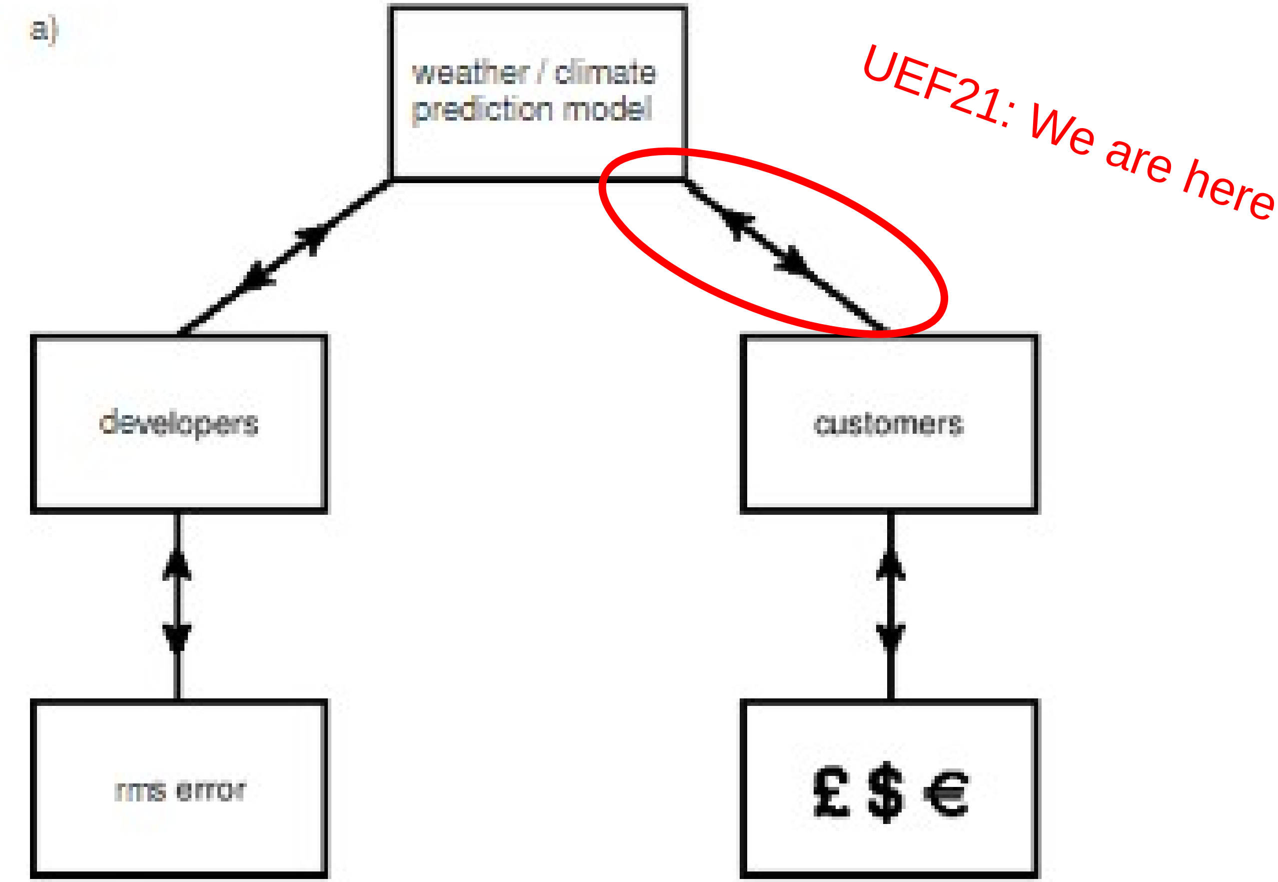
Beyond skill scores: Why we should include end-users in the model validation process.

Josh Dorrington

Based on *‘Beyond skill scores: exploring sub-seasonal forecast value through a case-study of French month-ahead energy prediction’*, Dorrington, Finney, Palmer & Weisheimer, **2020**, QJRMS, doi.org/10.1002/qj.3863

What makes a forecast 'good'?

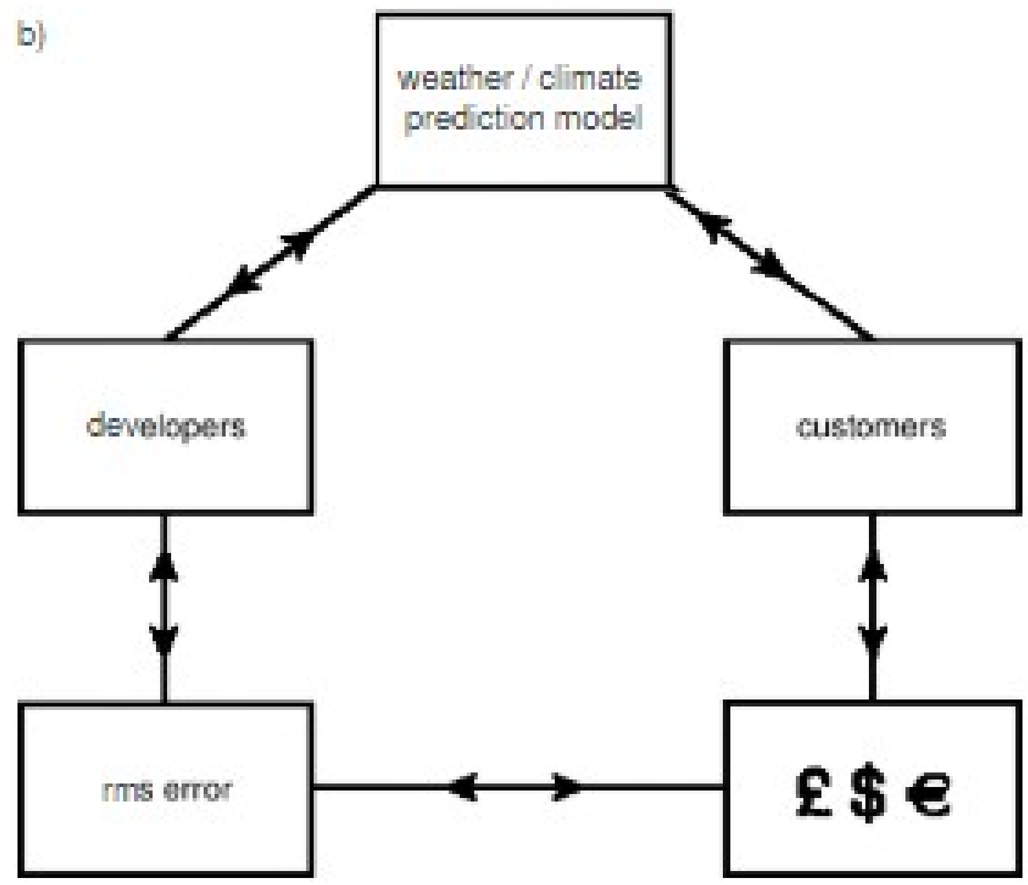
Academics, operational meteorologists, model developers and end-users are all part of the same project – to make better decisions based on weather knowledge.
But how well do we really communicate with each other?



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

What makes a forecast ‘good’?

Can we make it easier to move between meteorological scores and applied forecast value?



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

Using subseasonal forecasts

- Midlatitude subseasonal skill is low, and probabilistic in nature

- Often assessed using large spatial and temporal mean quantities
- There is a large technical overhead required to identify if subseasonal forecasts can help for a specific use case
- Useful predictions are probably being wasted!

Does this really matter?

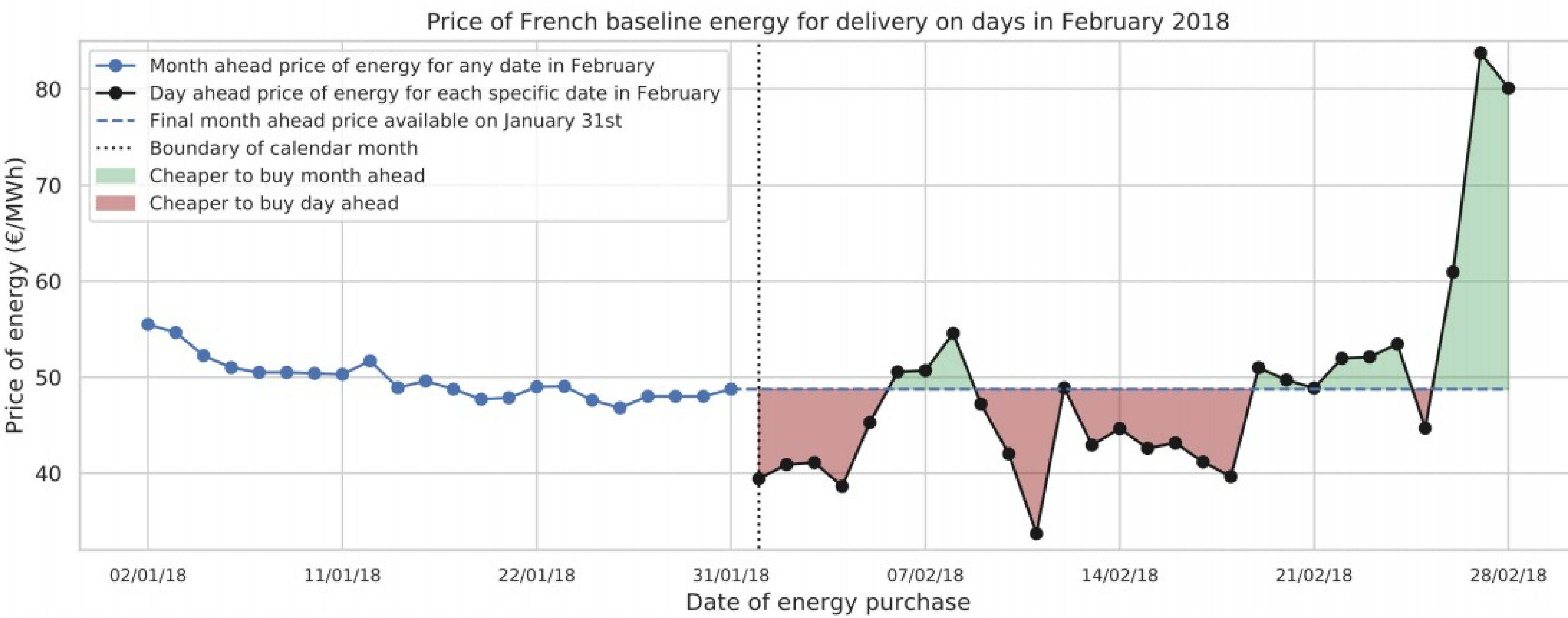
- Do we actually need to do this?
- Integrating users into development could be time-consuming, would need specialist skills, domain specific datasets, and is less conceptually simple
- We have a range of target variables and neat well-behaved skill scores, surely our conclusions will generalise well?
- Let's look at a specific end user case study and see how important this is for an actual example

Trading French Energy

A use case relevant for energy providers and traders.

Based on a (slightly idealised) real-world weather-relevant problem:
If I need electricity on a particular day, should I buy it the month before I need it, or the day before?

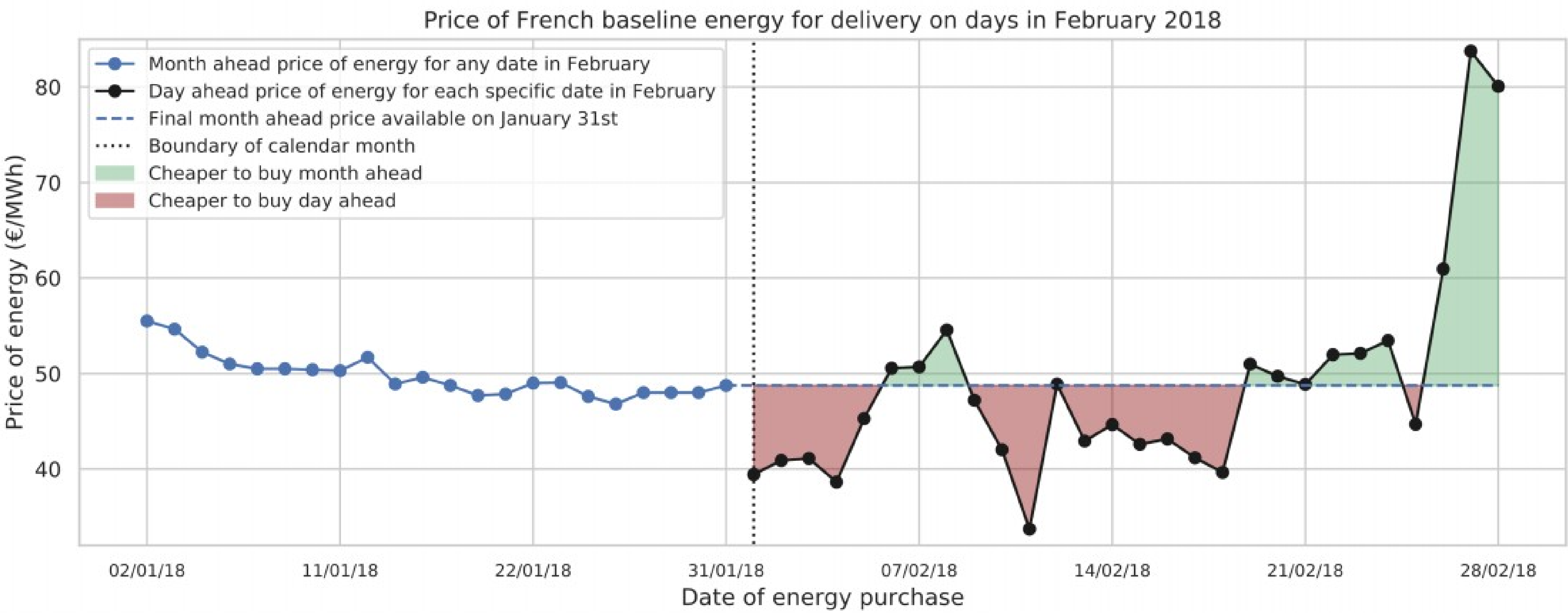
Monthly traded energy measured in TWh → €10’s millions



Trading French Energy

Non-meteorological quirks

- High spatio-temporal resolution – daily, national averages
- Decisions have to be made by specific calendar dates
- The application isn't based on a single lead time, but mixes days 1-45



Procedure

- Build a minimal model of the application
- Test it with hindcast data to get a lower-bound on actual predictability
- Compare to equivalent purely meteorological scores; do they imply qualitatively similar forecast value for this application?

Data

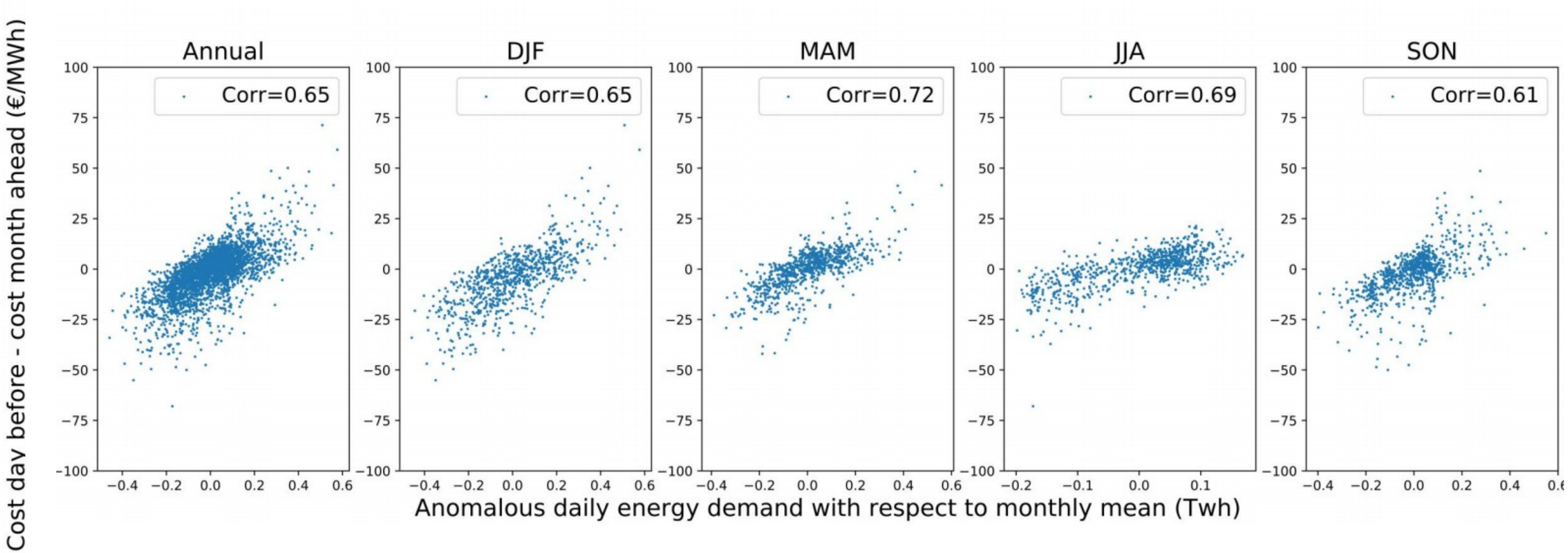
- Used 12:00 T2m, Area averaged over land in [5W-8E,42N-51N] as a univariate predictor
- Used publicly available French price[1] and demand data[2] for the period 2010-2018
- ERA5 reanalysis was used to find T2m → demand relationship
- Evaluated 2 Subseasonal forecasts, EC45 and GEFS and 1 Seasonal forecast, SEAS5
- All forecasts were calibrated using quantile mapping

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

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

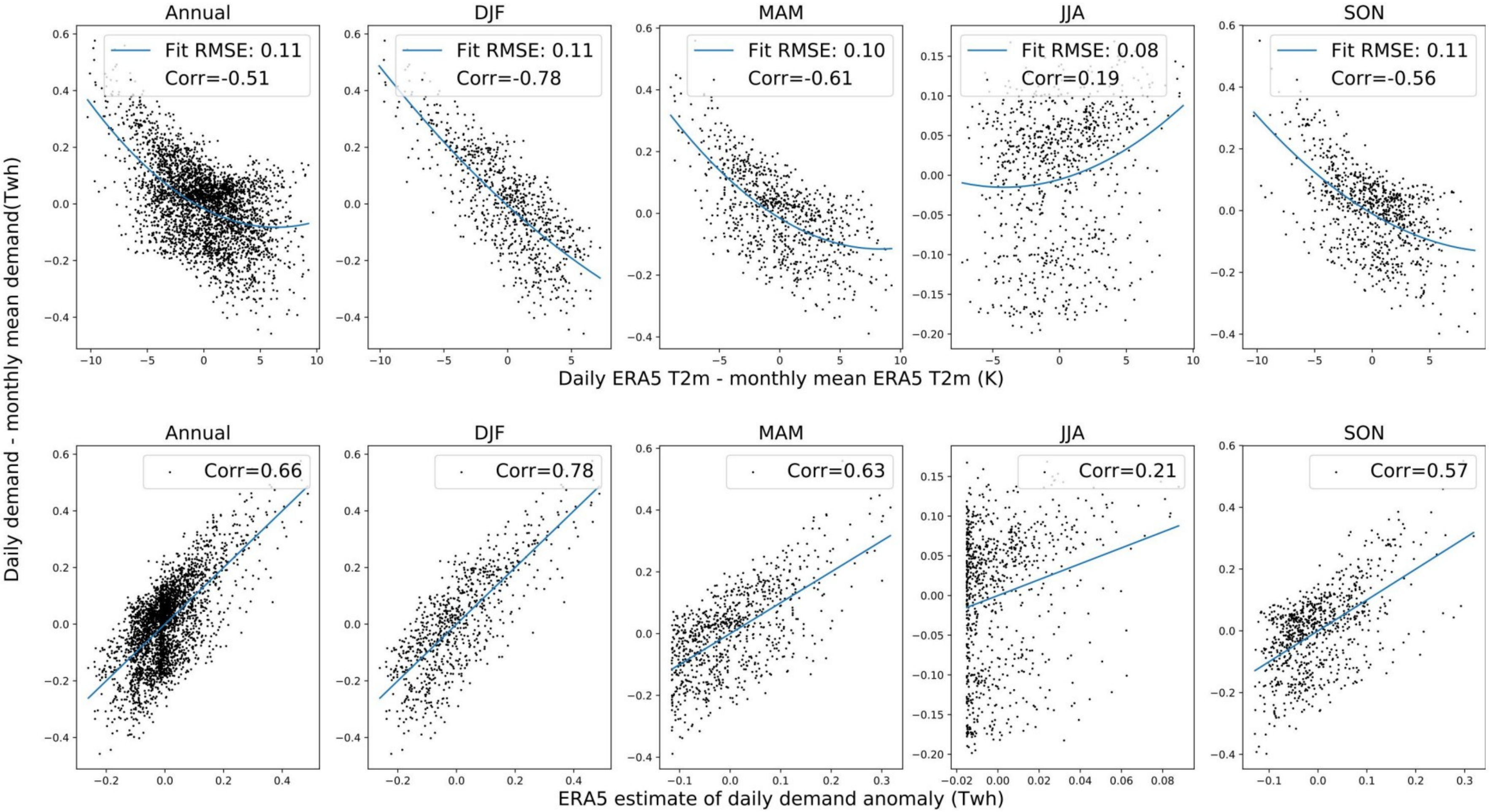
Demand is a good predictor of price

- 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

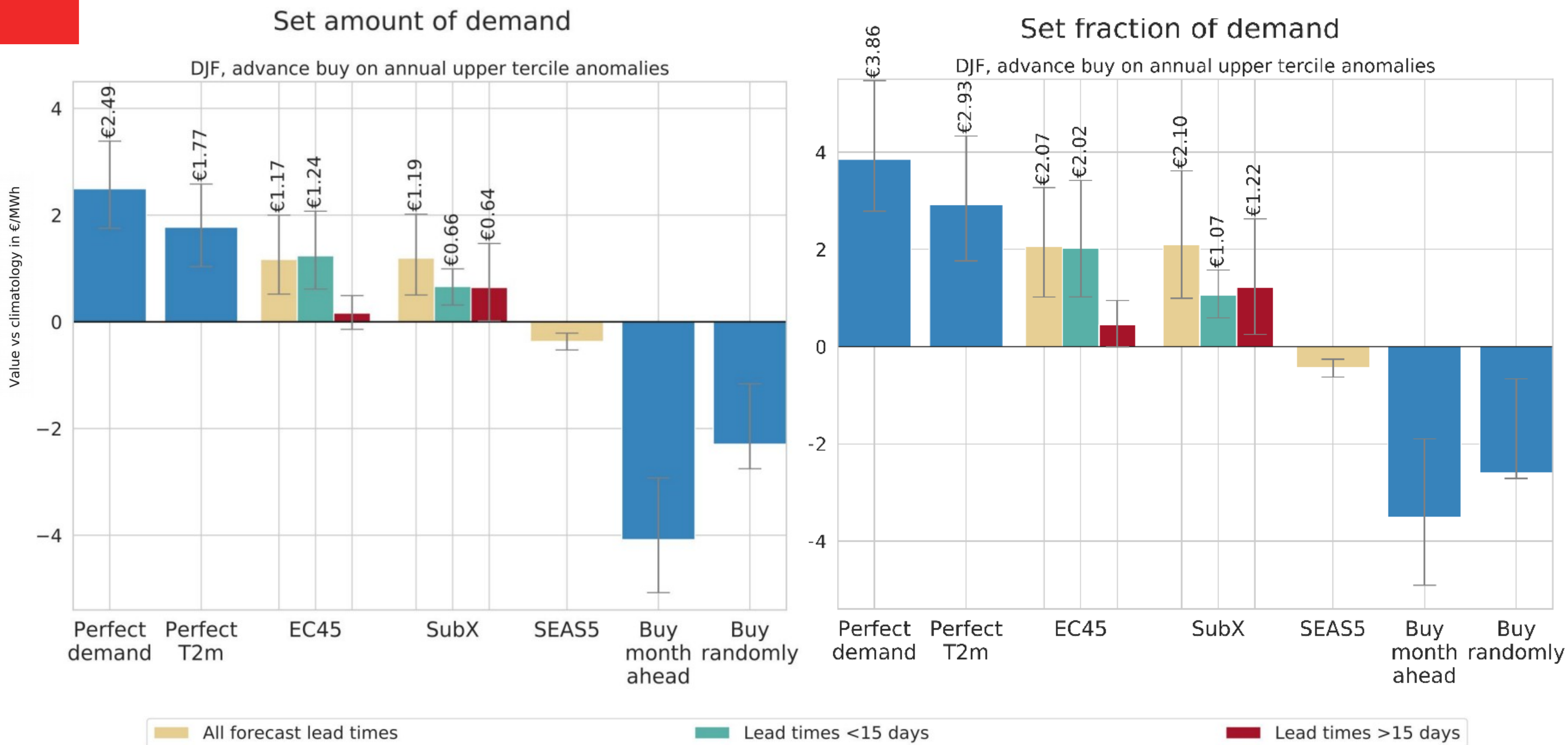
- We look at daily anomalies with respect to monthly averages
- Find quadratic fits of T2m make reasonable predictors in most seasons
- Deliberately keeping it conceptually simple – truly realisable skill will almost certainly be higher



UNIVERSITY OF OXFORD

Lake Street CONSULTING working with the weather

So how much is a forecast worth?



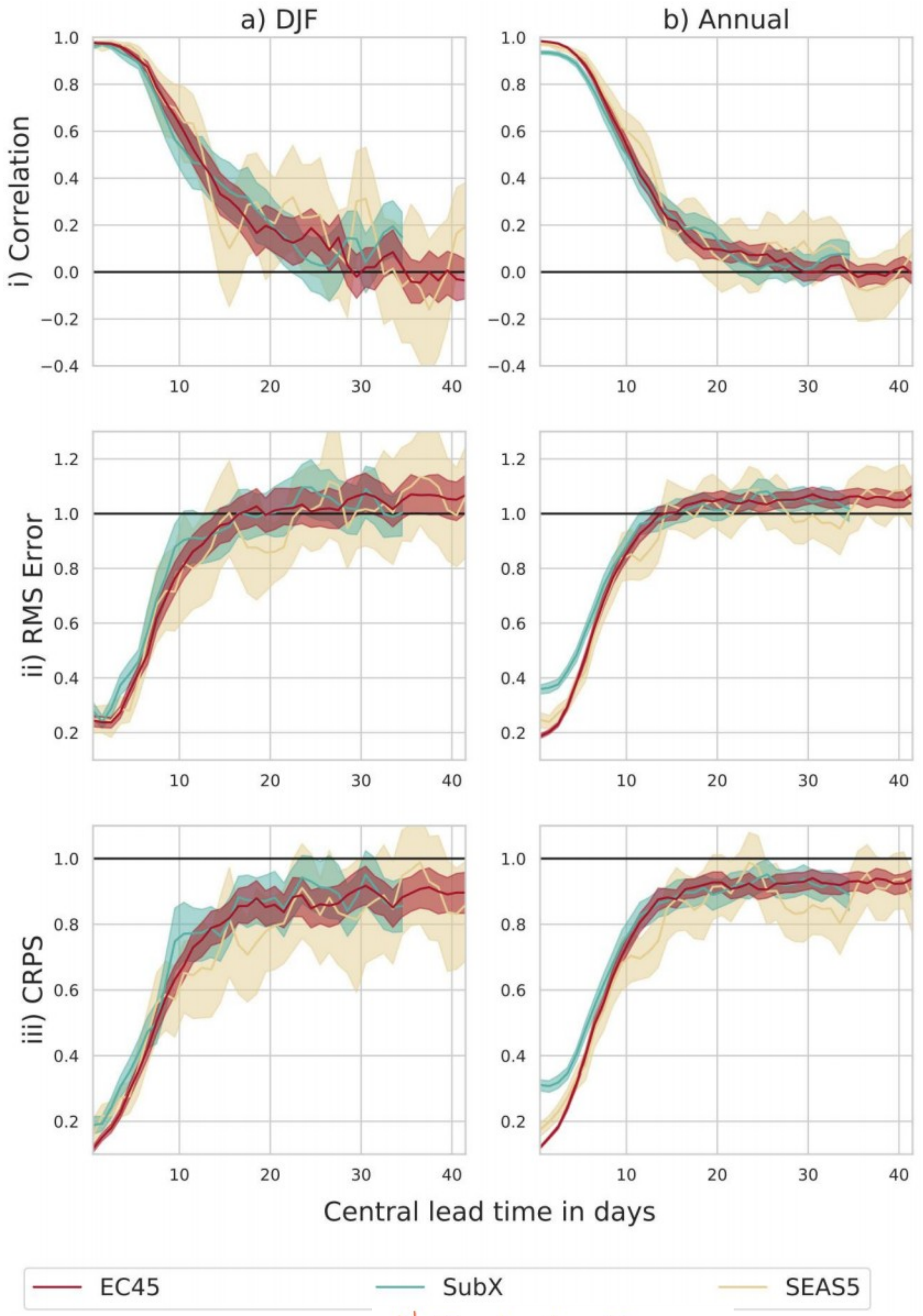
- Value shown relative to climatological action
- Extending into the subseasonal range can make up for less frequent initialisations
- For GEFS 15+ day forecasts had valuable skill on their own (35-40% of perfect T2m forecast)

Conclusion of case study

- Even this minimal model found a potential benefit from subseasonal forecast data, at least in DJF
- Someone interested in this use case should consider more detailed in-house analysis
- So do pure meteorological skill scores tell the same story?

Conventional meteorological scores

- We look at three generic, commonly used scores: anomaly correlation skill, root mean square error (RMSE) and continuous rank probability score (CRPS).
-
- Evaluated daily T2m, averaged over France, as for the case study, using the same hindcast data.
- Subseasonal models show correlation skill signicantly above zero out to days 22 and 27 for SubX and EC45 respectively, suggesting possible skill
- RMSE and CRPS are more pessimistic, DJF error has saturated in both cases by day 15, suggesting no extended range skill.



UNIVERSITY OF OXFORD

Lake Street CONSULTING working with the weather

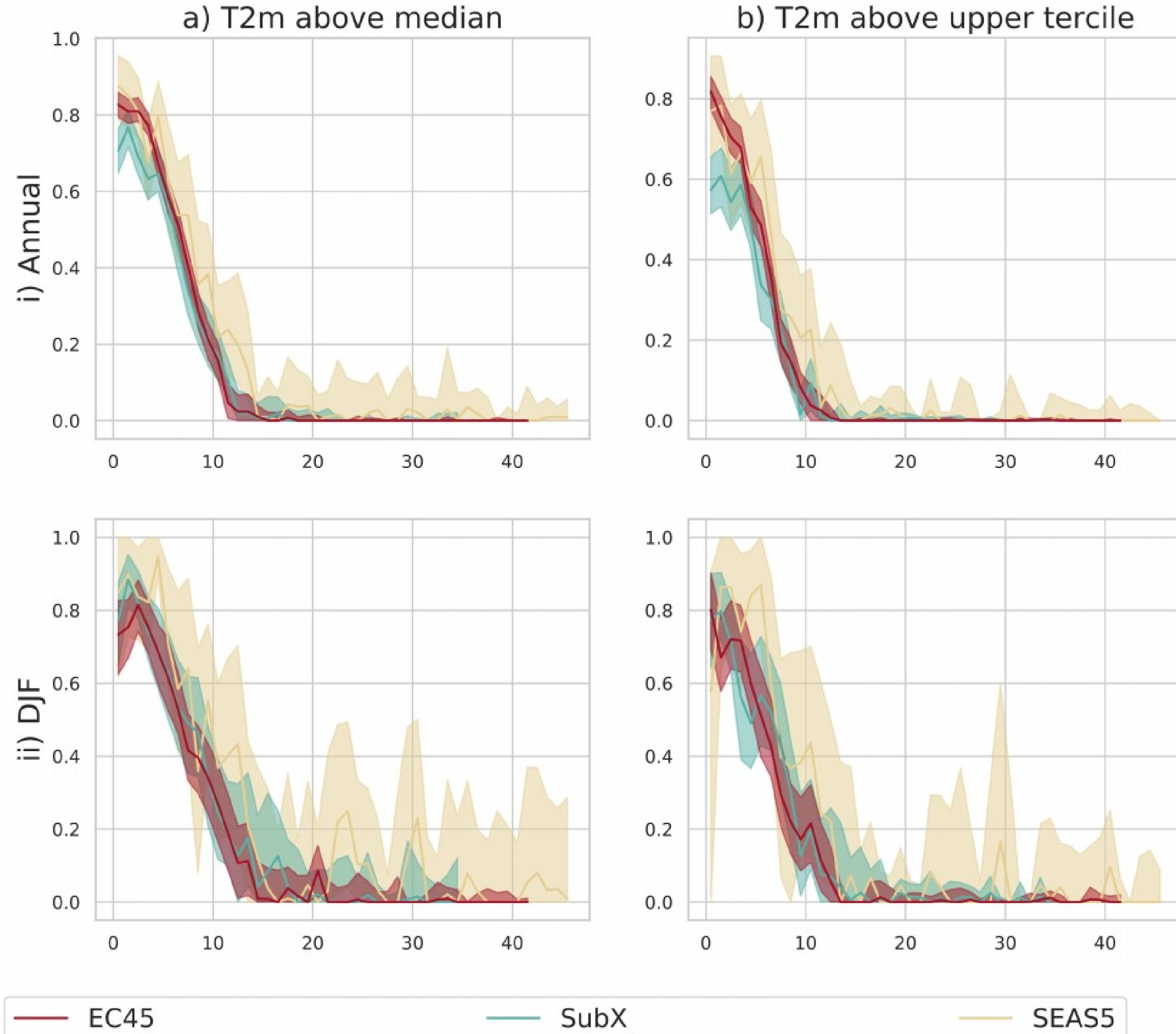
Potential Economic Value – a middle ground?

- A commonly used abstraction of forecast value to a user.
- An action with cost *C* is be taken to avoid a loss *L*, if an event *e* is predicted with a probability >p_{thresh}
- The PEV is defined from the confusion matrix:

$$
M^{(\text{conf})} = \begin{array}{c} \\ \text{Event} \\ \text{No Event} \end{array} \begin{array}{c} \begin{array}{cc} \text{Predicted} & \text{Not Predicted} \end{array} \\ \begin{pmatrix} P_{\text{True Positive}} & P_{\text{False Negative}} \\ P_{\text{False Positive}} & P_{\text{True Negative}} \end{pmatrix} \end{array}
$$

- From our price data we find a cost-loss ratio ~0.65, much higher than in idealised examples or extreme event applications
- Knowing something about users' cost-loss ratios makes a huge difference!

Potential Economic Value



- Value has decayed by days 10-12
- No evidence at all of extended-range skill!
- The assumption of constant *C*, and *L* is not well justified

UNIVERSITY OF OXFORD

Lake Street CONSULTING working with the weather

How useful are forecasts of DJF French daily T2m at week 3+?

“At least for one model, including extended range forecasts adds 30+% the value of a perfect forecast”

“There is no value for your application.”

“Most metrics show no skill, although there is marginal correlation skill. So maybe?”

So what’s the point?

- It’s hard for users to know if S2S forecasts might be valuable for them – reducing their societal benefit
- There are non-meteorological factors that impact forecast value that we might not tend to think about – initialisation date and day of the week for example.
- Applications are full of thresholds, cutoffs, and nonlinearities: this can make purely meteorological scores misleading
- Worst case hypothetical – a model update is introduced that improves a benchmark score but degrades end-user value

What could we do about it?

- A suggestion: **Add simplified end-user case studies to the forecast skill card**
- Work with users to build a catalogue of simplified applications, like in the example we've used here
- Run them routinely as part of the model validation process
- The skillcard of 2025?: