Knowledge-guided machine learning reveals European month-ahead temperature is substantially more predictable than suggested by numerical models

Sem Vijverberg[1, 2], Jannes Ingen[1, 2], Dim Coumou[1, 2]

1. Vrije Universiteit Amsterdam, Amsterdam, NOORD-HOLLAND, Netherlands

2. spin-off company institute for environmental studies, Beyond Weather, Amsterdam, Noord Holland, Netherlands

Being able to predict European temperature weeks up to months in advance can help numerous energy stakeholders with future price estimations and optimizing resource allocation. However, traditional numerical weather predictions struggle to achieve skillful predictions at subseasonal-to-seasonal (S2S) timescales in the mid-latitudes, especially in Europe. The application of machine learning for S2S forecasting has been maturing in recent years. Here we apply the state-of-the-art to predict European temperature at 1 to 3 months lead-time. First, we cluster EU temperature reanalysis (ERA-5) data to find domains that show a similar variability. This step increases the likelihood that a physical mechanism is driving this common variability within a clustered domain. By targeting the spatial mean of these domains, we increase the signal-to-noise ratio and thus help the ML model to learn meaningful relationships. In a second step, we apply a response-guided dimensionality reduction method to find features in 3 different climate variables; surface level soil moisture, sea surface temperature, and ocean salinity. These 3 variables are chosen because they influence atmospheric and surface weather and they have an enhanced memory (higher thermal inertia, slow mixing) compared to the more chaotic atmospheric variables. Finally, we apply both regularized regressions and tree-based models to predict continuous and tercile forecasts. Using a one-step-ahead cross-validation and a strict splitting of training and test data to prevent information leakage, we emulate how the prediction system would have performed in an operational-like setting. Across the European domain, we observe - on average - a doubling in the out-of-sample correlation coefficient compared to the best performing numerical weather prediction, which is the seasonal forecast system (SEAS-5) of the ECMWF. With information about the features and by focusing on 'high-signal' forecasts, we identify windows of opportunity, leading to a further improvement and higher confidence in certain temperature forecasts. Hence, by combining expert knowledge with the strengths of machine learning techniques, we show that European month-ahead temperature forecasts are substantially more predictable than suggested by the numerical model SEAS-5.