Beyond skill scores: exploring sub-seasonal forecast value through a case study of French month-ahead energy prediction.

Joshua Dorrington[1], Isla Finney[2], Tim Palmer[1], and Antje Weisheimer[1,3]

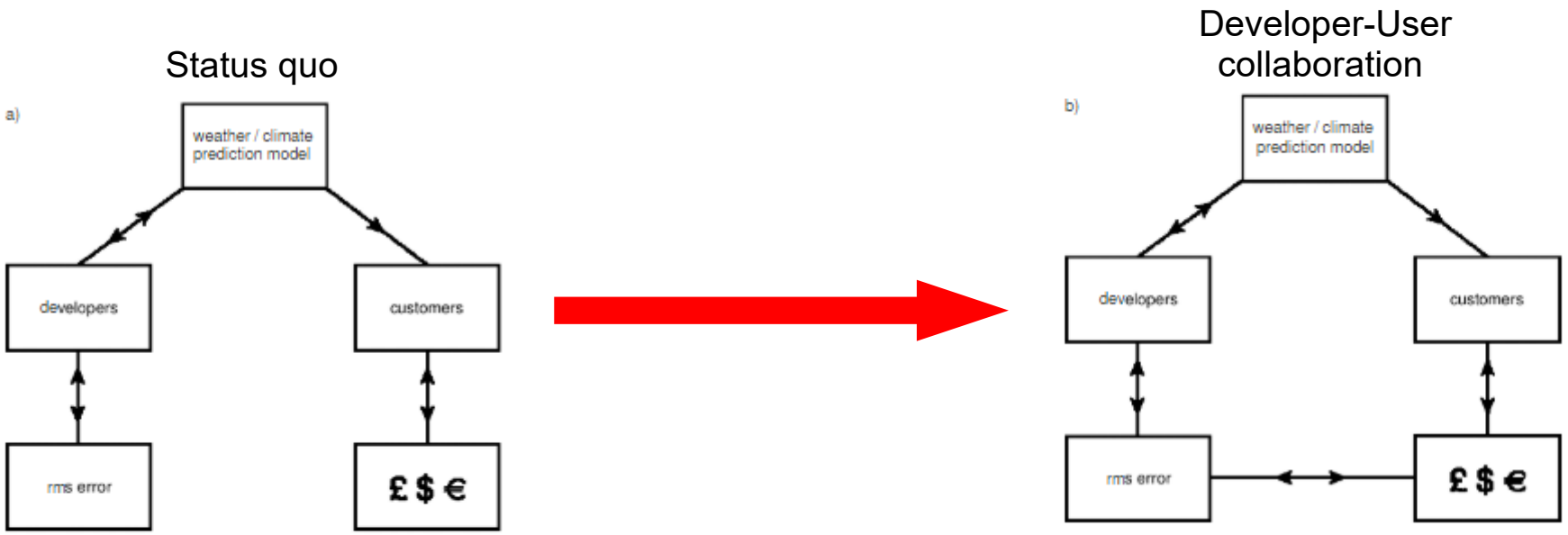
Under review in QJRMS, arXiv preprint at https://arxiv.org/abs/2002.01728

1 Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom

2 Lake Street Consulting Ltd

3 European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

What makes a forecast 'good'?



- Weather forecasts are made for the benefit of end-users, who will use those predictions to make decisions. For users, the value of a forecast is in the average money saved/ damages avoided/ lives saved. Value will be specific to meteorological variable, timescale, and sharpness/reliability for probabilistic forecasts.
- On S2S timescales, for which operational systems are relatively new, it is not always clear who is able to use these forecasts, and what the details of their use cases are. So how can we assess the value of S2S forecasts?
- Model developers use convenient, conceptually clear, and abstracted scores: correlation, RMSE, etc. These are obviously important, but are we missing something?
- Forecasting would benefit from real collaboration between model verification and end-user communities. Identifying the right goalposts makes it easier to provide real world benefits.

Schematic adapted from T.N. Palmer,
The economic value of ensemble forecasts as a tool for risk assessment: From days to decades (2002) with permission from the author

Our approach

- Is this a real concern? Do we actually need to go through the complex process of assessing user needs in detail? Or can we just use simple skill-scores, treating them as a reliable proxy of value to end users?
- We address this by taking a specific example of an end-user application for S2S forecasts; month ahead prediction of French energy demand. We develop a simplified model of this forecast use case based on surface temperature (T2m) with the aid of real energy demand and energy price data, and calculate the value of subseasonal hindcasts (in euros/MWh) within this framework.
- We derive a cost-loss ratio for this use case and use it to calculate the potential economic value (PEV) of surface temperature forecasts as is sometimes done in the subseasonal literature.
- We look at the skill of T2m using common meteorological scores as is commonly done in model development.
- Finally, we compare the conclusions for forecast value each of these 3 approaches provides, and highlight key differences.

Data

- We use ERA5 reanalysis and hindcast data from the EMC GEFS subseasonal system run as part of the SubX project, the ECMWF 46 day extended range system and the SEAS5 seasonal system, covering the period 1999-2018.

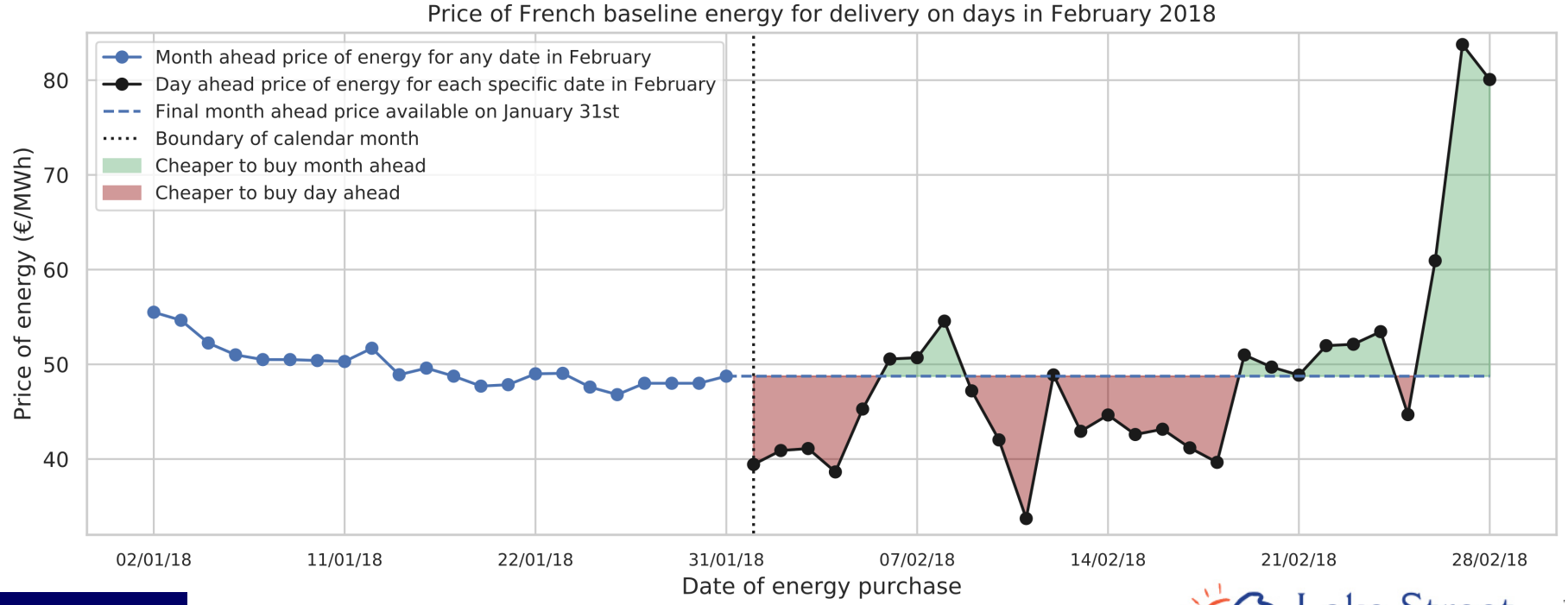
Name	Originating Centre	Forecast Period Used	Initialisation Frequency	No. of Annual/DJF Initialisations	Time Range of Initialisation Dates	Ensemble Size
EC45	ECMWF	46 days	2/week	2146/734	03/01/1999 - 30/05/2018	11
SEAS5	ECMWF	46 days	1/month	219/78	01/01/1999 - 01/08/2018	25
GEFS	EMC (SUBX)	35 days	1/week	1017/336	01/06/1999 - 30/05/2018	11

- French price[1] and demand data[2] for the period 2010-2018 was used for the end-user example.

1 http://clients.rte-france.com/lang/an/visiteurs/vie/vie_stats_conso_inst.jsp
2 http://www.eex.com/en/products/power-derivatives-market/power-futures/power-futures-products

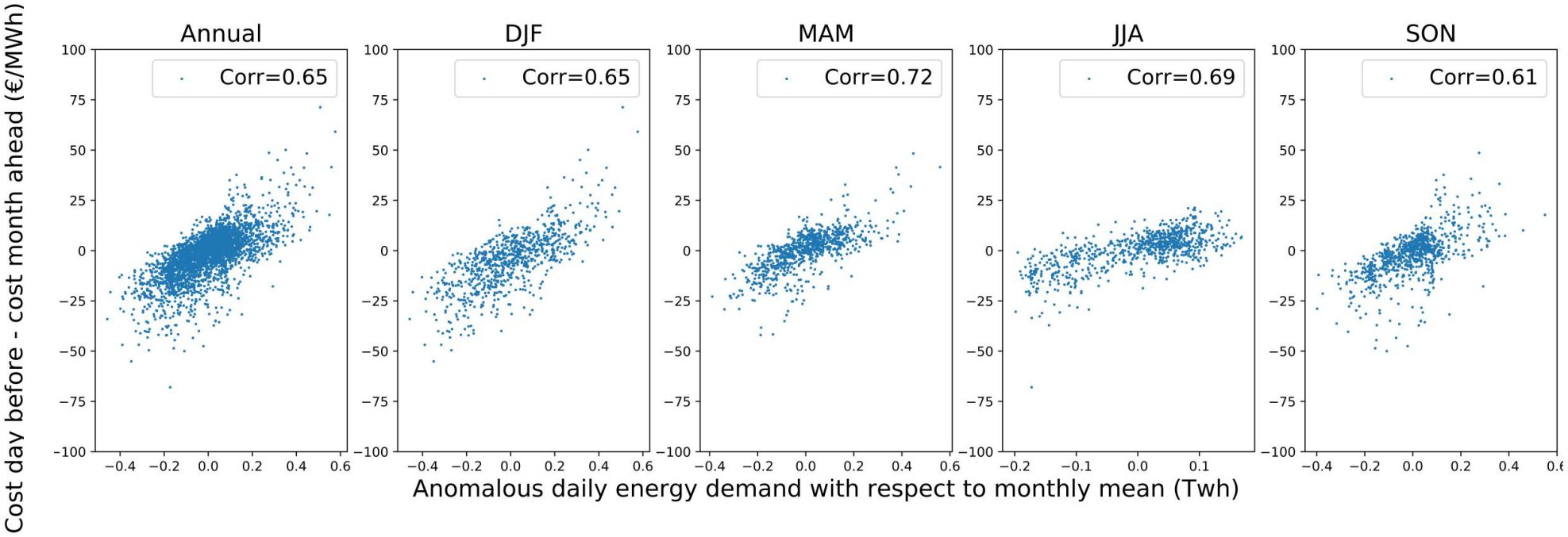
Month ahead French energy demand

- A specific example picked due to author experience with the industry.
- French energy for future days is traded ahead of time on a market. Within the current month, most energy is bought and sold in daily units, and the price of these will vary wildly in response to expected changes in demand. For the next month however, energy is traded with a single averaged price for all days, making it much less volatile.
- A common strategy used in this industry by traders or energy providers is to decide whether to buy energy at the end of the calendar month (the last chance to get the stable, roughly climatological price) or to buy it the day before delivery.
- Forecasts will be used to decide if a given day in the following month is likely to have above average energy prices (so buy month ahead), or below average (so buy day ahead).



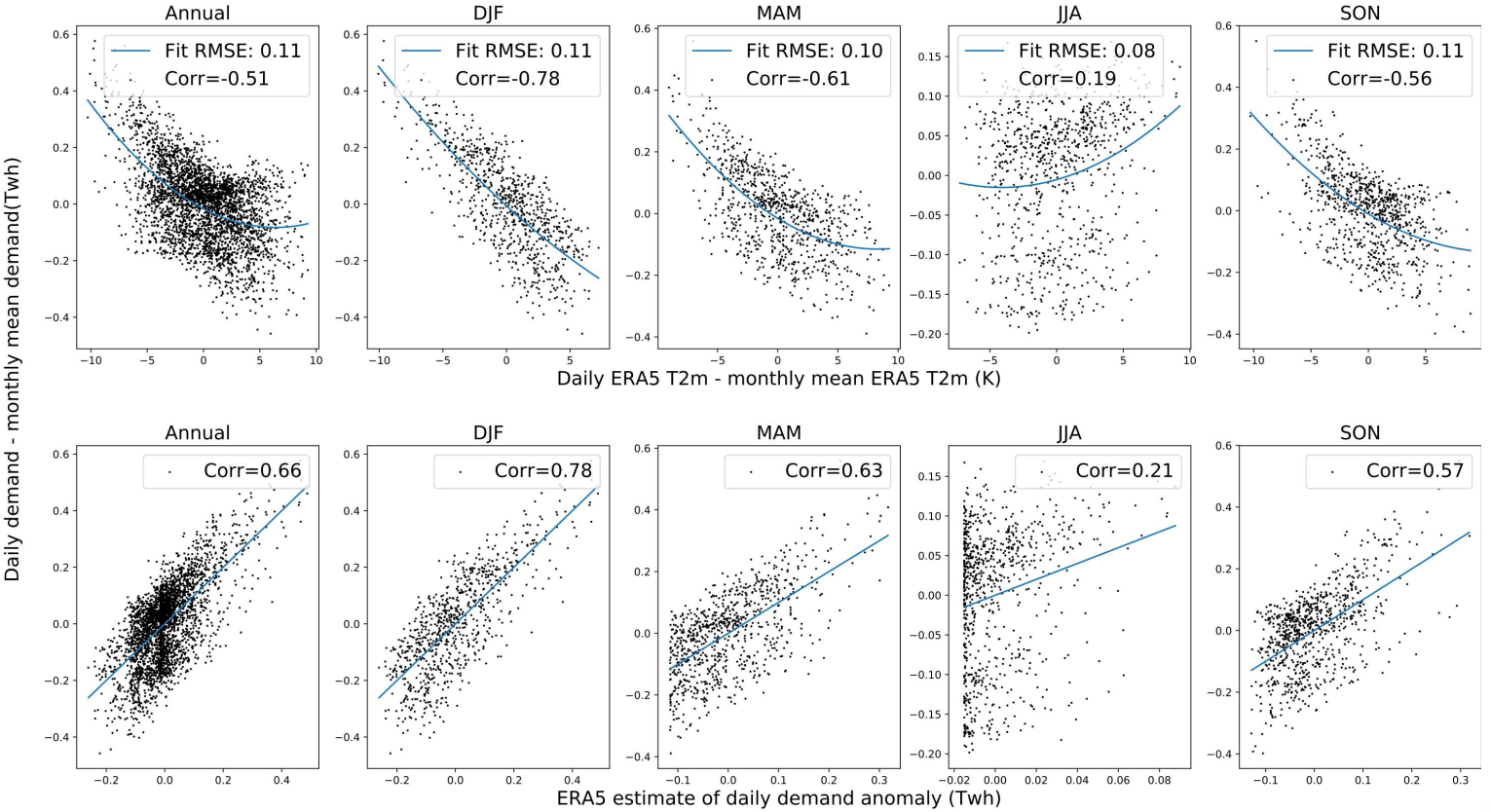
Demand is a good predictor of price

- We look at the anomaly of daily French energy demand with respect to the mean value for the calendar month and find it predicts a reasonable amount of the difference in price between the day ahead and month ahead prices of energy.
- Macroeconomic factors such as global price of fossil fuels will also heavily affect the price independent of demand.



Temperature is a good predictor of demand

- We can predict a large part of the meteorological component of energy demand, using only the surface temperature (T2m). We look at daily T2m averaged over land in the region [5W-8E,42N-51N].
- Again we subtract the mean temperature for the calendar month, and find especially during DJF we can explain a sizeable fraction of the demand variability. Including other variables such as relative humidity, surface winds etc. would likely lead to a better fit.
- Because of the simplicity of this fitted model, any forecast value we show will be only a lower bound on what is potentially achievable by a real user in the energy sector.



Details of the trading strategy

- The decision on whether to buy energy for a given date a month ahead or not, must be made at the end of the previous month. For the first week of a month therefore, we can use a week 1 forecast to make our decisions, while for the fourth week we must use a week 4 forecast.
- For each day in the record (2010-2018) we take the most recent available forecast from each forecasting system from the previous month. We use that to make an estimate of anomalous demand for each day.
- If the demand anomaly is greater than a threshold d , we buy energy at the month-ahead price.
- Otherwise we buy energy at the day-ahead price.
- Two types of user exist, those who are always buying a certain set amount of energy (such as an energy trader might), and those who are buying a set fraction of the daily energy (such as energy provider might).
- The value of the forecasts to the set-amount user, averaged over the 9 year period is:

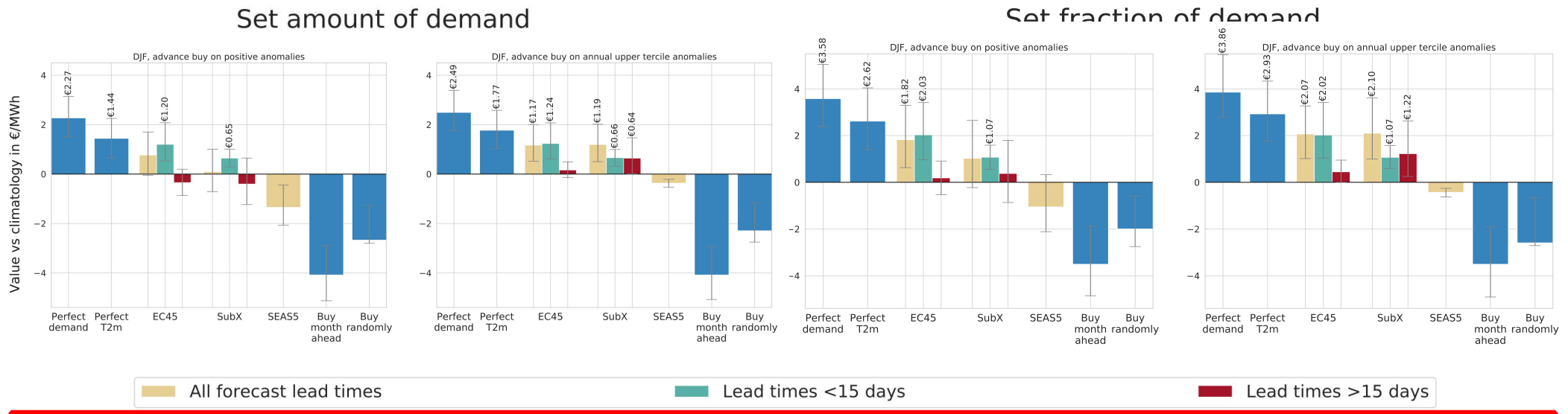
$$C_{\text{set amount}} = \frac{1}{T}\sum_{t}^{T}\left(H[P_D(t) - d] \cdot p_{\text{month}} + H[d - P_D(t)] \cdot p_{\text{day}}\right)$$

while for the set-fraction user, it is:

$$C_{\text{set fraction}} = \frac{1}{T}\sum_{t}^{T}\bar{D}_t \cdot \left(H[P_D(t) - d] \cdot p_{\text{month}} + H[d - P_D(t)] \cdot p_{\text{day}}\right)$$

Where p_{month} is the month-ahead price, p_{day}, the day-ahead price, and H the Heaviside step function.

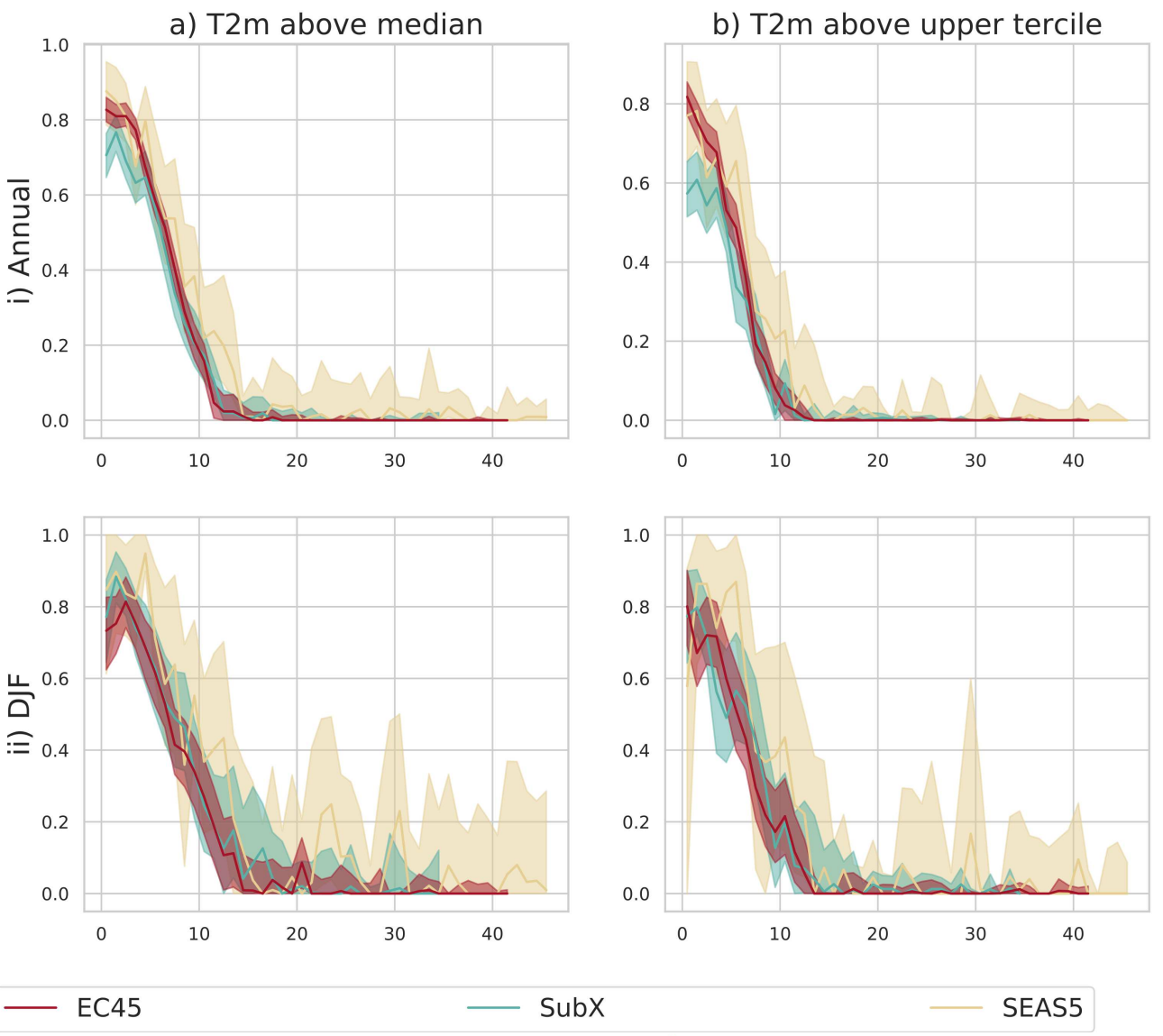
So how much is a forecast worth?



- Due to the simplicity of our model, most of the value we find is limited to DJF so we focus on that here.
- We look at the value of our forecasts relative to the best possible climatological strategy (always buying on the day ahead price).
- As well as looking at the value of the forecast systems holistically we consider two counterfactuals: what if our forecast only ran for 15 days? What if we had no forecasts with lead times less than 15 days? Thus we separate out the benefit of the extended range (week 3+).
- SEAS5 has no value! Why? Because its initialised on 2nd of the month, so not good for month ahead prediction (min lead time of 26 days).
- When using a threshold *d* of upper tercile anomalies the GEFS SubX system shows statistically significant value using only the day 15+ forecasts (~40% of the saving a perfect temperature forecast would provide).
- Using only 2 week forecasts, EC45 has higher value than SubX because of its higher initialisation frequency, but once the extended range forecasts are included, they become equally valuable (for upper tercile anomalies)

PEV – An abstracted application

- Potential economic value, or PEV, is an abstraction of a binary decision making process that a user might make, paying a cost *C* to avoid a loss *L* if a extreme event occurs. It is relatively frequently used in the meteorological literature. How realistic is it?
- One issue with PEV is the ratio *C/L* is very important for determining how much value a user can gain from a forecast. Normally no assumption about *C/L* is made. For our case we can calculate it directly from our user example.
- *C* : The average amount by which month ahead price exceeds day ahead price
- *L*: The average amount by which (day ahead price | high demand) exceeds (month ahead price | high demand)
- We find values of *C/L* ranging from 0.6-0.7 are appropriate for our example, much higher than typically assumed!
- We find PEV=0 (no value) by day 12 at the latest

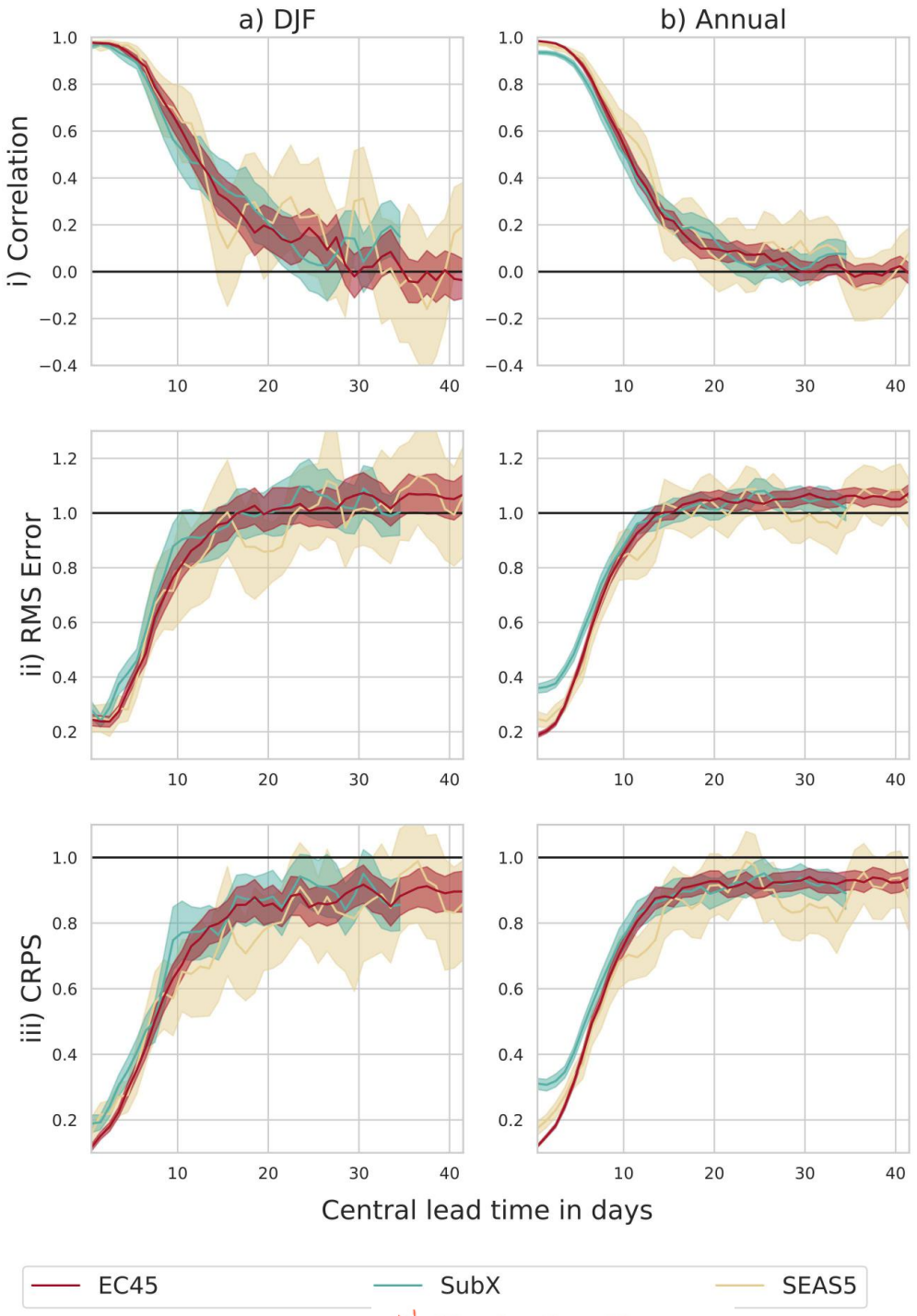


UNIVERSITY OF OXFORD

Lake Street CONSULTING working with the weather

Conventional meteorological scores

- We look at correlation skill, root mean square error (RMSE) and continuous rank probability score (CRPS) as these are common, general purpose skill scores with desirable properties (i.e. they are proper) for meteorological applications.
- Correlation skill remains signicantly above zero out to days 22 and 27 for SubX and EC45 respectively, supporting the low but nonzero value we saw in week 3+ in our example.
- RMSE and CRPS are more pessimistic, DJF error has saturated in both cases by day 15, suggesting no extended range skill.



Conclusions

- We have compared 3 methods of evaluating the value of sub-seasonal daily French T2m forecasts, using scores progressively more abstracted from application.
- We find value in one model's week 3+ forecasts using a realistic energy trading framework (for DJF), even with a relatively poor and untuned decision strategy, and only a univariate predictor.
- Correlation skill is also significantly non-zero at week 3+ but other scores including the quasi-realistic PEV don't show any signs of value in extended range daily T2m. Our hypothetical energy provider would probably have never tried to use subseasonal forecasts if he had seen these skill scores!
- Beyond our simple example, we hope to emphasise that forecasts do not exist in a vacuum, they are only as good or bad as they are useful to users and that does not always match up with our traditional ways of assessing skill. Additionally even seemingly minor operational points, such as date of forecast initialisation can make a big difference to users; if SEAS5 was initialised on the 29th of every month instead of the 2nd it would likely have been much more valuable for our application!
- Model developers should try and develop a small number of simplified applications such as we have here, in collaboration with users, **and use them routinely as target scores when verifying forecasts**.