S2S4E
Climate Services for Clean Energy

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

A new approach to subseasonal multi-model forecasting: Online prediction with expert advice

Paula LM Gonzalez[1] – David J Brayshaw[1] – Florian Ziel[2]

[1] NCAS – Climate / University of Reading, United Kingdom

[2] University of Duisburg-Essen, Environmental Economics, Germany

MOTIVATION

- Given the abundance of sub-seasonal and seasonal forecasting systems available, it is quite standard to try to gain skill through a 'smart' combination of their predictions.

- Standard multi-model combination techniques assign 'static' weights to the different predictions (most typically, uniform or skill-based weights). This is a limitation due to changing skill of the forecasting systems (e.g., seasonal, model updates, state dependence).

Online prediction with expert aggregation:

- a family of machine learning algorithms that allow to **combine predictors or 'experts' with evolving weights** by progressively minimizing a loss function (typically, the 'pinball loss').

Advantages of online methods:

→ the multi-model combination or 'mixture' is able to **adjust to preserve skill (minimize loss)** under certain conditions.
→ one can train a **different mixture of the experts for different quantiles of the distribution** and obtain a robust 'forecasting system'.
→ when provided with **inappropriate experts** (e.g., irrelevant or with no skill), the method is able to **discard** them.

DATA AND METHODS

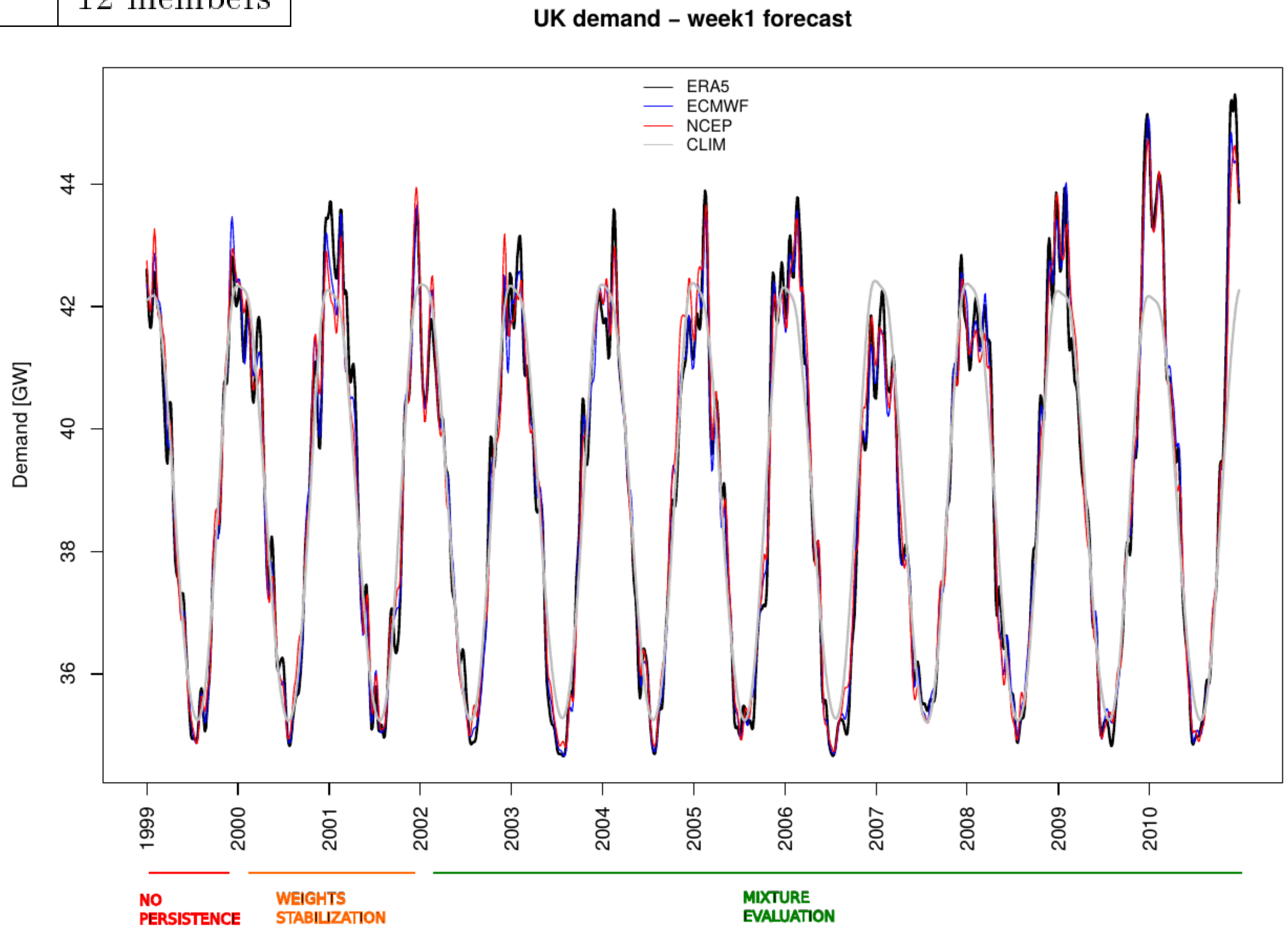
From the S2S hindcast dataset we have used two models in the following setup:

MODEL	RANGE	PERIOD	FREQUENCY	SIZE
ECMWF ENS-extended	0-46 days	past 20 years	2/week	11 members
NCEP CFSv2	0-42 days	1999-2010	daily	4 members
lagged NCEP CFSv2	0-39 days	1999-2010	2/week	12 members

The hindcast output is adjusted through a lead-dependent **mean bias correction and a variance inflation** (Doblas-Reyes et al. 2005)

We use **country-aggregate daily electricity demand** derived from t2m using a weather-dependent demand model developed by Bloomfield et al. (2020, Met. Apps.) and compare it with the corresponding values from **ERA5**.

The common period between the models is 1999-2010 and for the **evaluation of the methodologies** we can consider **2002-2010 (9 years)**.



INTRODUCTION TO METHODS

- The online learning algorithms

-**BOA**: Bernstein online aggregation
-**MLpol**: Polnomial potential aggregation

Were applied in two setups:

- **full**: considering all experts
- **NWP-only**: considering only the experts from the hindcast systems

- Additionally, we include the '**exponentiated gradient**' method as a reference, which is a sequential learning algorithm **previously used in weather and climate → EGA_NWP**

- A different model was trained for each quantile in **Qgrid**=0.05:0.05:0.95

EXPERTS

ENSEMBLE BASED

- QUANTILES of the ensemble distribution:
- q10,q35,q50,q65,q90 (for each S2S ensemble)
- FCST_MX (captures seasonality and range of models)
- FCST_MN

REANALYSIS BASED

- QUANTILES of the climatology (ERA5 1.5 deg – 11yrs as loo): q10,q35,q50,q65,q90
- PERS (persistence of weekly value for forecast days -7 to 0)
- PERS_1yr (persistence of past-year weekly demand)
- SEAS_MX (captures seasonally-varying range of obs)
- SEAS_MN

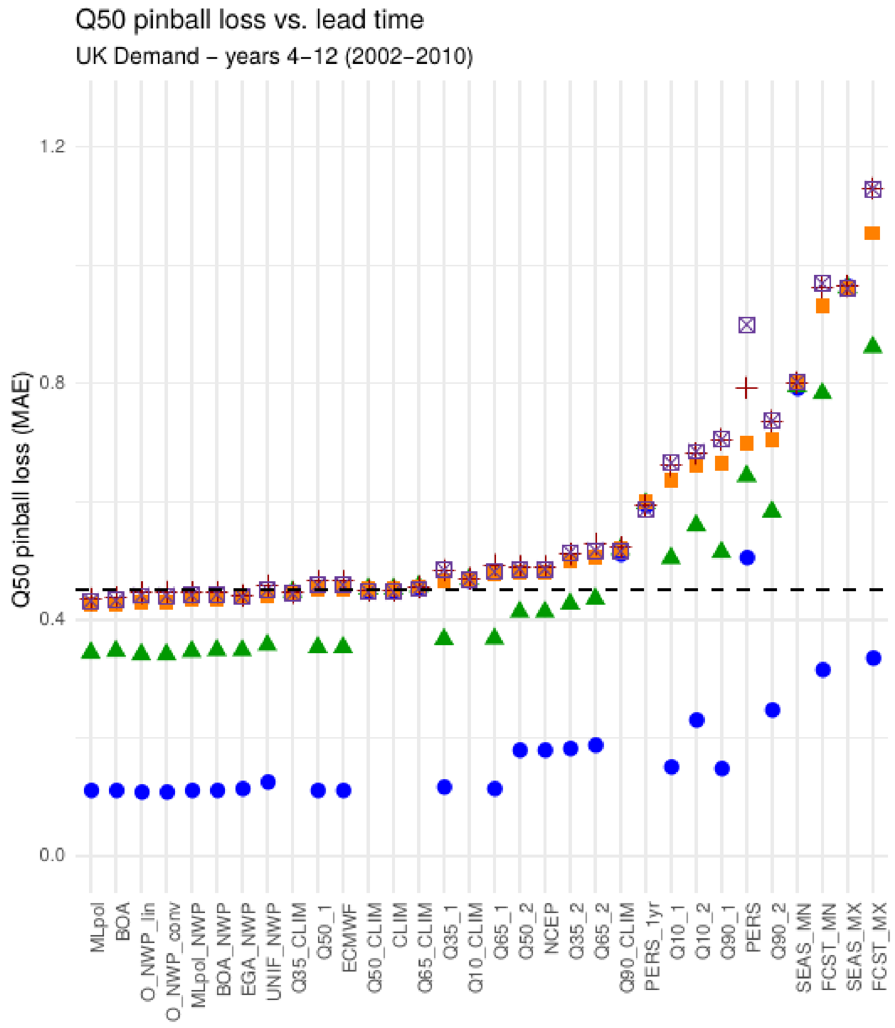
REFERENCE FORECASTS

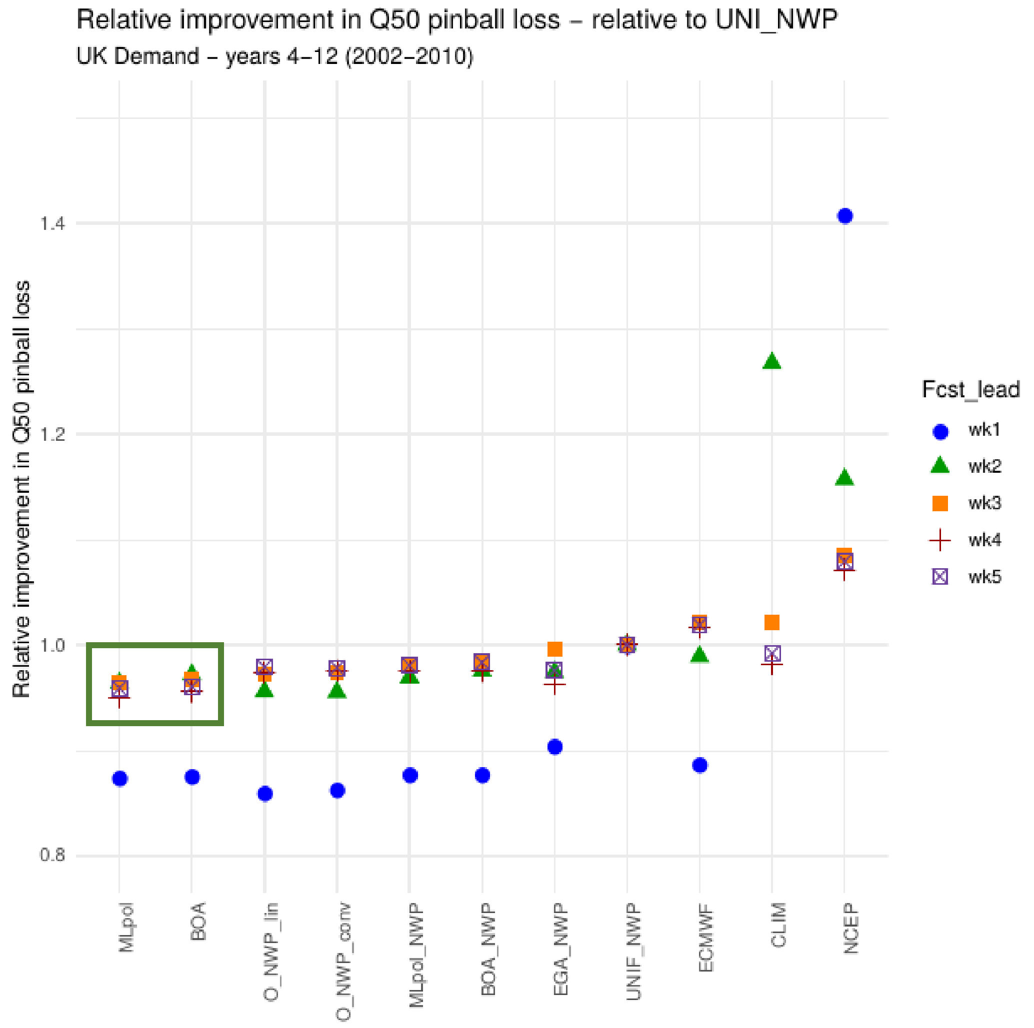
- UNIF_NWP (uniform combination of ECMWF & NCEP)
- CLIMATOLOGY (estimates for full Qgrid from 11yrs loo)
- ORACLE_NWP_linear (optimal mixture of models – full period)
- ORACLE_NWP_convex (requires 0<Wi<1 & sum(Wi)=1)

S2S4E
Climate Services for Clean Energy

RESULTS: DETERMINISTIC SKILL

Q50 pinball loss (Mean absolute error) and the relative improvements w.r.t. the uniform combination



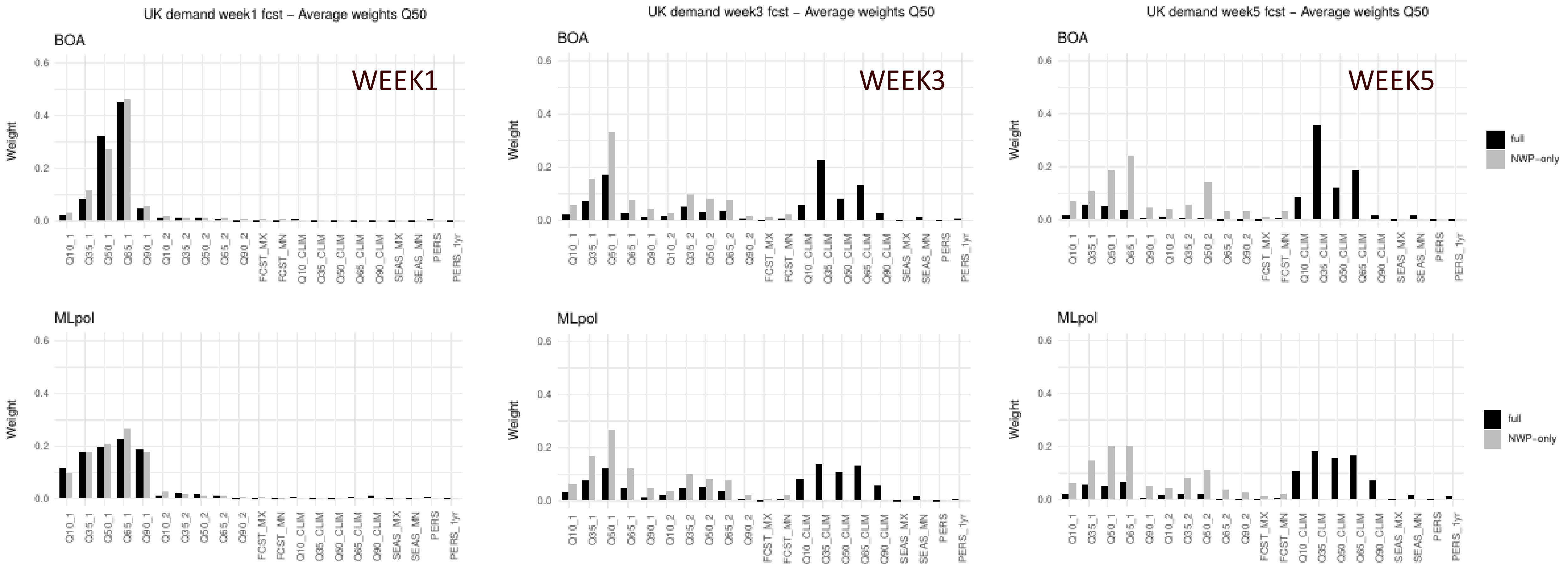


- All the ML combinations and the oracles are more skilful than the uniform combination for every lead
- For week 3, BOA and MLpol show around a 4% increase in skill wrt UNIF_NWP, but EGA_NWP ~0%

S2S4E Climate Services for Clean Energy

RESULTS: DETERMINISTIC SKILL

Composition of the 'mixtures' → time-averaged weights for Q50

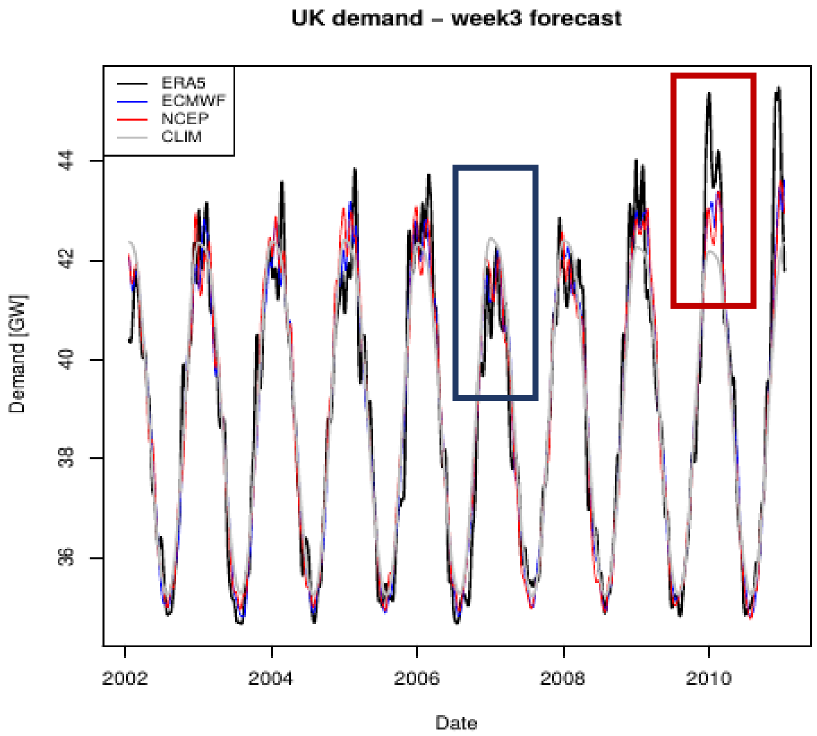


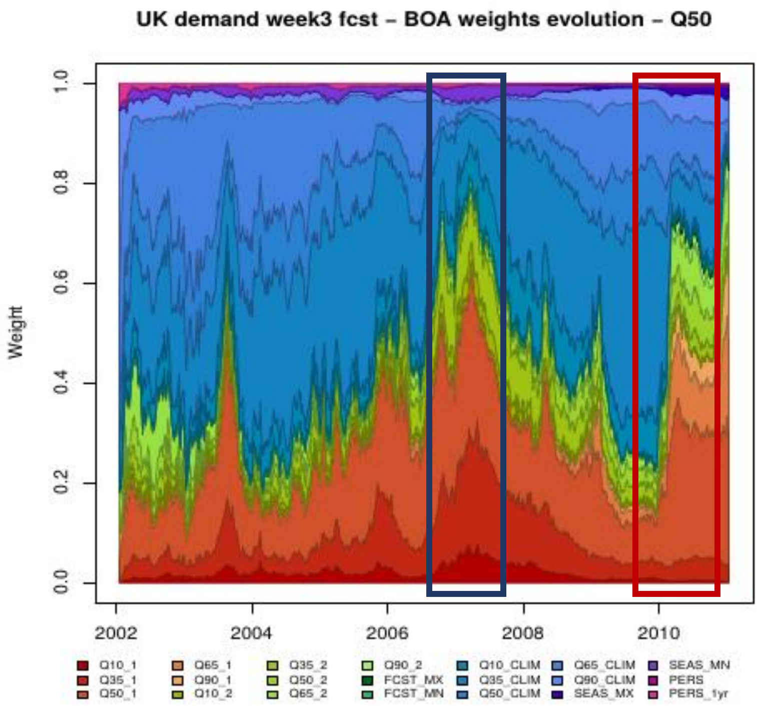
- Initially, ECMWF gets most weights but as lead time evolves, other experts become relevant, mainly CLIM
- In the case of the _NWP combinations, NCEP gets more weight as lead evolves
- Though BOA and MLpol have very similar skills, the composition of the mixtures show differences

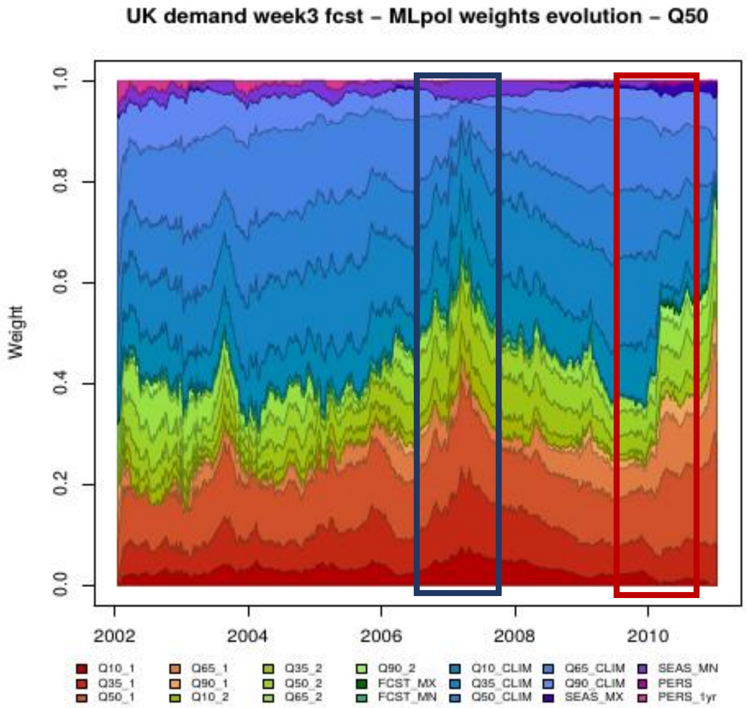
S2S4E Climate Services for Clean Energy

RESULTS: DETERMINISTIC SKILL

Evolution of the Q50 weights: case studies







- Week 3 would on average have high weights on CLIM, but in cases were the forecasts differ from CLIM and get closer to ERA5, CLIM weight drops in favour of ECMWF quantiles.
- In the case of 2006/2007 winter, demand was lower than CLIM but fcsts were larger than ERA5, so the **lower quantiles** of NCEP get the biggest weight increases.
- In the case of 2009/2010 winter, demand was higher than CLIM but forecasts were smaller than ERA5, so the **upper quantiles** of NCEP get the highest weight increases.
- BOA adjusts more quickly than MLpol.

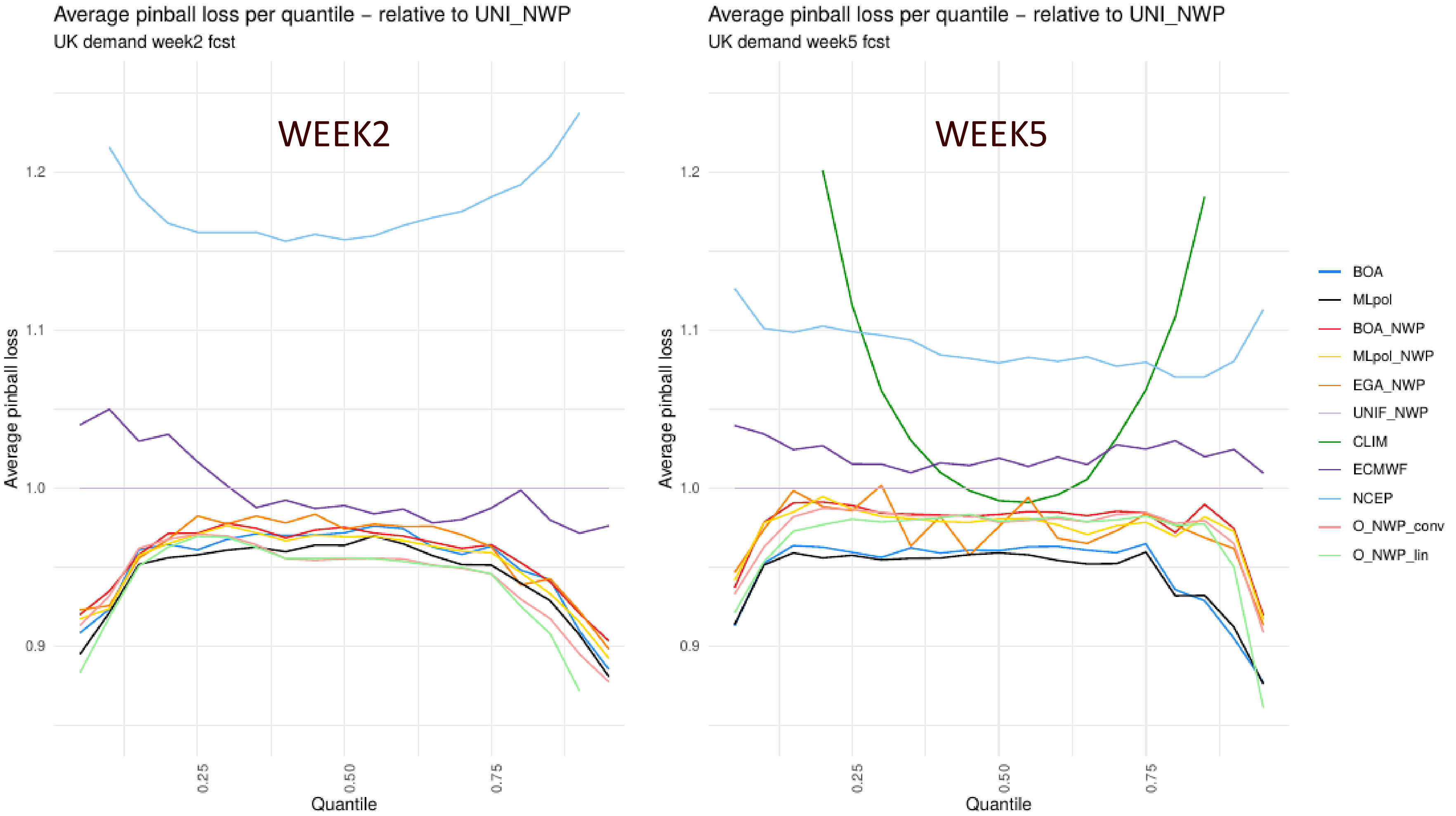
S2S4E Climate Services for Clean Energy

RESULTS: PROBABILISTIC SKILL

Q50 doesn't tell the whole story …

For each quantile in Qgrid, the relative pinball loss wrt UNIF_NWP

- Week 2 → BOA, MLpol, oracles all quite similar
- Week 5 → BOA and MLpol show clearly higher improvements
- EGA_NWP more unstable due to 'fixed' learning rate



S2S4E Climate Services for Clean Energy

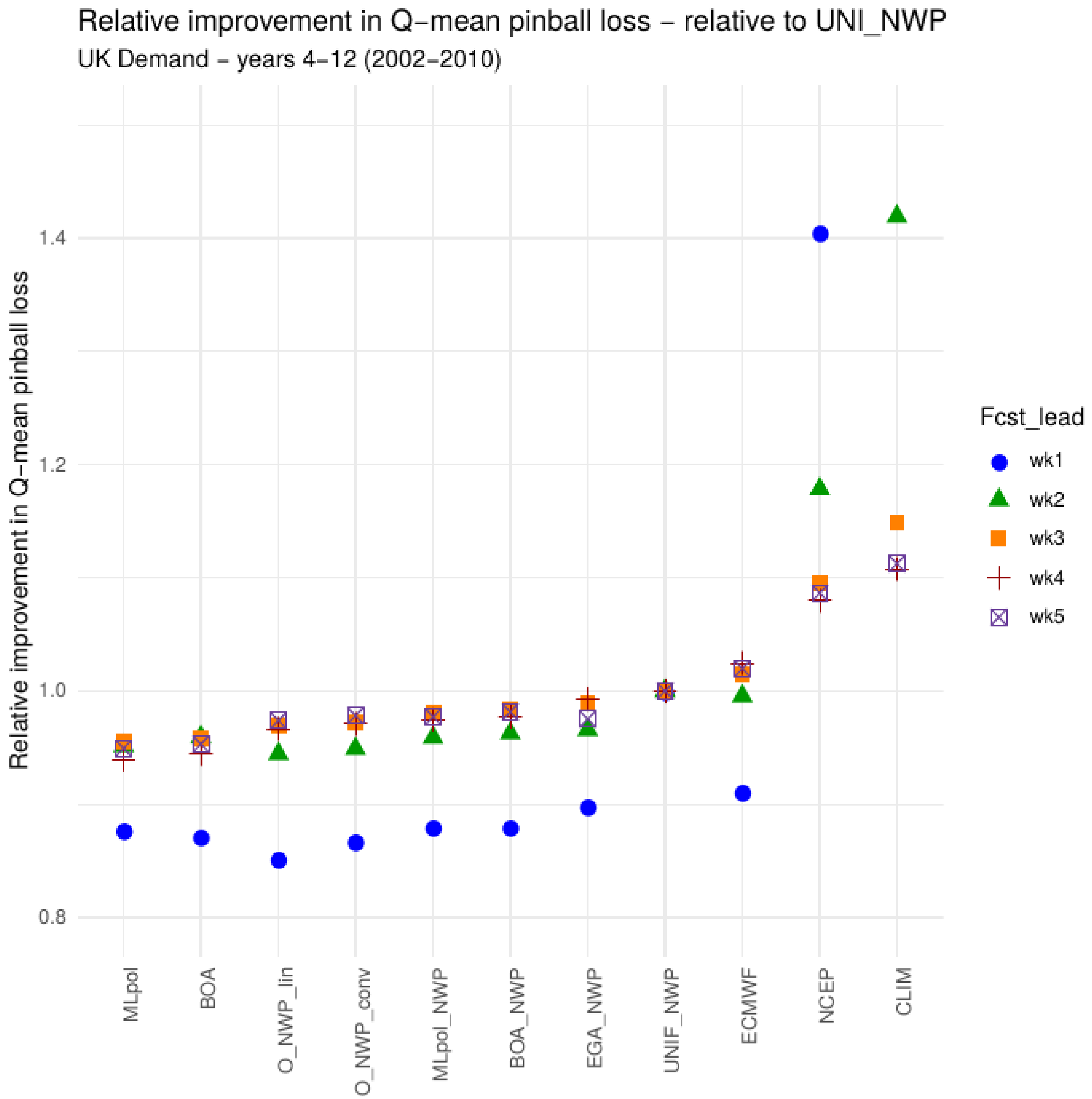
RESULTS: PROBABILISTIC SKILL

Q-mean pinball loss → ~ CRPS (for fine Qgrid)
Relative improvements w.r.t. UNIF_NWP

- BOA & MLpol beat all the other mixtures for weeks 2-5

- Week 3: skill increase for MLpol (best mixture) is 5% and slightly higher for weeks 4-5

- EGA_NWP shows improvements wrt UNIF_NWP but is unable to beat MLpol/BOA for any lead time



4E
ervices
Energy

RESULTS: PROBABILISTIC SKILL

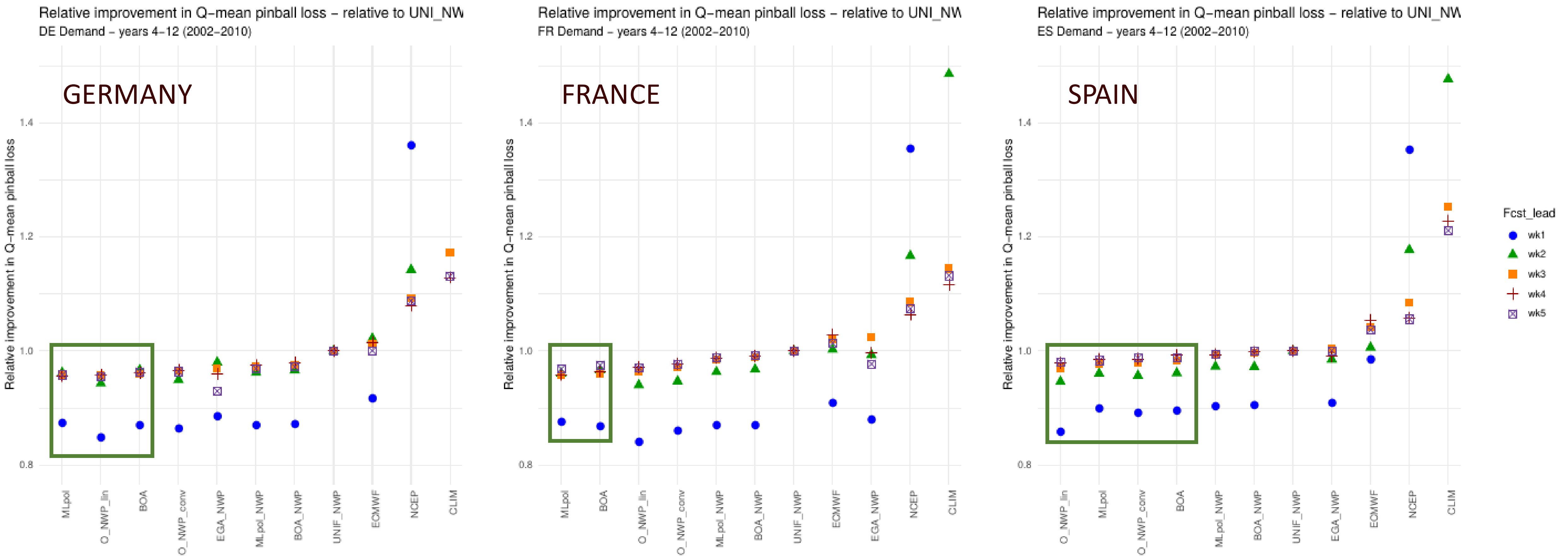
Statistical significance of the skill improvements
Diebold Mariano test (Diebold & Mariano 1995)

COMPARISON	WEEK1	WEEK2	WEEK3	WEEK4	WEEK5
BOA vs. UNIF_NWF	**100.00**	**99.97**	**99.99**	**100.00**	**100.00**
MLpol vs. UNIF_NWF	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
BOA vs. BOA_NWP	**97.79**	71.73	**99.59**	**99.92**	**99.84**
MLpol vs. MLpol_NWP	72.87	89.01	**99.60**	**99.99**	**99.92**
MLpol vs. BOA	7.83	**99.93**	*92.62*	**99.25**	*97.58*
MLpol_NWP vs. BOA_NWP	50.59	**99.91**	**99.14**	*96.76*	**99.97**
BOA_NWP vs. UNIF_NWP	**100.00**	**99.98**	*98.36*	**99.60**	*98.64*
MLpol_NWP vs. UNIF_NWP	**100.00**	**100.00**	**99.61**	**99.90**	**99.76**
EGA_NWP vs. UNIF_NWP	**100.00**	**99.70**	76.85	63.88	89.63

- **BOA & MLpol** are significantly more skilful than the uniform combination for **every lead** (same for NWP versions)
- **BOA & MLpol** are significantly more skilful than their **NWP-only** counterparts for lead **weeks 3-5**
- **MLpol** is significantly more skilful than **BOA** for lead **weeks 2-5** (same for NWP versions)
- **EGA_NWP** is significantly more skilful than the uniform combination for lead **weeks 1-2**

RESULTS: PROBABILISTIC SKILL

Generalization of the results: other countries



- MLpol and BOA **only beaten** in some cases by the **oracle combinations** → they provide an **upper limit for skill** because they require knowledge of the full period (unrealistic)
- The combinations that use **reanalysis-based experts** always have higher skill
- The magnitude of the **skill increases are consistent** (~2-5%)

S2S4E
Climate Services for Clean Energy

SUMMARY AND DISCUSSION

▪ The analysis presented here shows **very promising results** from the application of online prediction with expert advice to electricity demand. The BOA and MLpol methods show **skill improvements for leads beyond week 3**, a horizon rarely beaten by ECMWF at the country level.

▪ The full extent of the benefits of these methodologies is seen through their application to the complete Qgrid (probabilistic skill rather than deterministic).

▪ In the case of UK demand, MLpol and BOA provided skill enhancements of around 5% for weeks 3-5.

▪ The mixtures were significantly more skilful when they included reanalysis-based experts. We hypothesize that this is due to the fact that these experts provide the system with a memory-like effect of how the recent past behaved with respect to climatology and can therefore adjust the weights accordingly.

▪ The algorithms tested here beat the EGA sequential learning benchmark, which has been previously used in weather and climate prediction.

▪ The application of the methods to 3 other large countries yielded analogous results.

ANY QUESTIONS AND COMMENTS ARE WELCOME !

p.gonzalez@reading.ac.uk