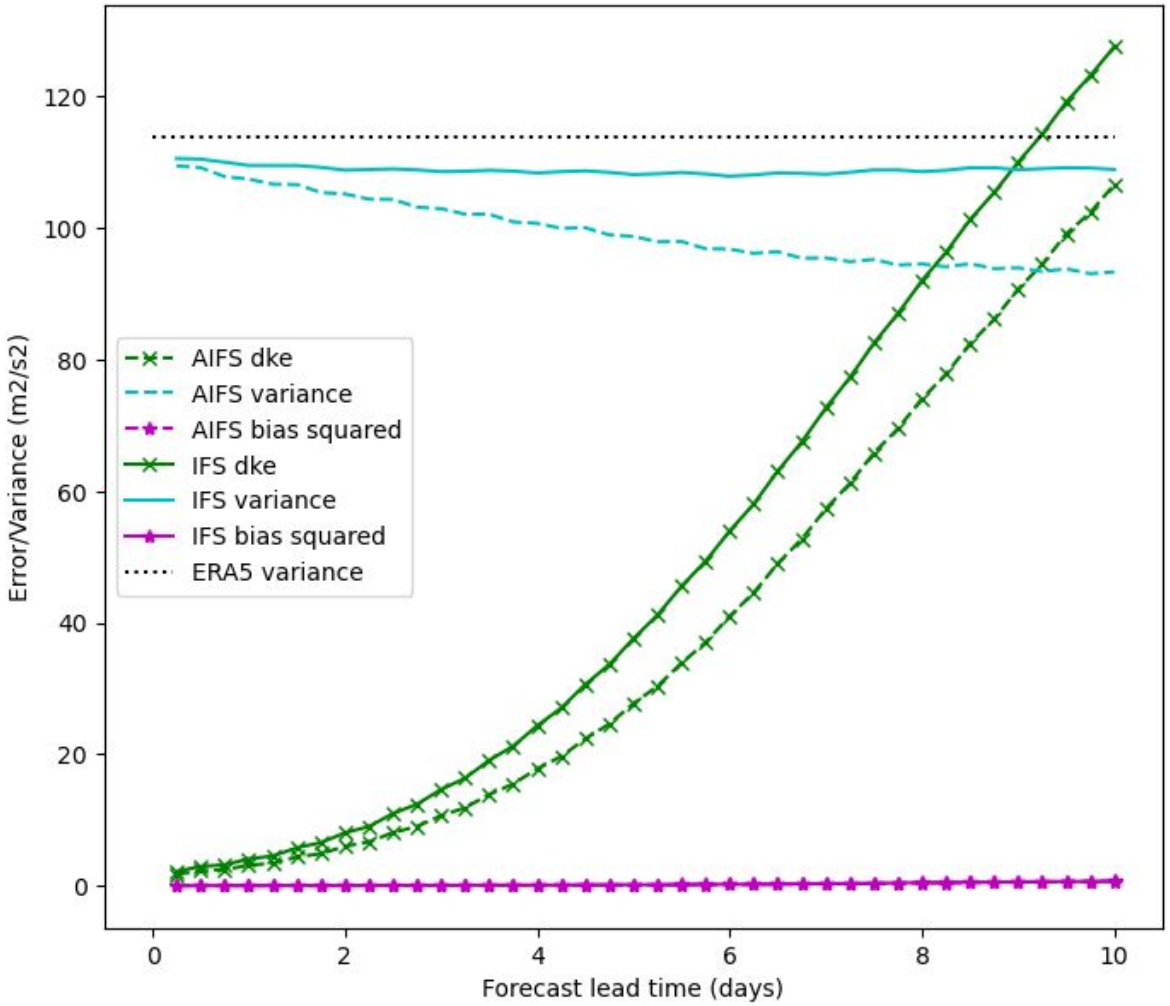
Comparing IFS and AIFS

- Error growth is slower in AIFS but they are comparable.
- Similar crossover time in both models, of about 6 days.

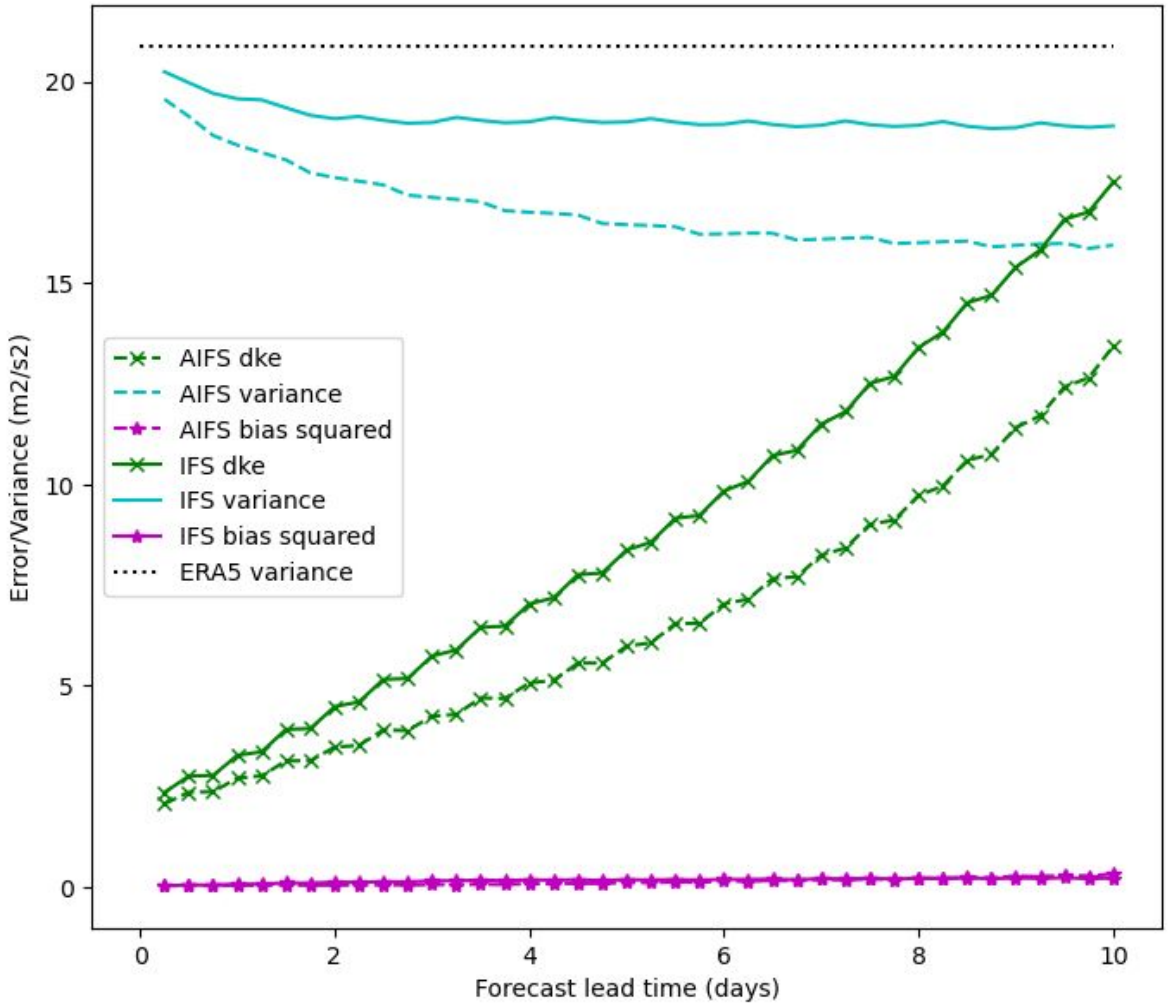


Splitting into components

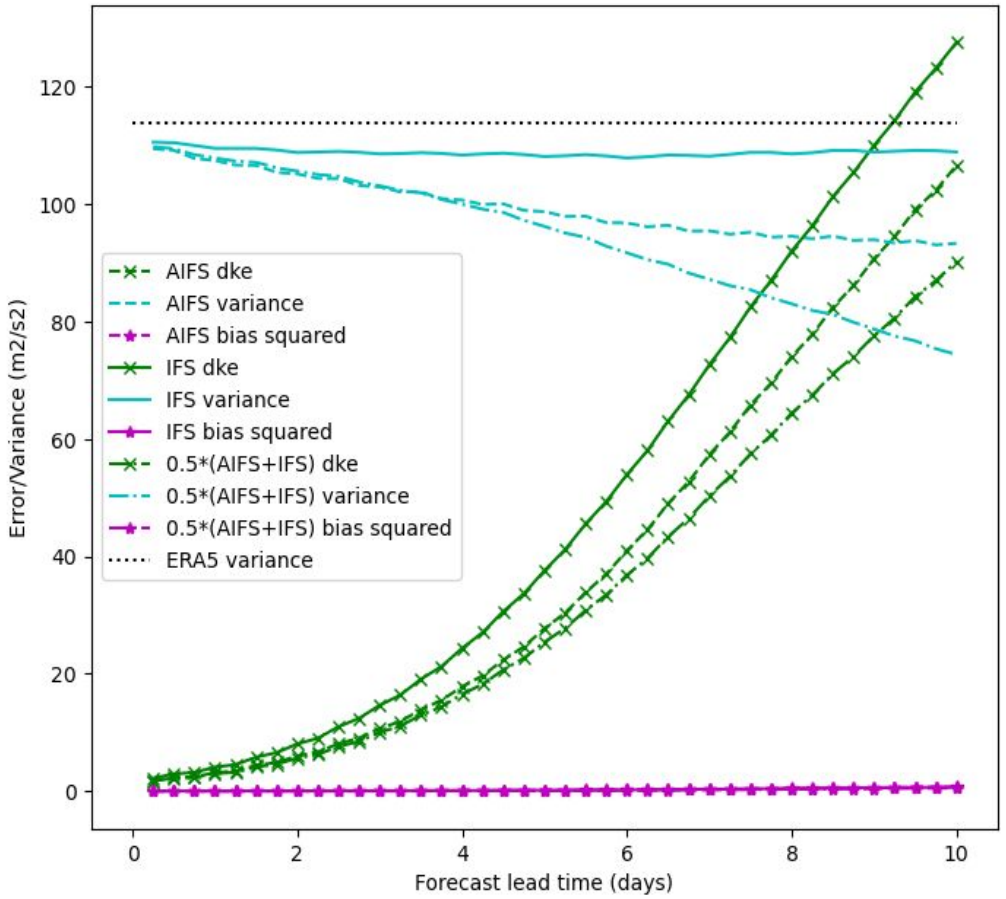
Middle latitudes



Tropical latitudes

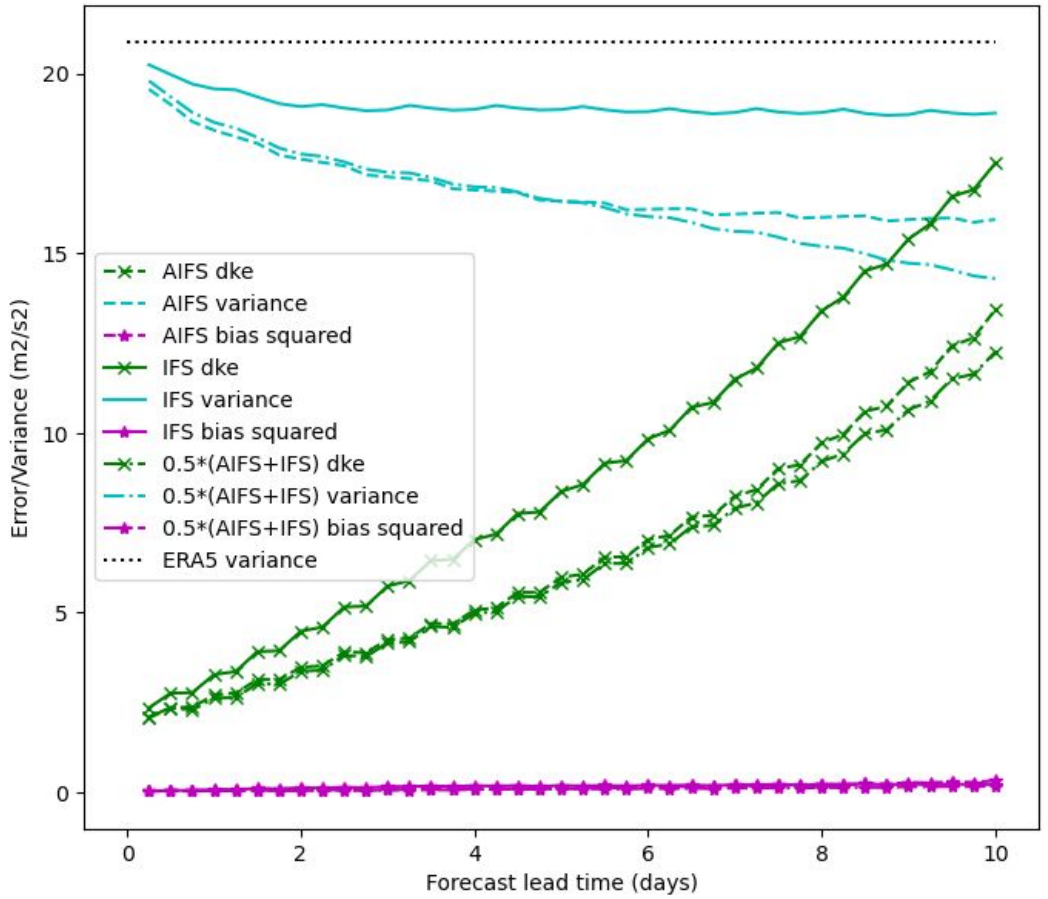


- Try using an average of the two forecasts.
  - This provides an improvement over AIFS alone.
    - Small at tropical latitudes, but almost as large as difference to IFS at middle latitudes.
  - Deviation from AIFS appears only after about 5 days.



Middle latitudes

Tropical latitudes



Summary I

- Systematic demonstration that weather and seasonal forecasts perform better on shorter scales at middle latitudes and on longer scales at tropical latitudes.
  - This has policy implications:
    - forecasting methods may need to be tailored separately to different regions of the world to a greater extent than is currently the case.
  - Forecasts based on machine learning look remarkably similar to those based on physical models, suggesting that this issue is still relevant for users relying on machine learning methods.

Summary II

- The result holds for both FSS verification of precipitation and upper-air kinetic energy error growth.
  - Links previous theoretical work with practical application.
- ECMWF machine-learning-based model has lower error growth than physical model.
  - Also better FSS, although this is less the case for larger scales (not shown).
  - However, forecast can be improved further by combining the two.
    - Suggests physical model does include information not in ml-based model.
    - This should be investigated further using ensemble forecasts.