Additive Covariance Matrix Models: Modelling Regional Electricity Net-Demand in Great Britain

V. Gioia\textsuperscript{1}, M. Fasiolo\textsuperscript{2}, J. Browell\textsuperscript{3}, R. Bellio\textsuperscript{4}

\textsuperscript{1} University of Trieste, Department of Economics, Business, Mathematics and Statistics
\textsuperscript{2} University of Bristol, School of Mathematics
\textsuperscript{3} University of Glasgow, School of Mathematics and Statistics
\textsuperscript{4} University of Udine, Department of Economics and Statistics

vincenzo.gioia@units.it

Abstract

Forecasts of regional electricity net-demand, consumption minus embedded generation, are an essential input for reliable and economic power system operation, and energy trading. While such forecasts are typically performed region by region, operations such as managing power flows require spatially coherent joint forecasts, which account for cross-regional dependencies. Here, we forecast the joint distribution of net-demand across the 14 regions constituting Great Britain’s electricity network. Joint modelling is complicated by the fact that the net-demand variability within each region, and the dependencies between regions, vary with temporal, socio-economic and weather-related factors. We accommodate for these characteristics by proposing a multivariate Gaussian model based on a modified Cholesky parametrisation, which allows us to model each unconstrained parameter via an additive model. Given that the number of model parameters and covariates is large, we adopt a semi-automated approach to model selection, based on gradient boosting. In addition to comparing the forecasting performance of several versions of the proposed model with that of two non-Gaussian copula-based models, we visually explore the model output to interpret how the covariates affect net-demand variability and dependencies.

The code for reproducing the results in this paper is available at \url{https://doi.org/10.5281/zenodo.7315105}.

Keywords: Covariance Matrix Regression Modelling; Generalized Additive Models; Modified Cholesky Decomposition; Multivariate Electricity Net-Demand Forecasting.
1 Introduction

Electricity networks are changing from centralised systems, where power is generated by large power plants connected to the transmission network and consumed mostly on the distribution network, to decentralised networks where significant generation and storage is connected directly to distribution networks. The growth of this embedded generation means that the transmission network now needs to serve the net-demand of customers, that is their demand net of local production. In Great Britain (GB), embedded production comes mostly from domestic and small-to-medium commercial solar and wind farms, as well as small thermal power plants. The lack of visibility of these units at the transmission level, combined with the weather-dependent nature of renewable generation leads to considerable challenges in energy trading and power system operation (Huxley et al., 2022).

The purpose of this work is to support such key operations by proposing an interpretable modelling approach that provides probabilistic, spatially coherent short-term net-demand forecasts. The energy industry is conservative by nature due to the need to maintain security of supply (see Chapter 8 of von Meier, 2006). As a result, new processes will only be adopted if they are trusted, and interpretability plays an important role in building trust. Further, interpretability is of critical importance when extrapolation is required, for example during exceptional events such as extreme temperatures, or when the model’s predictions must be decomposed into the contribution of several effects.

Predicting power flows on the electricity transmission network is a key motivating application for probabilistic, spatially coherent modelling in energy forecasting. This is important for both network operators, who are responsible for system security, and traders who must be aware of spatial variation in prices. Power flows are influenced by the injection and offtake of power from the network, as well as network configuration. They are also constrained by the physics of the network and must be forecasted to identify and mitigate any risk of exceeding thermal or stability limits. Therefore, spatial probabilistic forecasts of supply and demand are required to forecast power flows, and quantify uncertainty and risk associated with these constraints. Further, as the configuration of the network may change, any forecasting system must be flexible enough to allow the aggregation of supply and demand on the fly to calculate flows across relevant boundaries (Tuinia et al., 2020).

Motivated by the need for probabilistic joint net-demand forecasts, we consider joint modelling of net-demand across the 14 regions constituting GB’s transmission network, which are shown in Figure 1. The net-demand in each region is the aggregate of the net-demand across many Grid Supply Points (GSPs), the latter being the interfaces between the transmission system and either a distribution network or a high-voltage consumer. Correctly modelling the dependency structure between regions is critical as an error in such a structure would pollute downstream predictions of power flows. Further, joint probabilistic forecasts of regional net-demand can be flexibly post-processed to produce forecasts tailored
to the needs of different analyses. For example, when sustained winds blow in the North of the country, the boundary between Scotland and the North of England is of particular interest. Indeed, the power flows across this boundary can be substantial, their direction and size being driven mostly by wind generation in Scotland, where embedded wind capacity far exceeds regional demand. Similarly, when the sun shines across the country, power flows from the South of England, where most of the country’s solar generation units are installed, to London and the Midlands. During winter peak hours, the boundaries between London and its surroundings are characterised by heavy power flows directed toward the Capital, driven by high demand and low generation capacity in the city.

While National Grid’s 2021 ten-year statement (National Grid, 2021) gives a detailed description of tens of transmission grid boundaries and explains under which circumstances the power lines crossing them become heavily loaded, the examples given above are meant to convey the fact that power flow analysis on operational time scales needs regional net-demand forecasts that can be aggregated, or more generally post-processed, to match the particular scenario of interest. For concreteness, in this work we consider the aggregation into the five GSP macro-regions shown in Figure 1, which are motivated by the boundaries mentioned above and match closely the critical boundaries presented in a Regional Trends and Insights report from the National Grid (National Grid, 2018, see Figure 1 therein).

Forecasts of power flows, which are composite variables of demand and supply, among many other factors, can only be calculated if a multivariate predictive distribution of said quantities is available. Hence, joint forecasts of net-demand across the GSP regions shown in Figure 1, or across an appropriate, scenario-dependent aggregation of them, are essential to help the network operator to take early action when managing the risk of breaching constraints. However, the structure of spatial dependency in net-demand is complex, as it is influenced by both socio-economic and weather effects, and is time-varying. Figure 2a-b illustrates the issue. In particular, the conditional regional standard deviations and inter-regional correlations of net-demand, predicted one day ahead by one of the models proposed in this paper. Figure 2a corresponds to New Year’s Eve, a day where net-demand forecast uncertainty is particularly high and correlation is strong between densely populated areas, such as London and the West Midlands. In contrast, Figure 2b corresponds to the 20th of August and shows that net-demand is predicted to lead to a quieter day, from a network management perspective, with weak spatial dependency in forecast uncertainty.

Figure 2a-b makes clear that capturing the time-varying nature of regional net-demand dynamics is essential to produce operationally useful joint forecasts. However, several other factors affect the joint distribution of regional net-demand, in addition to daily and yearly seasonalities. For example, the right column of Figure 2 shows the joint configuration of the macro-regional net-demand variabilities and dependencies during storm Hector. In particular, Figure 2d shows the prediction obtained by conditioning on the seasonal factors and weather forecasts corresponding to this time period. Figures 2c–e have been obtained by
Figure 1: A map of the GSP groups forming GB’s electricity grid. The letters are the designation used by the electricity market in GB, while the numbers correspond to the position of each GSP group in the response vector, $y_i$ (see Section 3.2). The colours represent five macro-regions namely Scotland (red), Northern (green), Midlands (light blue), Southern (blue) and London (yellow). The data on GSPs boundaries have been obtained from https://data.nationalgrideso.com.

respectively decreasing and increasing the regional wind speed and precipitation forecasts by 25%. They show that weather has a strong effect on the joint distribution of net-demand. Specifically, a strengthening of the storm is predicted to lead to higher variability in Scotland and to stronger correlations between the latter and the other macro-regions.

Having motivated the need for an interpretable covariance matrix modelling framework able to provide the spatially coherent net-demand forecasts required by power flow analysis, we now outline the modelling approach proposed here. We jointly model GB regional net-demand using a multivariate Gaussian model, based on a covariance matrix parametrisation that allows us to model each of its unconstrained parameters via a separate additive model, containing both parametric and smooth spline-based effects. In particular, the covariance matrix of the regional net-demand vector is parametrised via the modified Cholesky decomposition (MCD) of Pourahmadi (1999). The wiggliness of the smooth effects is controlled via smoothing penalties, whose strength is controlled via smoothing parameters. Model fitting is performed via two nested iterations, the regression coefficients being estimated via maximum a posteriori (MAP) methods, while the smoothing parameters are selected by maximising a Laplace approximation to the marginal likelihood (LAML).

The proposed model can be seen as a multi-parameter generalized additive model
Figure 2: Conditional standard deviations (nodes) and correlations (edges) across the 14 GSP regions (a–b) or macro-regions (c to e), predicted by the model from Section 3.3. The edges corresponding to correlations lower than 0.25 have been made transparent. The plots correspond to 7am on 31/12/18 (a), midnight on 20/08/18 (b) and 10am on 14/06/18 (c to e). Plot d is based on regional wind and precipitation forecasts, while c and e correspond to, respectively, a 25% decrease and increase of such forecasts. Note that the colour scales used for the nodes and the edges are different between a-b and c-e.

(GAM, Hastie and Tibshirani, 1987) or as a generalized additive model for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005). Additive models are popular modelling tools in electricity demand forecasting (see, e.g., Fan and Hyndman, 2012), in part because they strike a balance between predictive performance and interpretability, the importance of the latter in this context having been discussed above. While ensuring interpretability is challenging under the richly parametrised model considered here, the
MCD parametrisation provides some degree of interpretability when the response vector has some, not necessarily unique, intrinsic ordering, as is the case for regional net-demand. We further enhance the interpretable exploration of the model by summarising its output via the accumulated local effects (ALEs) of Apley and Zhu (2020).

The proposed joint regional net-demand model has 119 distributional parameters, controlling the mean vector and the covariance matrix of a conditional multivariate Gaussian distribution. Each parameter can be modelled via parametric and smooth effects of several covariates, hence the space of possible models is large. While the effects controlling the mean vector can be chosen based on expert knowledge or previous research, manual selection of an additive model for each of the remaining 105 parameters is unrealistic. Here, we leverage the interpretation of the MCD’s parameters to choose the set of candidate effects that could be used to model each parameter. Then, we use gradient boosting (Friedman, 2001) to order the effects on the basis of how much they improve the fit, and we choose the number of effects modelling the MCD elements by maximising the forecasting performance on a validation set. The results show that the semi-automatic effect selection procedure just outlined leads to satisfactory predictive performance and to model selection decisions that are largely in agreement with intuition (e.g., wind speed and solar irradiance are selected to model net-demand variability in, respectively, Scotland and the South of England).

To our knowledge, this is the first applied statistical paper to consider additive modelling of each parameter of the mean vector and of an unconstrained covariance matrix parametrisation, in a context where the response vectors are not low-dimensional and have heterogeneous elements, i.e. they are not lagged values of the same variable. Additive modelling of multivariate responses has been proposed by Klein et al. (2015), who consider bivariate Gaussian and t-distributions based on a variance-correlation decomposition. Marra and Radice (2017) propose a fitting framework where bivariate responses are modelled via copulas with continuous margins, and all distributional parameters are modelled additively. Also relevant are the covariate-dependent copula approaches of Vatter and Nagler (2018) and Hans et al. (2023), who provide examples featuring respectively four- and two-dimensional responses. Similarly to Hans et al. (2023), Strömer et al. (2023) use the gradient boosting methods of Thomas et al. (2018) for model fitting, and consider several two-dimensional response models, including the bivariate Gaussian one. Klein et al. (2022) propose a copula-related approach that makes minimal distributional assumptions on the margins. While their model did not originally include penalised smooth effects, which are essential for the application considered here, it is conceivable that such effects could be included. Such an addition, if supported by sufficiently scalable fitting methods and software, could make their approach a serious alternative to our proposal.

In a generalized linear modelling (GLM) context, Pourahmadi (1999) uses the MCD to parametrise a multivariate Gaussian model in eleven dimensions. They are interested in capturing temporal dependencies in longitudinal data, which allows them to impose a strong
structure on the covariance matrix model. Beyond the bivariate case, Bonat and Jørgensen (2016) model directly the covariance matrix using covariates and tune the optimiser to avoid generating indefinite matrices, while Browell et al. (2022) use covariates-dependent covariance functions which, in some cases, do not provide positive definiteness guarantees and require perturbing the estimated covariance matrix to achieve it.

From a methodological perspective, our work is closely related to Muschinski et al. (2024), who consider non-parametric regression with multivariate Gaussian responses. They propose modelling of the elements of several Cholesky-based parametrisations, including the MCD. But, they fit the model using MCMC methods, rather than direct optimisation methods as done here and, while they compare a set of manually-chosen covariance matrix models on a weather forecasting application, here we consider semi-automatic variable selection to handle a much larger set of candidate covariates. Further, they aim at capturing temporal rather than spatial dependencies, the latter being the focus of this work.

The rest of the paper is structured as follows. Section 2 introduces, in a general setting, the proposed multivariate Gaussian model structure and fitting methodology. It also summarises the inferential framework and motivates the use of ALEs for model output exploration. Section 3 focuses on the regional net-demand modelling application. In particular, the data is introduced in Section 3.1, while Section 3.2 describes the bespoke, boosting-based model selection approach proposed here. The output of the final model is explored in Section 3.3, while the forecasting performance of the proposed model is assessed in Sections 3.4 and 3.5. Section 4 summarises the main results.

2 Multivariate Gaussian Additive Models

2.1 Model Structure

Let \( y_i = (y_{i1}, \ldots, y_{id})^\top \), for \( i = 1, \ldots, n \), be independent response vectors, normally distributed with mean \( \mu_i \) and covariance matrix \( \Sigma_i \). The \( q = d + d(d + 1)/2 \) unique elements of \( \mu_i \) and of (a suitable parametrisation of) \( \Sigma_i \) are modelled via \( \eta_i \), a \( q \)-dimensional vector of linear predictors. The \( j \)-th element of \( \eta_i \) is modelled via

\[
\eta_{ij} = Z_{ij}^\top \psi_j + \sum_l f_{jl}(x_{il}^{S_{jl}}), \quad \text{for } j = 1, \ldots, q,
\]

where \( Z_{ij}^\top \) is the \( i \)-th row of the design matrix \( Z^j \), \( \psi_j \) is a vector of regression coefficients, \( x_i \) is an \( s \)-dimensional vector of covariates and \( S_{jl} \subset \{1, \ldots, s\} \). Hence, for example, if \( S_{jl} = \{2, 4\} \) then \( x_{il}^{S_{jl}} \) is a two dimensional vector formed by the second and fourth element


of $x_i$. Each $f_{jl}$ is a smooth function, built via

$$ f_{jl}(x^{S_{jl}}) = \sum_k b_{jl}^k(x^{S_{jl}})\alpha_{jl}^k, \quad (2) $$

where $b_{jl}^k$ are spline basis functions of dimension $\text{card}(S_{jl})$, while $\alpha_{jl}^k$ are regression coefficients. Denote with $\alpha$ the vector of all such coefficients in the model. The wiggliness of the effects is controlled by an improper multivariate Gaussian prior on $\alpha$. The prior is centered at the origin and its precision matrix is $S^\lambda = \sum_n \lambda_n S_n$, where the $S_n$’s are positive semi-definite matrices and $\lambda = (\lambda_1, \lambda_2, \ldots)^\top$ is a vector of positive smoothing parameters. See Wood (2017) for a detailed introduction to GAMs, smoothing splines bases and penalties.

Let us temporarily drop index $i$ to simplify the notation. In this work we use the following parametrisation of $\mu$ and $\Sigma$ in terms of $\eta$: $\mu_j = \eta_j$ for $j = 1, \ldots, d$, while the remaining elements of $\eta$ parametrise an MCD of $\Sigma^{-1}$ (Pourahmadi, 1999). In particular,

$$ \Sigma^{-1} = T^\top D^{-2} T, \quad (3) $$

where $D^2$ is a diagonal matrix with $D_{jj} = \exp(\eta_{j+d})$, for $j = 1, \ldots, d$, and

$$ T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\
\eta_{2d+1} & 1 & 0 & \cdots & 0 \\
\eta_{2d+2} & \eta_{2d+3} & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\eta_{q-d+2} & \eta_{q-d+3} & \cdots & \eta_q & 1 \end{pmatrix}. \quad (4) $$

Parametrisation (3) is unconstrained, that is the resulting covariance matrix $\Sigma$ is positive definite for any finite $\eta$, which facilitates model fitting. Other unconstrained parametrisations could have been used, such as those discussed by Pinheiro and Bates (1996) and Pourahmadi (2011), but the MCD approach is particularly attractive in the context of this work. First, the fitting methods described in Section 2.2 require the first two derivatives of the log-likelihood w.r.t. $\eta$ and, under the MCD parametrisation, the multivariate Gaussian log-likelihood can be written directly in terms of $\eta$, which eases the computation of such derivatives. Second, the MCD parametrisation has a regression-related interpretation, which can be exploited when the response vector has some intrinsic ordering. In particular, assume w.l.o.g. that $E(y) = 0$ and that $y$ follows the regression models

$$ y_l = \sum_{k=1}^{l-1} \phi_{lk} y_k + \epsilon_l, \quad \text{for} \quad l = 2, \ldots, d, $$

where $y_1 = \epsilon_1$, $\text{var}(\epsilon_l) = \sigma_l^2$ and $\text{cov}(\epsilon_l, \epsilon_k) = 0$ for $l \neq k$. Pourahmadi (1999) shows that $T_{lk} = -\phi_{lk}$ and $D_{kk}^2 = \sigma_k^2$, for $k = 1, \ldots, d$, and $l = k + 1, \ldots, d$. Hence, the elements of $T$
can be interpreted as the regression coefficients of the elements of \( y \) on their predecessors, while the non-zero elements of \( D^2 \) are the residual variances of such regressions.

Note that the interpretation of the MCD elements depends on the ordering of the elements of the response vector. Hence, the MCD parametrisation is particularly attractive when the response vector has some natural ordering, as is the case when dealing with chronologically ordered responses. While in this work we consider a response vector that does not have a unique natural ordering, in Section 3.2 we will discuss how the spatial nature of GB regional net-demand data allows us to exploit the interpretation of the MCD parametrisation to guide the development of a multivariate model.

2.2 Model Fitting

Let us indicate the set of all response vectors \( y_1, \ldots, y_n \) simply with \( y \) and with \( \beta \) the vector of all regression coefficients in the model, which include \( \alpha \) and all the unpenalised coefficients vectors \( \psi_j \). Let \( \tilde{S}^\lambda \) be the prior precision matrix of \( \beta \), that is an enlarged version of \( S^\lambda \) padded with zeros so that \( \alpha^\top S^\lambda \alpha = \beta^\top \tilde{S}^\lambda \beta \). Then, up to an additive constant that does not depend on \( \beta \), the Bayesian posterior log-density of the model from Section 2.1 is

\[
\mathcal{L}(\beta) = \log p(\beta | y, \lambda) = \sum_{i=1}^{n} \log p(y_i | \beta) - \frac{1}{2} \beta^\top \tilde{S}^\lambda \beta, \tag{5}
\]

where \( \log p(y_i | \beta) \) is the \( i \)-th log-likelihood contribution.

For fixed smoothing parameters, \( \lambda \), we obtain MAP estimates of the regression coefficients by maximising the log-posterior (5), using Newton’s algorithm. The latter requires the gradient and Hessian of the log-posterior w.r.t. \( \beta \), which are provided in the Supplementary Material A (henceforth SM A). The real challenge is selecting the smoothing parameters themselves. We do it by maximising an approximation to the log marginal likelihood, \( \hat{\mathcal{V}}(\lambda) = \log \int p(y | \beta) p(\beta | \lambda) d\beta \). In particular, we consider the LAML criterion

\[
\hat{\mathcal{V}}(\lambda) = \mathcal{L}(\hat{\beta}) + \frac{1}{2} \log |\tilde{S}^\lambda|_+ - \frac{1}{2} \log |H| + \frac{M_p}{2} \log(2\pi), \tag{6}
\]

with \( M_p \) being the dimension of the null space of \( \tilde{S}^\lambda \), \( |\tilde{S}^\lambda|_+ \) the product of its positive eigenvalues, \( \hat{\beta} \) the maximiser of \( \mathcal{L}(\beta) \) and \( H \) its negative Hessian, evaluated at \( \hat{\beta} \).

We maximise \( \hat{\mathcal{V}}(\lambda) \) via the generalized Fellner-Schall method of Wood and Fasiolo (2017), under which the \( u \)-th smoothing parameter is updated using

\[
\lambda_{u}^{\text{new}} = \frac{\text{tr}[(\tilde{S}^\lambda)^{-1} \tilde{S}_u] - \text{tr}(H^{-1} \tilde{S}_u \hat{\beta}^\top \hat{S}_u \hat{\beta})}{\hat{\beta}^\top \hat{S}_u \hat{\beta}} \lambda_{u}^{\text{old}}, \tag{7}
\]
where \((\hat{S}^\lambda)^{-1}\) is the Moore-Penrose pseudoinverse of \(\hat{S}^\lambda\) and \(\hat{S}_u\) is \(S_u\) after padding it with zeros. That is, if we indicate with \(\beta_u\) the subvector of \(\beta\) that is penalised by \(S_u\), then \(\beta_u^\top S_u \beta_u = \beta^\top \hat{S}_u \beta\). An advantage of update (7) is that it does not require computing the derivatives of \(\hat{V}(\lambda)\) w.r.t. \(\lambda\). In particular, as detailed in Wood et al. (2016), computing the gradient of \(\hat{V}(\lambda)\) requires the third derivatives of log-likelihood w.r.t. each element of \(\eta\), which leads to computational effort of order \(O(n^{q^2}/3)\) \((\approx 2 \times 10^{10}, \text{if} q = 119 \text{and } n \approx 8 \times 10^4 \text{as in the application considered here})\). Hence, for moderately large dimension \(d\) of the response vector, a quasi-Newton iteration for maximising \(\hat{V}(\lambda)\) would be too computationally intensive, at least under naïve evaluation of the likelihood derivatives.

### 2.3 Inference and Effect Visualisation

The uncertainty of the fitted regression coefficients, \(\beta\), can be quantified via the approximate Bayesian methods detailed in Wood et al. (2016), which we summarise here. In particular, standard Bayesian asymptotics justify approximating \(p(\beta|y, \lambda)\) with a Gaussian distribution, \(N(\hat{\beta}, V_\beta)\), centered at the MAP estimate and with covariance matrix \(V_\beta = -H^{-1}\). Such a posterior approximation does not take into account the uncertainty of the smoothing parameters estimates, which are considered fixed to the LAML maximiser. Wood et al. (2016) use a Gaussian approximation to \(p(\lambda|y)\) and propagate forward the corresponding smoothing parameter uncertainty to obtain an approximation to the unconditional posterior, \(p(\beta|y)\). In principle, this approach could be adopted for the model class considered here, but the formulae provided by Wood et al. (2016) require the Hessian of \(\hat{V}\) w.r.t. \(\lambda\) which involves the fourth derivative of log-likelihood w.r.t. each element of \(\eta\). While it might be possible to reduce the analytical effort needed to obtain such derivatives by automatic differentiation (see, e.g., Griewank and Walther, 2008) the computational cost mentioned in Section 2.2 would still be an obstacle.

Given that the smooth effects are linear combinations of the regression coefficients, it is simple to derive pointwise Bayesian credible intervals for the effects, the asymptotic frequentist properties of such intervals having been studied by Nychka (1988). However, each effect acts directly on a linear predictor, the latter being non-linearly related to one or more elements of \(\Sigma\). As explained in Section 2.1, the MCD parametrisation is related to a set of regressions involving the elements of the response vector. This fact aids interpretability only if the response vector has some natural ordering. While this is to some extent the case in the application considered here (see Section 3.2), communicating modelling results to non-statisticians is more likely to be effective if framed in terms of widely-used concepts such as covariances and correlations, rather than parametrisation-specific quantities. Hence, we use the accumulated local effects (ALEs) of Apley and Zhu (2020) to quantify the effect of a covariate on \(\Sigma\) or on the corresponding correlation matrix, \(\Gamma\).

In contrast to the partial dependence plots (Friedman, 2001), ALEs avoid making an
extrapolation error when the covariates are correlated. This is explained by Apley and Zhu (2020), who also provide formulas for estimating ALEs, and quantify their uncertainty via bootstrapping. Here, we exploit the results of Capezza et al. (2021) for multi-parameter GAMs to approximate the posterior variance of ALEs. In SM A.4 we provide more details on ALEs, while in Section 3.3 we use them to visualise a model for GB regional net-demand.

### 3 Joint Multivariate Regional Net-Demand Modelling

#### 3.1 Data Description and Modelling Setting

We consider data on regional net-demand in GB, from five years, 2014 to 2018. Net-demand is the load measured at the interface between transmission and distribution networks. In GB these interfaces are called Grid Supply Points (GSPs) and are grouped into 14 GSP groups. Let $y_{ij}$, for $i = 1, \ldots, n$, be the standardised net-demand of GSP group $j$ measured at a 30min resolution. In addition to net-demand, the data contain the covariates listed in Table 1 or transformations thereof. Some covariates are common to all GSP groups, while others are region-specific, such as those derived from the hourly day-ahead weather forecasts produced by the operational ECMWF-HRES model. Gridded weather predictions are summarised via the regional features reported in Table 1.

Browell and Fasiolo (2021) use the data just described to model the conditional distribution of $y_{ij}$, separately for each of the $d = 14$ regions. They do so using a composite modelling approach, where the raw residuals of a Gaussian GAM are modelled via linear

| General covariates                                      | Covariates derived from weather forecasts                   |
|--------------------------------------------------------|-------------------------------------------------------------|
| $\text{dow}_i$ day of the week factor                   | $\text{rain}_{ij}$ mean precipitation (m/day)               |
| $\text{dow}^+_i$ $\text{dow}_i$ with additional factor levels accounting for public holidays | $\text{temp}_{ij}$ temperature (K) at cell with highest regional population density |
| $\text{t}_i$ time since the 1st January 2014            | $\text{temp}^5_{ij}$ 48 hours rolling mean of $\text{temp}_{ij}$ |
| $\text{shol}_{ij}$ school holidays, three levels factor to distinguish Christmas from other holidays | $\text{irr}_{ij}$ mean solar irradiance (W/m$^2$) times embedded solar generation capacity (MW) |
| $\text{tod}_i$ time of day ($\in \{0, 0.5, \ldots, 23.5\}$) | $\text{wsp}_{10}^{ij}$ mean wind speed at 10 meters (m/s) |
| $\text{wcap}_i$ GB embedded wind generation capacity (MW) | $\text{wsp}_{100}^{ij}$ mean wind speed at 100 meters (m/s) |
| $\text{doy}_i$ day of the year ($\in \{1, \ldots, 366\}$) |                                                             |
| $\text{n2ex}_i$ N2EX day-ahead electricity price (£/MWh) |                                                             |
| $y_{ij}^{24}$ net-demand at a 24 hours lag              |                                                             |

Table 1: Covariates used to model GB regional net-demand.
quantile regression. Extreme quantiles are modelled using a GAMLSS model based on the generalized Pareto distribution. Here, we are interested in modelling the joint distribution of the $d$-dimensional response vector $y_i$. We consider a multivariate Gaussian model $y_i \sim N(\mu_i, \Sigma_i)$, where $\mu_i$ and $\Sigma_i$ are controlled by the linear predictors vector $\eta_i$, as described in Section 2.1, and each element of $\eta_i$ is modelled via (1).

The proposed model has $q = 119$ linear predictors and each of them could be modelled via any of the covariates described above. Hence, model selection is challenging. As explained in Section 3.2, we use 2014-2016 data to generate a long list of candidate covariate effects, ordered in terms of decreasing importance. We choose the number of effects to add to the final multivariate Gaussian model, that is where to stop along the ordered effect list, by maximising the out-of-sample predictive performance on 2017 net-demand. Having chosen the model structure, in Section 3.3 we explore the model output, and in Sections 3.4 and 3.5 we evaluate the accuracy of the resulting forecasts on 2018 data.

3.2 Semi-Automatic Model Selection

Browell and Fasiolo (2021) consider a progression of univariate GAMs based on an increasingly rich set of covariates and assess their performance on a day-ahead forecasting task. We use their results to choose a model for the first $d = 14$ elements of $\eta_i$ or, equivalently, $\mu_i$. In particular, we adopt the model formula

$$\eta_{ij} = g_{j1}(t_i) + g_{j2}(t_i^2) + g_{j3}(\text{dow}_{ij}^+) + g_{j4}(\text{shol}_{ij}) + g_{j5}(\text{g}_{ij}^{24}) + g_{j6}(\text{wsp}_{ij}^{10})$$

$$+ f_{j1}^{20}(\text{doy}_i) + f_{j2}^{35}(\text{tod}_i) + f_{j3}^{10}(\text{n2ex}_i) + f_{j4}^{35}(\text{temp}_i) + f_{j5}^{35}(\text{temp}^S_i) + f_{j6}^{10}(\text{rain}^{1/2}_i)$$

$$+ \text{wcap}_i \times f_{j7}^{20}(\text{wsp}_{ij}^{100}) + f_{j8}^{5}(\text{irr}_{ij}) + f_{j9}^{30}(\text{tod}_i, \text{dow}_{ij}^+) + f_{j10}^{20}(\text{tod}_i, \text{shol}_{ij})$$

$$+ f_{j11}^{5.5}(\text{n2ex}_i, \text{tod}_i) + f_{j12}^{5.5}(\text{temp}_{ij}, \text{tod}_i) + f_{j13}^{5.5}(\text{rain}^{1/2}_i, \text{tod}_i) + f_{j14}^{10.10}(\text{doy}_i, \text{tod}_i), \quad (8)$$

for $j = 1, \ldots, d$. Here $g_{j1}$ to $g_{j6}$ are parametric (linear) effects, while $f_{j1}$ to $f_{j14}$ are smooth effects. In particular, $f_{j1}$ to $f_{j8}$ are univariate smooth effects, the spline bases dimensions being indicated by the superscripts. Effects $f_{j9}$ and $f_{j10}$ are smooth-factor interactions, where a different univariate smooth is defined for each level of the $\text{dow}_{ij}^+$ or $\text{shol}_{ij}$ factor variables. The last four effects in (8) are bivariate tensor-product smooths, where the dimension of each marginal basis is indicated by the superscripts. All smooth effects are built using cubic regression spline bases, except for $f_{j7}^{20}(\text{doy}_i)$, which uses a B-spline basis with an adaptive P-spline penalty. The latter allows the smoothness of the effect to vary with $\text{doy}_i$, see Section 5.3.5 of Wood (2017) for details. Model (8) is similar to the “GAM–point” model of Browell and Fasiolo (2021) but their model lacks the interaction between $\text{doy}_i$ and $\text{tod}_i$, and uses $\text{rain}_{ij}$ rather than $\text{rain}_{ij}^{1/2}$, the transformed version leading to more even use of basis functions (rainfall is rightly skewed). Further, they use a parametric effect, based on basic trigonometric functions of $\text{doy}_i$, to model the annual seasonality. The

12
approach used here models seasonality more flexibly, which is particularly important around year-end and in densely populated regions, such as London (see the results in Section 3.3).

It is challenging to develop a model for the remaining 105 elements of $\gamma$. As explained in Section 2.1, the elements of the $T_i$ and $D_i$ matrices correspond to the parameters of a set of linear models, where the $j$-th element of $y_i$ is regressed on its predecessors $y_{ij-1}, \ldots, y_{i1}$. Hence, the parameters of the decomposition depend on the ordering of the elements of $y_i$. For the GSP net-demand data, a sensible ordering can be chosen based on the location of the GSP regions. As Figure 1 shows, we order the regions North to South hence $y_{i1}$ and $y_{i14}$ are, respectively, net-demand in the North of Scotland and in the South West of England. Under such an ordering, neighbouring regions, which are more likely to be affected by similar weather and socio-economic events, occupy nearby positions in $y_i$. Of course, variations on the proposed ordering could be considered, for example one could think about swapping the order of regions 12 and 13, which are at a similar latitude, or about using a South-to-North ordering. More radically, one could experiment with orderings based purely on the socio-economic characteristics of each region, without taking spatial distances into account. While there might be orderings that lead to a significantly better forecasting performance than that achieved here under the proposed North-to-South ordering, performing a systematic assessment of the effect of ordering is infeasible under the complex model considered here, due to the substantial cost of model selection and fitting.

Given that the search for an ‘optimal’ ordering is impracticable, we design a semi-automatic model section procedure that does not explicitly take the ordering into account. In fact, while Pourahmadi (1999) proposed highly structured models for $D_i$ and $T_i$ which rely on the interpretation of their elements, and thus on their ordering, we use gradient boosting to choose which matrix elements should be modelled, and the effects that should be used to do so. The proposed approach is related to the method of Strømer et al. (2022) who use non-cyclical component-wise gradient boosting (Thomas et al., 2018) to determine the effects’ importance, and then run a further boosting procedure based on a subset of selected effects, chosen using a user-defined importance threshold. In contrast, we use the out-of-sample predictive performance to determine the number of effects to include in the final model and we fit the latter using the methods from Section 2.2, rather than boosting.

In the following we describe the proposed model selection approach, and we refer to SM B.1 for further details. For $j = 1, \ldots, d$, we fit a univariate Gaussian GAM, $y_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$, using net-demand data from the 1st of January 2014 to the 31st of December 2016, with $\mu_{ij} = \eta_{ij}$ modelled via (8). Then, define $\epsilon_{ij} = y_{ij} - \eta_{ij}$ and let $\Sigma$ be the empirical covariance matrix of such residuals. Having fixed $\eta_j$, for $j = 1, \ldots, d$, gradient boosting spans the candidate effects appearing in

$$
\begin{align*}
\eta_{ij} &= \bar{\eta}_{ij} + g_{j1}(t_i) + g_{j2}(d_{i1}) + g_{j3}(d_{i2}) + f_{j4}^{10}(d_{i3}) + f_{j5}^{10}(d_{i4}) + f_{j6}^{10}(d_{i5}) + w_{cap_i} \times f_{j7}^5(w_{sp_{il}^{10}}) + f_{j8}^5(i_{r{i}}^{il}) + f_{j9}^5(t_{em_{il}}^{il}) + f_{j10}^5(rain_{il}^{1/2}) + f_{j11}^5(n2ex_i) ,
\end{align*}
$$

(9)
for \( j = d + 1, \ldots, q \), where we indicate with \( \tilde{\eta}_{ij} \) the elements of the MCD of \( \Sigma^{-1} \), which serve as fixed offsets, with \( l_j \) the row of \( D_i \) or \( T_i \) on which the \( j \)-th linear predictor appears, and with \( \text{wsp}_1^{100}, \text{irr}_d \), and so on the weather forecasts corresponding to the \( l_j \)-th region. Thus, we exploit the interpretation of the MCD parametrisation to specify the candidate weather forecasts modelling the non-trivial elements of the \( l_j \)-th row of \( D_i \) and \( T_i \). Indeed, the forecasts for the North of Scotland are used to model only \( (D_i)_{11} \), those for the South of Scotland are used to model \( (D_i)_{22} \) and second row of \( T_i \), and so on. To see the reasoning behind this choice, recall from Section 2.1 that the linear predictors appearing on the \( l_j \)-th row of \( T_i \) or \( D_i \) are related to, respectively, the coefficients or the residual variance of the regression of the \( l_j \)-th element of \( y \) on its predecessors. Hence, it seems reasonable to use the weather forecasts for the \( l_j \)-th region to model the effect of the preceding regions on \( l_j \).

Note that (9) contains only a subset of the effects appearing in (8). In particular, no bivariate tensor-product smooth effect is used to model \( \eta_j \), \( j = d + 1, \ldots, q \). This choice is motivated by the fact that we are performing model selection across 105 linear predictors, so it is important to limit the number of candidate effects to ensure computational feasibility and statistical parsimony. For the same reason, the number of basis functions used to construct the effects in (9) is kept low. Further, the two terms \( g_{j1} \) and \( g_{j2} \) effectively form a single candidate effect in the model selection process.

To select the model for \( \eta_j \), with \( j = d + 1, \ldots, q \), we first run gradient boosting for \( M = 3000 \) iterations on the 2014-2016 data. At each iteration, the linear predictors fit the training data slightly better, which eventually leads to over-fitting. Having verified that over-fitting starts well before 3000 steps, we find the iteration \( M^* \in \{1, 2, \ldots, 3000\} \) at which the out-of-sample performance on 2017 data is optimal. The output of gradient boosting at step \( M^* \) is a list containing the selected effect-linear predictor pairs and the cumulative log-likelihood gains obtained by adding them to the boosting model (see SM B.1 for details). Assuming that the priority with which an effect-linear predictor pair should be added to the final model is proportional to its cumulative log-likelihood gain, the \( L \) pairs corresponding to the \( L \)-largest gains should be included in the final model, for some \( L \geq 0 \). Let \( L_1 = 0, L_2 = 5, L_3 = 10, \ldots, \) be a grid of potential values for the total number of effects, \( L \). To determine \( L \), we optimise the predictive performance of the full multivariate Gaussian model on 2017 data. This is done by first fitting univariate Gaussian GAMs and adopting a 1-month block rolling origin forecasting procedure starting from the 1st of January 2017 to predict the value of \( \eta_{ij} \), for \( j = 1, \ldots, d \), covering the whole of 2017. Then, by using the same rolling procedure, for each candidate value of \( L_j \) we fit the multivariate Gaussian model \( y_i \sim N(\mu_i, \Sigma_i) \) on 2017 data via the methods from Section 2.2, obtaining the out-of-sample predictions for \( \eta_{ij} \), \( j = d + 1, \ldots, q \), and the day-ahead predictions for \( \mu_i \) and \( \Sigma_i \) are used to compute the out-of-sample log-likelihood. The procedure suggests including \( L = 80 \) effects. Running gradient boosting for 3000 steps and evaluating the predictive performance take, respectively, around 42 and 5 hours when run in parallel on a
workstation with a 12-core Intel Xeon Gold 6130 2.10GHz CPU and 256GBytes of RAM.

3.3 Model Selection Results

Figure 3 shows the effects selected to model each element of $D_i$ and $T_i$. Recall that, under the interpretation detailed in Section 2.1, the elements on the $l$-th row of $T_i$ are the coefficients of the regression of $y_{it}$ on $y_{it-1}, \ldots, y_{i1}$, while the $l$-th diagonal element of $D_i$ is the corresponding residual variance. Hence, the effects acting on $D_i$ are not directly modelling the regional variances (i.e., the diagonal elements of $\Sigma_i$), but the residual variance of the net-demand in each region, after having conditioned on the preceding regions. Similarly, the effects acting on $T_i$ modulate the dependence between regions but they do not directly control correlations, which depend on the elements of both $D_i$ and $T_i$. The effects of several covariates on regional variances and correlations are shown in Figure 4 which contains a set of ALE plots, obtained by fitting a model containing the effects shown in Figure 3 to all the data available (2014-2018). Note that the 95% credible bands showed in Figure 4 are based on the Gaussian posterior approximation described in Section 2.3, which does not take into account the variability induced by the model selection procedure used here. Hence, the intervals might have coverage levels lower than the nominal ones.

Considering Figure 3, note that most of the effects act on the diagonal elements of $D_i$ and that the time of day, $\text{tod}_i$, affects all such elements. It is not surprising to see that the cumulative log-likelihood gain of the $\text{tod}_i$ effect is particularly large in highly urbanised areas, such as the Midlands (R. 8 and 9) and London (R. 11). The red ALE in Figure 4a shows the effect of daily seasonality in London, which is characterised by high net-demand variance during peak hours. The same effect has a similar, but flatter, shape in the South of Scotland (R. 2). In the South of England (R. 13) the effect has a single peak and it is even stronger than in London. As we discuss later, this is likely related to the high capacity of embedded solar generation relative to electricity consumption.

The time of day is used to model also several elements of $T_i$. The strongest such effect, in terms of cumulative log-likelihood gain, acts on the element corresponding to London and the West Midlands (R. 11 and 8). The ALE of $\text{tod}_i$ on the correlation between these regions is shown in green in Figure 4b. It shows that prediction errors in these urbanised regions are more correlated during the morning and evening ramps than in the middle of the day. The red curve shows that the correlation between London and the South of England (R. 13) has a similar pattern, although with milder daily oscillations.

The effect of the day of the year, $\text{doy}_i$, is used to model many elements of $D_i$ and $T_i$. It is not surprising to see this effect appearing on the 8th and 11th row of Figure 3, which correspond to the highly urbanised West Midlands (R. 8) and London (R. 11). As the green ALE in Figure 4c shows, net-demand forecast uncertainty is very high in London at the end of the year, due to holidays that have a sizeable, hard-to-model effect on demand.
patterns. Furthermore, as the effects in Figure 4d show, the uncertainty between regions is also highly correlated during this period, meaning that forecast errors are likely to have the same sign across regions as they are driven by the same underlying behavioural effects.

In accordance with intuition, wind speed, wsp$_{100}$, is selected to model the elements of $D_i$, corresponding to regions with a high penetration of embedded wind generation, such as the South East of England (R. 12), and the North (R. 1) and South (R. 2) of Scotland. The ALEs of wsp$_{100}$ in these regions are shown in Figure 4e and could be interpreted as follows. At low wind, the variability of wind production is low because little or no
generation is occurring, but it increases at modest wind speeds that are sufficient for power generation to occur, while being in a range where generation is highly sensitive to wind speed. Then variability decreases for high wind speeds, where power production is less variable as turbines self-regulate to maintain their maximum power production. At very high wind speeds, wind turbines may automatically shut down, and small differences in wind speed may result in large differences in production, leading to greater forecast uncertainty.

It is perhaps surprising that wind speed is not selected to control the element of $T_i$ controlling the dependency between the two Scottish GSPs, which both contain large amounts of embedded wind. However, capacity as a fraction of peak load is considerably higher in the North than in the South of Scotland. Further, the fact that $(T_i)_{21}$ does not depend on $wsp_{ij}^{100}$, does not mean that the correlation between the two Scottish regions stays constant as $wsp_{ij}^{100}$ changes, as illustrated by the green curve in Figure 4f. The plot shows that the correlation is proportional to the variance in these regions, hence wind speed controls both the size and correlation between prediction errors. Interestingly, the blue curve in Figure 4f shows that the net-demand in the South East of England is less correlated with that in London (R. 11) as wind speed in the former region increases, which suggests that this covariate affects the net-demand patterns of these two regions quite differently.

The time of day and solar irradiance, $irr_{ij}$, are both strongly related to solar energy production, hence it is interesting to see that the effects of both variables are selected to model the elements of $D_i$ corresponding to several Southern regions, which have high embedded solar generation capacity. The ALEs of $irr_{ij}$ on the net-demand variability in South Wales (R. 7), South (R. 13) and South-West (R. 14) England are shown in Figure 4g. Note that the horizontal scales are different because the installed solar capacity differs between regions. The shape of these effects is similar and could be interpreted as follows. Variability is low at low or high levels of irradiance which correspond to, respectively, heavily clouded (or night) and clear sky conditions. Variability is highest at intermediate levels of irradiance, which might correspond to partial or broken cloud conditions. However, the shape of the effects may also be affected by the correlation between irradiance and temperature and by changes in installed solar capacity over the study period, hence it is important not to over-interpret them. Solar irradiance is selected to control the dependency between the South of Wales and the South-West of England via the corresponding element of $T_i$. This is interesting because, while the two regions are separated by the Bristol channel, they are geographically close, hence likely to be affected by similar weather patterns, and they both feature very high solar penetration relative to peak load.

As explained above, the set of effects selected by the semi-automated procedure proposed in Section 3.2 matches intuition in many respects. However, looking at Figure 3, note that more than half of the selected effects are related to calendar variables, namely progressive time, time of day, day of year and day of the week. Further, external temperature has not been selected to model any element of $T_i$ or $D_i$. Hence, it is interesting
Figure 4: Left: ALEs of the time of day (a), day of year (c), forecasted wind speed (e) and solar irradiation (g) on the standard deviation of net-demand in a selected group of regions. Right: ALEs of the same covariates on a selected group of pairwise correlations.
to analyse how the predictive performance of the model depends on the set of candidate covariates that are considered by the model selection procedure. In Sections 3.4 and 3.5 we analyse this issue and we compare the proposed model with two non-Gaussian alternatives.

3.4 Validation on Regional Net-Demand Forecasting

Here we assess the predictive performance of several alternative models obtained via the selection procedure proposed in Section 3.2. In particular, we consider three sets of candidate effects. The first set (Full) includes all the effects appearing in (9), hence it corresponds to the model analysed in Section 3.3. Then we consider a calendar-only model (Cal), obtained by using only the first five effects in (9), that is from \( g_{j1}(t_i) \) to \( f_{j2}(tod_i) \), and a larger model (Cal+Ren) which also includes the effects of wind speed and solar irradiance. For each set of candidate effects, we use data from 2014 to 2017 to perform model selection, as described in Section 3.2. Most of the selected effects in Figure 3 appear on the diagonal, hence we consider also a Cal+Ren Diag model, which is based on the same candidate effects as Cal+Ren, but where gradient boosting is allowed to model only the diagonal elements of \( D \), while \( T \) is kept constant. See SM B.2 for more details on the models discussed above.

Having selected the model structure, we assess performance on 2018 data. This is done by first fitting each model to data up to the end of 2017 and then forecasting net-demand during January 2018. We then refit the models to data up to the 31st of January 2018 and forecast net-demand for February. By iterating this rolling forecasting origin procedure, we obtain day-ahead predictions covering the whole of 2018, for each model. To speed up computation, we first fit model (8) to the net-demand from each region using separate univariate Gaussian GAMs and then fit the corresponding residuals vectors using each of the covariance matrix models described above, via the methods described in Section 2.2.

We also include in the comparison three models where the marginal distribution of each GSP region’s net-demand is modelled with an increasing amount of flexibility. Two of these models are useful to check whether relaxing the Gaussian assumption improves predictions. In the gaulss+cop model, the net-demand of each region is modelled separately via a univariate location-scale Gaussian model. In shash+cop the net-demand of each region is modelled via the four-parameter sinh-arcsinh distribution of Jones and Pewsey (2009), which nests the Gaussian but allows for asymmetry and fat-tails. The shash+gpd+cop model produces the same predictions of shash+cop between quantiles 0.05 and 0.95, but uses a generalised Pareto distribution (GPD) beyond these. Having fitted the univariate models separately to each GSP group, we evaluate the corresponding conditional c.d.f.s to obtain uniform residuals. Then we use a static Gaussian copula to model the correlation structure of the resulting 14-dimensional residual vectors. See SM B.3 for more details.

As for the multivariate Gaussian models, the Gaussian, sinh-arcsinh and GPD models are fitted to the raw residuals (responses minus estimated mean) of Gaussian GAMs based
on formula (8), hence their location parameters are kept constant to avoid fitting a location-like parameter twice. The effects used to model the scale parameters of the gaulss+cop model are chosen from the same pool of candidates used for Cal+Ren, following the approach described in Section 3.2 but with the following modifications. We set all the elements of $T$ to zero which implies that the diagonal elements of $D$ are directly controlling the marginal variance of each GSP group’s net-demand (see Section 2.1). Then, as for Cal+Ren Diag, we allow gradient boosting to model only the elements of $D$. In this way, the models for the Gaussian scale parameter are selected to optimise the marginal fit to each region’s net-demand. The resulting effects are used to model the scale parameters of the sinh-arcsinh distribution as well, while the parameters controlling the skewness and kurtosis are modelled only via intercepts. This is because our attempts to manually select a model for them led to a worse performance than what is reported below. Each GPD model is fitted to only 5% of the data, hence we model its scale parameter using only a smooth effect of $\text{tod}_i$, while the shape parameter is kept constant.

We use the day-ahead multivariate predictions of each model to compute the performance metrics reported in Table 2. We consider the log score (i.e., the negative log-likelihood), the log score under independence (i.e., the sum of the negative marginal log-likelihoods of each GSP group), the marginal continuous ranked probability score (CRPS) and marginal pinball losses for quantiles 0.001 and 0.999 (each summed over the GSP groups, see Gneiting and Raftery (2007) for a detailed introduction to both losses), and the $p$-variogram score (Scheuerer and Hamill, 2015) with $p = 0.5$ and $p = 1$.

Considering Table 2, note that the predictive performance of Cal+Ren is superior to that of Cal on all scores, demonstrating the importance of including covariates that are strongly related to embedded renewable generation. The significance of the improvement is demonstrated by Figure 5a. Here, non-parametric bootstrapping with week-long blocks is used to quantify the variability of the differences in log scores between several pairs of models (SM B.4 provides analogous plots for the remaining scores). The boxplots show

| Model            | Log  | Log Ind | CRPS  | Pin 0.001 | Pin 0.999 | Var 0.5 | Var 1.0 |
|------------------|------|---------|-------|-----------|-----------|---------|---------|
| Cal              | -3669 | -326.8 | 2582  | 21.67     | 44.51     | 9515    | 10293   |
| Cal+Ren          | -4082 | -744.2 | 2571  | 20.24     | 33.97     | 9298    | 10012   |
| Full             | -4071 | -771.9 | 2570  | 19.25     | 34.07     | 9263    | 9965    |
| Cal+Ren Diag     | -3843 | -606.2 | 2574  | 20.97     | 35.45     | 9299    | 10014   |
| gaulss+cop       | -3786 | -646.0 | 2576  | 21.05     | 32.07     | 9387    | 10125   |
| shash+cop        | -3920 | -790.7 | 2574  | 20.66     | 30.31     | 9355    | 10097   |
| shash+gpd+cop    | -3975 | -821.6 | 2575  | 20.41     | 30.59     | 9362    | 10116   |

Table 2: Performance scores on 2018 test data, when forecasting the joint distribution of net-demand across the 14 GSP groups. The best score in each column is underlined.
that the gain obtained by including the effect of wind and solar irradiation is significant. Instead, extending the set of candidate effects as done in Full does not lead to any gain under the log score, which motivates our choice of basing the copula models on the same pool of candidate effects used for Cal+Ren.

Table 2 and Figure 5a show that modelling only the elements of D as done by Cal+Ren Diag leads to a worse performance, which highlights the importance of modelling T as well. Interestingly, gaulss+cop does worse than Cal+Ren Diag under the total log score, but better on the independent log score. As both models are Gaussian, this suggests that modelling the marginal variances directly and assuming that the correlation structure is constant, as done by gaulss+cop, leads to better marginal predictions but to a worse multivariate fit, relative modelling the D factor of the MCD parametrisation, as done by Cal+Ren Diag (recall that D affects the marginal variances as well as the correlation structure). Similarly, the shash+cop and shash+gpd+cop models have better marginal log scores than Cal+Ren (which uses the same pool of covariates), but worse total log and variogram scores. The fact that the CRPS loss, which is insensitive to tail behaviour (Taillardat et al., 2023), is better under the Cal+Ren model suggests that this model is

![Figure 5: a: Bootstrapped differences in the total log score between several pairs of models. Negative values mean that the first method is better than the second (e.g., Cal+Ren does better than Cal). b: Half-hourly marginal log losses of Cal+Ren (grey) along the test set (2018) and differences between the loss of Cal+Ren and that of shash+gpd+cop model (black). The black ticks mark the start and the end of the Beast from the East cold wave.](image-url)
better at predicting intermediate marginal net-demand quantiles while, as demonstrated also by the pinball scores, the non-Gaussian marginal models do better in the upper tail.

To verify this, Figure 5b shows the independent log score of the Cal+Ren model (grey line) and the differences between the Cal+Ren and shash+gpd+cop under the same score (black line). The largest gain of shash+gpd+cop relative to Cal+Ren occurs during the Beast from the East cold wave, which is delimited by the black ticks at the bottom of Figure 5b. All the models considered here struggle to forecast the effect of this weather extreme on net-demand, as the training data does not contain any cold wave of similar magnitude. The better performance of shash+gpd+cop under the independent log score is explained by the fact that this model has fatter tails (recall that the tails are modelled separately from the bulk of the net-demand distribution under this model). To check to what degree the results discussed here are affected by the Beast from the East, in SM B.4 we report a version of Table 2 obtained by excluding this exceptional period from the data. All the results discussed here are still valid but, as expected, excluding the cold wave reduces the benefit of adopting non-Gaussian marginal models.

### 3.5 Validation on Macro-Regional Net-Demand Forecasting

This work is motivated by the need for spatially coherent probabilistic net-demand forecasts in power flow studies and, as explained in Section 1, the transmission grid boundaries of interest in such studies vary depending on, for example, the status of the network. Hence, it is interesting to verify whether the results discussed above still hold when the forecast is post-processed to match the needs of an operationally relevant scenario. While considering realistic scenarios would require covering engineering aspects that are well beyond the scope of this work, in Section 1 we proposed aggregating the GSP regions into the five macro-regions shown in Figure 1, which were motivated by some of the boundaries of interest described in the 2021 National Grid’s Ten Year Statement (National Grid, 2021).

| Model      | Log Log | Log Ind | CRPS | Pin 001 | Pin 999 | Var 0.5 | Var 1.0 |
|------------|---------|---------|------|---------|---------|---------|---------|
| Cal        | 3236    | 4529.7  | 2008 | 14.84   | 42.77   | 2215    | 4871    |
| Cal+Ren    | 3091    | 4351.1  | 2000 | 13.63   | 34.73   | 2181    | 4782    |
| Full       | 3056    | 4337.9  | 1999 | 12.92   | 34.58   | 2170    | 4756    |
| Cal+Ren Diag | 3226   | 4469.3  | 2004 | 14.22   | 37.77   | 2188    | 4802    |
| gaulss+cop | 3184    | 4384.8  | 2003 | 14.64   | 36.59   | 2183    | 4788    |
| shash+cop  | 2002    | 14.26   | 35.40 | 2181    | 4785    |
| shash+gpd+cop | 2003  | 13.94   | 37.47 | 2183    | 4792    |

Table 3: Performance scores on 2018 test data, when forecasting the joint distribution of net-demand across the five GSP macro-regions. The best score in each column is underlined.
Joint macro-regional net-demand forecasts are easy to obtain because they are linear transformations of the regional forecasts. Table 3 shows the performance of each model when forecasting the joint distribution of macro-regional net-demand. Note that the scores are on a different scale due to the aggregation of net-demand (5 macro-regions vs 14 regions) and that the log scores have not been computed under two of the models, because the p.d.f. of linear combinations of correlated sinh-arcsinh-distributed random variables is, to our best knowledge, not analytically available. The results are similar to those obtained at regional level, but here the Full model produces the best forecasts under all the scores. Due to the high cost of operating power systems (and the volume of energy traded in wholesale markets), even marginal improvements in forecast performance and associated decision-making can yield substantial economic and operational benefits.

It is also possible to linearly transform the joint regional forecasts to obtain marginal probabilistic forecasts of differences in net-demand between regions or macro-regions. Such forecasts could be of particular interest in the context of power flow analyses focused on specific transmission grid boundaries. In SM B.4 we assess the performance of the models under three operationally motivated scenarios.

4 Conclusion

Forecasts of supply and demand are essential inputs to predict and manage power flows on electricity networks, as well as prices and other important variables. Given the imperative to maintain a reliable electricity supply, these predictions must enable risk to be quantified and managed. As the complexity of energy systems increases, the heuristic approaches widely used today are becoming inadequate and will have to be replaced by explicit probabilistic forecasts of power flows (Morales et al., 2014).

Motivated by the need for spatially coherent, probabilistic net-demand forecasts to support energy system operations, we have focused on joint day-ahead forecasting of net-demand across the GSP regions comprising GB’s transmission system. To accommodate for the dynamic nature of the net-demand covariance matrix, we let the elements of its MCD parametrisation vary with a number of temporal and weather-related covariates. To perform effect selection for a model comprising more than one hundred linear predictors, we leverage the interpretability of the chosen parametrisation and we combine it with a semi-automatic effect selection method, based on gradient-boosting. The results on the test set show that additive covariance matrix models significantly outperform, in terms of the total log, CRPS and variogram scores, two non-Gaussian models where the correlation matrix is static. However, the non-Gaussian models provide better predictions of extremely high quantiles, which suggests that a promising direction for future research might be adapting dynamic covariance matrix models to a non-Gaussian context.
A further direction for future work would be to extend the model presented here to capture temporal, in addition to spatial, dependencies. In particular, the covariance matrix models used here implicitly assume that regional net-demand residual vectors are uncorrelated in time. While the mean vector model (8) contains the effect of lagged net-demand, which is meant to capture part of the intra-regional temporal dependencies, more complex temporal effects could be captured by extending the covariance matrix to explicitly model the longitudinal nature of the data considered here. Such an extension should lead to models able to generate multivariate net-demand trajectories that are coherent both in space and in time, thus supporting important operations (e.g. determining the schedules for power-generating units) that must consider both spatial and temporal constraints.

Acknowledgements

MF was supported by Électricité de France, JB by the EPSRC (EP/R023484/1 and EP/R023484/2) and RB by the Departmental Strategic Plan (PSD) of the University of Udine, Department of Economics and Statistics (2022-2025).

References

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(4), 1059–1086.

Bonat, W. H. and B. Jørgensen (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5), 649–675.

Browell, J. and M. Fasiolo (2021). Probabilistic forecasting of regional net-load with conditional extremes and gridded NWP. *IEEE Transactions on Smart Grid* **12**(6), 5011–5019.

Browell, J., C. Gilbert, and M. Fasiolo (2022). Covariance structures for high-dimensional energy forecasting. *Electric Power Systems Research* **211**, 108446.

Capezza, C., B. Palumbo, Y. Goude, S. N. Wood, and M. Fasiolo (2021). Additive stacking for disaggregate electricity demand forecasting. *The Annals of Applied Statistics* **15**(2), 727–746.

Fan, S. and R. J. Hyndman (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems* **27**(1), 134–141.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(5), 1189–1232.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Griewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation* (2 ed.). Philadelphia, PA, USA: SIAM.

Hans, N., N. Klein, F. Faschingbauer, M. Schneider, and A. Mayr (2023). Boosting distributional copula regression. *Biometrics* **79**(3), 2298–2310.

Hastie, T. and R. Tibshirani (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association* **82**(398), 371–386.

Huxley, O., J. Taylor, A. Everard, J. Briggs, K. Tilley, J. Harwood, and A. Buckley (2022). The uncertainties involved in measuring national solar photovoltaic electricity generation. *Renewable and Sustainable Energy Reviews* **156**, 112000.

Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. *Biometrika* **96**(4), 761–780.

Klein, N., T. Hothorn, L. Barbanti, and T. Kneib (2022). Multivariate conditional transformation models. *Scandinavian Journal of Statistics* **49**(1), 116–142.

Klein, N., T. Kneib, S. Klasen, and S. Lang (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(4), 569–591.

Marra, G. and R. Radice (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis* **112**, 99–113.

Morales, J. M., A. J. Conejo, H. Madsen, P. Pinson, and M. Zugno (2014). *Integrating Renewables in Electricity Markets: Operational Problems*, Volume 205. New York, NY, USA: Springer.

Muschinski, T., J. G. Mayr, T. Simon, N. Umlauf, and A. Zeileis (2024). Cholesky-based multivariate Gaussian regression. *Econometrics and Statistics* **29**, 261–281.

National Grid (2018). Regional trends and insights. [https://www.nationalgrideso.com/document/118656/download](https://www.nationalgrideso.com/document/118656/download).

National Grid (2021). Electricity ten year statement. [https://www.nationalgrideso.com/document/223046/download](https://www.nationalgrideso.com/document/223046/download).
Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association* **83**(404), 1134–1143.

Pinheiro, J. C. and D. M. Bates (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* **6**(3), 289–296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**(3), 677–690.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science* **26**(3), 369–387.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 507–554.

Scheuerer, M. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review* **143**(4), 1321–1334.

Strömer, A., N. Klein, C. Staerk, H. Klinkhammer, and A. Mayr (2023). Boosting multivariate structured additive distributional regression models. *Statistics in Medicine* **42**(11), 1779–1801.

Strömer, A., C. Staerk, N. Klein, L. Weinhold, S. Titze, and A. Mayr (2022). Deselection of base-learners for statistical boosting—with an application to distributional regression. *Statistical Methods in Medical Research* **31**(2), 207–224.

Taillardat, M., A.-L. Fougeres, P. Naveau, and R. De Fondeville (2023). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting* **39**(3), 1448–1459.

Thomas, J., A. Mayr, B. Bischl, M. Schmid, A. Smith, and B. Hofner (2018). Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing* **28**(3), 673–687.

Tuinema, B. W., J. L. Rueda Torres, A. I. Stefanov, F. M. Gonzalez-Longatt, and M. A. M. M. van der Meijden (2020). Probabilistic power flow analysis. In *Probabilistic Reliability Analysis of Power Systems*, pp. 179–208. Cham, CH: Springer.

Vatter, T. and T. Nagler (2018). Generalized additive models for pair-copula constructions. *Journal of Computational and Graphical Statistics* **27**(4), 715–727.

von Meier, A. (2006). *Electric Power Systems: A Conceptual Introduction*. Hoboken, NJ, USA: John Wiley & Sons.
Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R.* (2 ed.). Boca Raton, FL, USA: Chapman & Hall/CRC.

Wood, S. N. and M. Fasiolo (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* 73(4), 1071–1081.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111(516), 1548–1563.
Supplementary Material to “Additive Covariance Matrix Models: Modelling Regional Electricity Net-Demand in Great Britain”

Note: The numbering of the tables and figures shown below follows from the main text. For example, when we mention Figure 1 below, we refer to Figure 1 in the main text. The first figure in the Supplementary Material below is Figure 6 because the main text contains 5 figures.

A Derivatives of the Log-Likelihood

A.1 Setting up the Notation

Consider a scalar-valued function $f$ of the $n$-dimensional vectors $\eta_1, \ldots, \eta_q$. We indicate with $f^k$ and $f^{kij}$ the vectors with $i$-th elements

$$f^k = \frac{\partial f}{\partial \eta_k} \quad \text{and} \quad f^{kij} = \frac{\partial^2 f}{\partial \eta_{ik} \partial \eta_{ij}},$$

where $\eta_{ik}$ indicates the $i$-th element of $\eta_k$. Each $\eta_k$ is a function of a corresponding $p_k$-dimensional vector $\beta_k$. For the derivatives of $f$ w.r.t. the elements of $\beta_k$, we use the compact notation

$$f^{kr} = \frac{\partial f}{\partial \beta_{kr}} \quad \text{and} \quad f^{krjs} = \frac{\partial^2 f}{\partial \beta_{kr} \partial \beta_{js}},$$

where $\beta_{kr}$ indicates the $r$-th element of $\beta_k$. Finally, we denote with $f^{\beta} = \nabla_{\beta} f$ the gradient of $f$ w.r.t. $\beta$ and with $f^{\beta \beta} = \nabla_{\beta}^\top \nabla_{\beta} f$ the matrix of second derivatives.

A.2 Gradient and Hessian w.r.t. $\beta$

To simplify the notation, let us indicate with $y$ the collection of all response vectors $y_1, \ldots, y_n$, and define $L(\beta) = \log p(\beta | y, \lambda)$. Recall that

$$L(\beta) = \sum_{i=1}^n \ell_i - \frac{1}{2} \beta^\top \bar{S}^\lambda \beta$$

where $\ell_i = \log p(y_i | \beta)$. The gradient and Hessian of the log-posterior w.r.t $\beta$ are

$$L^\beta(\beta) = \sum_{i=1}^n \ell_i^{\beta} - \bar{S}^\lambda \beta \quad \text{and} \quad L^{\beta \beta}(\beta) = \sum_{i=1}^n \ell_i^{\beta \beta} - \bar{S}^\lambda.$$
Let us define $\bar{\ell} = \sum_{i=1}^{n} \ell_i$. To provide formulas for $\bar{\ell}^{\beta}$ and $\bar{\ell}^{\beta\beta}$, let us assume that $\beta = (\beta_1^T, \ldots, \beta_q^T)^T$, where $\beta_j$ is the vector of regression coefficients specific to the $j$-th linear predictor, that is $\eta_j = X_j^T \beta_j$ where $X_j$ is an $n \times p_j$ model matrix. With this notation, the $j$-th sub-vector of $\bar{\ell}^{\beta}$ is

$$\bar{\ell}^{\beta_j} = X_j^T \ell \eta_j,$$

while the $j, k$-th block of the Hessian is

$$\bar{\ell}^{\beta_k \beta_j} = (\bar{\ell}^{\beta_k \beta_j})^T = X_k^T \text{diag}(\ell \eta_j \eta_k) X_j,$$

where $\text{diag}(\cdot)$ is the vector-to-matrix diagonal operator. The formulas provided so far apply to any GAM with multiple linear predictors and independent response vectors. In contrast, the expressions for $\ell^{\eta_j}$ and $\ell^{\eta_j \eta_k}$ are model-specific and are provided in the following section for a multivariate Gaussian distribution, with covariance matrix parametrised via the MCD.

### A.3 Derivatives w.r.t. $\eta$

Let us start by defining a few useful quantities. Let $G$ be a $((d-1)\times(d-1))$ lower triangular matrix such that $G_{jk} = C_{jk} + 2d \mathbb{1}_{k \leq j}$, where

$$C_{jk} = \begin{cases} \frac{j+1}{2} & k = j \\ C_{j(k+1)} - 1 & k < j \\ 0 & k > j \end{cases}$$

and $\mathbb{1}$ is the indicator function. Define the $((d-1)\times(d-1))$ lower triangular matrices $Z$ and $W$ such that $Z_{jk} = k \mathbb{1}_{k \leq j}$ and $W_{jk} = (j+1) \mathbb{1}_{k \leq j}$. Let $z = \text{rvech}(Z)$ and $w = \text{rvech}(W)$, where $\text{rvech}(\cdot)$ is the row-wise half-vectorisation operator, that is $\text{rvech}(Z) = (Z_{11}, Z_{21}, Z_{22}, Z_{31}, Z_{32}, Z_{33}, \ldots, Z_{(d-1)(d-1)})^T$. Let $Q_l$, for $l = 1, \ldots, d$, and $P_l$, for $l = 1, \ldots, d(d-1)/2$, be $d \times d$ matrices such that $(Q_l)_{ll} = e^{-\eta_{l+d}}$ and $(P_l)_{zul} = 1$, while all other elements are equal to zero.

Here, index $i$ is not needed, hence we drop it and we indicate the $i$-th log-likelihood component $\ell_i$ simply with $\ell$. Note that, given that we are focusing on an individual $i$, here $\eta$ is a $q$-dimensional vector and $q = d + d(d+1)/2$. If we omit the constants that do not depend on $\eta$ and we indicate with $r_k$ the $k$-th element of the residual vector, $r = y - \mu$, the Gaussian log-density can be written

$$\ell = -\frac{1}{2} \left\{ \text{tr}(\log D^2) + r^T T^T D^{-2} T r \right\}$$

$$= -\frac{1}{2} \sum_{j=1}^{d} \left\{ \eta_{j+d} + e^{-\eta_{j+d}} \left( \sum_{k=1}^{j-1} \eta_{G_{(j-1)k}} r_k + r_j \right)^2 \right\},$$

2
where we used \( \log |\Sigma| = \text{tr}(\log D^2) = \sum_{j=1}^{d} \eta_{j+d} \) and we implicitly assumed that the sum \( \sum_{k=1}^{j-1} \) should not be computed when \( j = 1 \) (we will use the same convention in several places below). Similarly, below we assume that \( \sum_{j=l+1}^{d} \) will not be computed when \( l = d \). Here we provide the first and second derivatives of \( \ell \) w.r.t. \( \eta \) both in compact matrix form and in an extended format, the latter being more useful for efficient numerical implementation.

With notation above, the elements of \( \ell_\eta = (\ell_\eta^1, \ldots, \ell_\eta^q)^\top = (\partial \ell / \partial \eta_1, \ldots, \partial \ell / \partial \eta_d)^\top \) are

\[
\ell_\eta^l = (T^\top D^{-2}\text{Tr})_l
= e^{-\eta_{l+d}} \left( \sum_{k=1}^{l-1} \eta_{G(l-1)k} r_k + r_l \right) + \sum_{j=l+1}^{d} e^{-\eta_{j+d}} \left( \sum_{k=1}^{j-1} \eta_{G(j-1)k} r_k + r_j \right) \eta_{G(j-1)l} ,
\]

for \( l = 1, \ldots, d \),

\[
\ell_\eta^l = \frac{1}{2} r^\top T^\top Q_{l-d} \text{Tr} - \frac{1}{2} e^{-\eta_l} \left( \sum_{k=1}^{l-d-1} \eta_{G(l-d-1)k} r_k + r_{l-d} \right)^2 - \frac{1}{2} ,
\]

for \( l = d + 1, \ldots, 2d \), and

\[
\ell_\eta^l = -r^\top P_{l-2d} D^{-2} \text{Tr}
= -e^{\eta_{l-2d+d}} \left( \sum_{k=1}^{w_{l-2d-1}} \eta_{G(w_{l-2d-1})k} r_k + r_{w_{l-2d}} \right) z_{l-2d} ,
\]

for \( l = 2d + 1, \ldots, q \).

The elements forming the upper triangle of \( \ell_\eta^m \) (here \( \ell_\eta^m = \partial^2 \ell / \partial \eta_l \partial \eta_m \)), are

\[
\ell_\eta^m = - (T^\top D^{-2} T)_{lm}
= - \left\{ e^{-\eta_{l+d}} + \sum_{k=l+1}^{d} e^{-\eta_{k+d}} \left( \eta_{G(k-1)l} \right)^2 \right\} \mathbb{1}_{\{m=l\}}
\]

\[
- e^{-\eta_{m+d}} \eta_{G(m-1)l} + \sum_{k=m+1}^{d} e^{-\eta_{k+d}} \eta_{G(k-1)l} \eta_{G(k-1)m} \mathbb{1}_{\{m>l\}} ,
\]

for \( l = 1, \ldots, d \), and \( m = l, \ldots, d \),

3
\[
\ell_{lm}^m = - (T^\top Q_{m-d} Tr)^l \\
\quad = -e^{-\eta_m} \left\{ \left( \sum_{k=1}^{l-1} \eta_{G(l-1)_k} r_k + r_l \right) \mathbb{1}_{\{m-d=l\}} \right. \\
\left. + \left( \sum_{k=1}^{m-d-1} \eta_{G(m-d-1)_k} r_k + r_{m-d} \right) \eta_{G(m-d-1)_l} \mathbb{1}_{\{m-d>l\}} \right\} ,
\]
for \( l = 1, \ldots, d \), and \( m = d+1, \ldots, 2d \),

\[
\ell_{lm}^m = (P_{m-2d} D^{-2} Tr + T^\top D^{-2} P_{m-2d} r)^l \\
\quad = e^{-\eta_{m-2d+l-d}} \left\{ r_{z_{m-2d}} \left( \mathbb{1}_{\{w_{m-2d}=l\}} + \eta_{G(w_{m-2d-1})} \mathbb{1}_{\{w_{m-2d}>l\}} \right) \right. \\
\left. + \left( \sum_{k=1}^{w_{m-2d}-1} \eta_{G(w_{m-2d-1})} r_k + r_{w_{m-2d}} \right) \mathbb{1}_{\{z_{m-2d}=l\}} \right\} ,
\]
for \( l = 1, \ldots, d \), and \( m = 2d+1, \ldots, q \),

\[
\ell_{lm}^m = -\frac{1}{2} r^\top T^\top Q_{l-d} Tr \\
\quad = -\frac{1}{2} e^{-\eta} \left( \sum_{k=1}^{l-d-1} \eta_{G(l-d-1)_k} r_k + r_{l-d} \right)^2 \mathbb{1}_{\{m=l\}} ,
\]
for \( l = d+1, \ldots, 2d \), and \( m = l, \ldots, 2d \),

\[
\ell_{lm}^m = r^\top P_{m-2d} Q_{l-d} Tr \\
\quad = e^{-\eta} \left( \sum_{k=1}^{l-d-1} \eta_{G(l-d-1)_k} r_k + r_{l-d} \right) r_{z_{m-2d}} \mathbb{1}_{\{w_{m-2d}=l-d\}} ,
\]
for \( l = d+1, \ldots, 2d \), and \( m = 2d+1, \ldots, q \), and finally

\[
\ell_{lm}^m = -r^\top P_{l-2d} D^{-2} P_{m-2d} r \\
\quad = -e^{-\eta_{w_{l-2d}}} + r_{z_{l-2d}} r_{z_{m-2d}} \mathbb{1}_{\{w_{m-2d}=w_{l-2d}\}} ,
\]
for \( l = 2d+1, \ldots, q \), and \( m = l, \ldots, q \).
A.4 Details on the Accumulated Local Effects

Recall that $\Sigma_i$ depends on the covariate vector $x_i$ via the linear predictor vector $\eta_i$. By omitting the subscript $i$ and expliciting the dependence on $x$, let us denote with $\omega(x)$ a generic element of $\Sigma$ or of the correlation matrix $\Gamma$, the elements of the latter being defined by $\Gamma_{jk} = \Sigma_{jk} / \sqrt{\Sigma_{jj} \Sigma_{kk}}$. Assuming that $\omega(x)$ is differentiable w.r.t. the $k$-th covariate, the main (first-order) accumulated local effects (ALEs) of $x_k$ is

$$\bar{\omega}_k(x) = \int_{x_k}^{x} \mathbb{E}_{x \setminus k} \{ \omega^k(z, x \setminus k)|x_k = z \} dz - \text{constant},$$

where $x \setminus k$ is $x$ with the $k$-th element excluded, $\omega^k = \partial \omega / \partial x_k$ and $\mathbb{E}_{x \setminus k}\{\cdot |x_k = z\}$ is the conditional expectation w.r.t. $p(x \setminus k | x_k = z)$. The choice of $x_k^{\text{min}}$ is unimportant, as changing it simply shifts the effect vertically, so in practice $x_k^{\text{min}}$ is set to just below the smallest observed value of $x_k$.

An uncentered first-order ALE is obtained by setting the constant term to zero, while a centered ALE has a mean equal to zero when averaged over the observed values of the covariate of interest. This is explained by Apley and Zhu (2020), who also provide formulas for obtaining estimated effects $\hat{\omega}_k(x)$, by approximating the integral above. For instance, consider an uncentered ALE and let $x_{ik}$ be the $i$-th observed value of $x_k$. Further, denote with $z_{0k}, \ldots, z_{Bk}$ a grid of values along $x_k$, with $z_{0k} = \min_{i=1,\ldots,n} x_{ik}$ and $z_{Bk} = \max_{i=1,\ldots,n} x_{ik}$, and let $n_k(1), \ldots, n_k(B)$ the number of $x_{ik}$ included, respectively, in the intervals $N_k(1) = [z_{0k}, z_{1k}), \ldots, N_k(B) = [z_{B-1k}, z_{Bk}]$. Then, the ALE of $x_k$ is estimated via

$$\hat{\omega}_k(x) = \sum_{v=1}^{v_k(x)} \frac{1}{n_k(v)} \sum_{\{i, x_{ik} \in N_k(v)\}} \{ \omega(z_{vk}, x_{i \setminus k}) - \omega(z_{v-1k}, x_{i \setminus k}) \}$$

with $\hat{\omega}_k(z_{0k}) = 0$ and $v_k(x) \in \{1, \ldots, B\}$ denoting the bin number to which an arbitrary value $x$ of $x_k$ belongs.

The uncertainty of ALEs can be quantified by propagating posterior parameter uncertainty via a standard asymptotic approximation. Recall that standard Bayesian asymptotics justify approximating $p(\beta | y, \lambda)$ with a Gaussian distribution, $N(\hat{\beta}, \mathbf{V}_{\hat{\beta}})$, centered at the MAP estimate and with covariance matrix $\mathbf{V}_{\hat{\beta}} = -\mathbf{H}^{-1}$, where $\mathbf{H}$ is the negative Hessian of $\log p(\beta | y, \lambda)$, evaluated at $\hat{\beta}$. Capezza et al. (2021) show that, in a GAM context, the delta method can be used to approximate the posterior variance of ALEs via

$$\text{var}\{\hat{\omega}_k(x)\} \approx \nabla_{\hat{\beta}}^T \hat{\omega}_k \mathbf{V}_{\hat{\beta}} \nabla_{\hat{\beta}} \hat{\omega}_k.$$
Suppose that the output of interest, \( \omega(x) \), is an element of \( \Sigma \). Define the vector \( \boldsymbol{\sigma}_{jk} = \{(\Sigma_1)_{jk}, \ldots, (\Sigma_n)_{jk}\}^\top \) and consider the vector containing the values of the \( j \)-th linear predictor at each observation, that is \( \eta_j = (\eta_{1j}, \ldots, \eta_{nj})^\top \). For \( j = 1, \ldots, q \), we have that \( \eta_j = \mathbf{X}^j \beta_j \) where \( \mathbf{X}^j \) and \( \beta_j \) are, respectively, the \( n \times p_j \) model matrix and the \( p_j \)-dimensional vector of regression coefficients belonging to the \( j \)-th linear predictor. The \( n \times p \) Jacobian matrix of \( \sigma_{jk} \) w.r.t. \( \beta \) is

\[
J_{jk} = \nabla_\beta \sigma_{jk} = (\nabla_{\beta_1} \sigma_{jk}, \ldots, \nabla_{\beta_q} \sigma_{jk})^\top,
\]

where \( p = \sum_{j=1}^q p_j \). The \( a \)-th block of \( J_{jk} \) is

\[
\nabla_{\beta_a} \sigma_{jk} = \nabla_{\eta_a} \sigma_{jk} \nabla_{\beta_a} \eta_a = \nabla_{\eta_a} \sigma_{jk} \mathbf{X}^a,
\]

for \( a = 1, \ldots, q \), and where \( \nabla_{\eta_a} \sigma_{jk} \) is an \( n \times n \) diagonal matrix with non-zero elements

\[
\left( \nabla_{\eta_a} \sigma_{jk} \right)_{ii} = \frac{\partial (\Sigma_i)_{jk}}{\partial \eta_{ia}}.
\]

Note \( \partial (\Sigma_i)_{jk} / \partial \eta_{ia} \) is the only parameterisation-dependent component of the Jacobian, thus in Section A.4.1 we provide the relevant formulas. When \( \omega(x) \) is an element of \( \Gamma \), the posterior variance of the ALEs is approximated similarly, since the Jacobian of \( \{(\Gamma_1)_{jk}, \ldots, (\Gamma_n)_{jk}\} \) w.r.t. \( \beta \) is computed analogously, but with \( \partial (\Gamma_i)_{jk} / \partial \eta_{ia} \) in place of \( \partial (\Sigma_i)_{jk} / \partial \eta_{ia} \). Formulas for \( \partial (\Gamma_i)_{jk} / \partial \eta_{ia} \) are provided in Section A.4.2.

### A.4.1 Derivatives of \( \Sigma \) w.r.t. \( \eta \)

Here the index \( i \) is not needed, hence we drop it. Consider the factorisation \( \Sigma = \mathbf{R} \mathbf{R}^\top \) where \( \mathbf{R} = \mathbf{L} \mathbf{D} \) and \( \mathbf{L} = \mathbf{T}^{-1} \). The partial derivative of the \((l, m)\) element of \( \Sigma \) w.r.t. \( \eta_j \), is

\[
\frac{\partial \Sigma_{lm}}{\partial \eta_j} = \sum_{k=1}^d \left( \frac{\partial \mathbf{R}_{lk}}{\partial \eta_j} \mathbf{R}_{mk} + \mathbf{R}_{lk} \frac{\partial \mathbf{R}_{mk}}{\partial \eta_j} \right),
\]

where

\[
\frac{\partial \mathbf{R}_{lk}}{\partial \eta_j} = 0, \quad \text{for} \quad j = 1, \ldots, d,
\]

\[
\frac{\partial \mathbf{R}_{lk}}{\partial \eta_j} = \frac{\partial \mathbf{L}_{lk} \mathbf{D}_{kk}}{\partial \eta_j} = \frac{1}{2} \mathbf{L}_{l(j-d)(j-d)} \mathbf{D}_{(j-d)(j-d)} \mathbf{1}_{(j-d)(j-d)} \text{, for} \quad j = d+1, \ldots, 2d,
\]

and

\[
\frac{\partial \mathbf{R}_{lk}}{\partial \eta_j} = \frac{\partial \mathbf{L}_{lk} \mathbf{D}_{kk}}{\partial \eta_j} = -\mathbf{L}_{ls} \mathbf{L}_{lk} \mathbf{D}_{kk}, \quad \text{for} \quad j = 2d+1, \ldots, q,
\]
with \( t \) and \( s \) being the row and column indices of the only element of \( T \) that depends on \( \eta_j \), for \( j = 2d + 1, \ldots, q \). Hence, we obtain
\[
\frac{\partial \Sigma_{lm}}{\partial \eta_j} = 0, \quad \text{for } j = 1, \ldots, d,
\]
\[
\frac{\partial \Sigma_{lm}}{\partial \eta_j} = L_{l(j-d)}L_{m(j-d)}D_{(j-d)(j-d)}, \quad \text{for } j = d + 1, \ldots, 2d,
\]
and
\[
\frac{\partial \Sigma_{lm}}{\partial \eta_j} = -L_{ls}(2\sum_{k=1}^{d} L_{tk}D_{kk}L_{mk} - L_{ms} \sum_{k=1}^{d} L_{tk}D_{kk}L_{tk}), \quad \text{for } j = 2d + 1, \ldots, q,
\]
where \( t \) and \( s \) are defined as above.

### A.4.2 Derivatives of \( \Gamma \) w.r.t. \( \eta \)

To simplify the notation indicate \((\Sigma_{ll})^{-1/2}\) with \( \Sigma_{ll}^{-1/2} \). The \((l, m)\) element of \( \Gamma \) is
\[
\Gamma_{lm} = \Sigma_{ll}^{-1/2} \Sigma_{lm} \Sigma_{mm}^{-1/2}, 
\]
and its partial derivative w.r.t. \( \eta_j \), for \( j = 1, \ldots, q \), is
\[
\frac{\partial \Gamma_{lm}}{\partial \eta_j} = \Sigma_{ll}^{-1/2} \frac{\partial \Sigma_{lm}}{\partial \eta_j} \Sigma_{mm}^{-1/2} - \frac{1}{2} \frac{\partial \Sigma_{lm}}{\partial \eta_j} \left\{ \Sigma_{ll}^{-3/2} \Sigma_{mm}^{-1/2} \frac{\partial \Sigma_{ll}}{\partial \eta_j} + \Sigma_{ll}^{-1/2} \Sigma_{mm}^{-3/2} \frac{\partial \Sigma_{mm}}{\partial \eta_j} \right\},
\]
and where the derivatives of the elements of \( \Sigma \) w.r.t. \( \eta_j \) are provided in Section A.4.1.

### B Further Details and Results

#### B.1 Details on the Model Selection Approach

In the following we provide more details on the model selection approach described in Section 3.2.

Denote with \( R_j \) a vector containing the indices of the candidate effects in (9), with the first element referring to both \( g_{j1} \) and \( g_{j2} \), so that these two terms effectively form a single effect in the model selection process. Let \( \eta_j \) be the \( j \)-th linear predictor and indicate with \( \eta \) the \( n \times q \) matrix containing all the linear predictors. Let \( X^r \), with \( r \in R_j \), be the model matrix of the \( r \)-th effect, and let \( S_r \) be the corresponding positive semi-definite
penalty matrix. Let $\Delta$ be a list of length $q$, its $j$-th element $\Delta_j$ being a vector of dimension $\text{card}(R_j)$ with all elements, $\Delta_{rj}$, initialised at 0.

Algorithm 1 details the steps of the gradient boosting procedure used to quantify the importance of each candidate effect-linear predictor pair. In particular, its output is $\Delta$, the list containing the cumulative log-likelihood gains achieved by each candidate effect-linear predictor pair. Note that in step 1.II.(a) we regress the log-likelihood gradient $u_j$ on the model matrix of each effect in $R_j$ using penalised least squares, with penalty $\zeta_r S_r$, where $\zeta_r > 0$. As explained by Hofner et al. (2011), penalisation is helpful to mitigate the selection bias in favour of effects with more parameters. In particular, the penalties of each effect should be scaled to make sure that all the effects have similar effective degrees of freedom. The latter are defined as

$$\text{edf}_r = \text{tr}\{X^r (X^r\top X^r + \zeta_r S_r)^{-1} X^r\top\},$$

and depend on $\zeta_r$. In particular, assuming that $X^r$ is of full rank $p_r$ and that $S_r$ has rank $s_r \leq p_r$, as $\zeta_r$ increases from zero to infinity the edf$_r$ decrease from $p_r$ to $p_r - s_r$.

For each effect appearing in the MCD model (9) such that $p_r > 4$, we choose $\zeta_r$ such that edf$_r = 4$. This is a one-dimensional numerical optimisation problem, which can be solved very rapidly prior to running Algorithm 1. The effect of progressive time $t_i$ is modelled using two parameters, hence penalisation is unnecessary. The penalty matrices $S_r$ correspond to cubic splines penalties (i.e., proportional to the integrated second derivative of the effect, $\int f''(x)^2 dx$) for all the smooth effects in model (9), while a standard ridge penalty is used for the effect of the factor dow$_i$. 
Algorithm 1 Quantifying the effects’ importance via gradient boosting

1: For $j = d + 1, \ldots, q$,

I. Compute the gradient $u_j$ of the log-likelihood w.r.t. $\eta_j$, that is

$$u_{ij} = \frac{\partial \log p(y_i | \eta_i)}{\partial \eta_{ij}}, \quad \text{for} \quad i = 1, \ldots, n,$$

where $\eta_i$ is the $i$-th row of $\eta$.

II. For $r \in \mathcal{R}_j$

(a) Regress $u_j$ on $X^r$ via

$$\hat{u}^r_j = X^r (X^r\top X^r + \zeta_r S_r)^{-1} X^r\top u_j.$$

(b) Update the corresponding linear predictor via $\tilde{\eta}_j^r = \eta_j + \nu \hat{u}^r_j$, where $\nu = 0.1$ is the learning rate. Let $\tilde{\eta}^r_j$ be the same as $\eta$ matrix, but with the $j$-th column set to $\tilde{\eta}_j^r$.

(c) Compute the corresponding change in log-likelihood

$$\delta_{rj} = \sum_{i=1}^{n}\{\log p(y_i | \tilde{\eta}^r_j) - \log p(y_i | \eta_i)\}.$$

2: Let $j^*$ and $r^*$ be the indices corresponding to the largest $\delta_{rj}$. Update the relevant linear predictor and the cumulative gain vector by doing

$$\eta_j \leftarrow \tilde{\eta}_j^{r^*}, \quad \text{and} \quad \Delta_{r^*j^*} \leftarrow \Delta_{r^*j^*} + \delta_{r^*j^*}.$$

3: Unless the maximum number of iterations $M$ has been reached, go back to step 2.
B.2 Details on the Multivariate Gaussian Models

As explained in Section 3.4, the Full, Cal+Ren, Cal and Cal+Ren Diag models are based on a conditional multivariate Gaussian distribution, parametrised via the MCD. The purpose of this section is to provide a self-contained introduction to these models.

Each element of the mean vector is modelled via formula (8), which we repeat here for convenience

\[ \eta_{ij} = g_{j1}(t_i) + g_{j2}(t_i^2) + g_{j3}(dow_{ij}) + g_{j4}(shol_{ij}) + g_{j5}(y_{ij}^{24}) + g_{j6}(wsp_{ij}^{10}) \\
+ f_{j1}^{10}(doy_i) + f_{j2}^{35}(tod_i) + f_{j3}^{10}(n2ex_i) + f_{j4}^{35}(temp_{ij}) + f_{j5}^{35}(temp_{ij}^5) + f_{j6}^{10}(rain_{ij}^{1/2}) \\
+ wcap_i \times f_{j7}^{20}(wsp_{ij}^{100}) + f_{j8}^{55}(irm_{ij}) + f_{j9}^{39}(tod_i, dow_{ij}) + f_{j10}^{20}(tod_i, shol_{ij}) \\
+ f_{j11}^{55}(n2ex_i, tod_i) + f_{j12}^{55}(temp_{ij}, tod_i) + f_{j13}^{55}(rain_{ij}^{1/2}, tod_i) + f_{j14}^{10,10}(doy_i, tod_i) , \]

for \( j = 1, \ldots, 14 \). Note that, given that the mean model is fitted in a preliminary step to save computational time (see Section 3.4), the mean vector fit is exactly the same for each of the models.

For the Full model, each element of the MCD parametrisation of the covariance matrix can be modelled via any of the effects appearing in formula (9), which we repeat here

\[ \eta_{ij} = \bar{\eta}_{ij} + g_{j1}(t_i) + g_{j2}(t_i^2) + g_{j3}(dow_{ij}) + f_{j1}^{10}(doy_i) + f_{j2}^{10}(tod_i) + \\
+ wcap_i \times f_{j4}^{55}(wsp_{ij}^{100}) + f_{j5}^{55}(irm_{ij}) + f_{j6}^{55}(temp_{ij}) + f_{j7}^{55}(rain_{ij}^{1/2}) + f_{j8}^{55}(n2ex_i) , \]

for \( j = 15, \ldots, 119 \), and where we indicate with \( l_j \) the row of \( D_l \) or \( T_l \) on which the \( j \)-th linear predictor appears. Model selection Algorithm 1 (see SM B.1) evaluates, at each step, the gain obtained by adding one of the effects appearing in the formula above to one of the non-zero elements of the \( T \) and \( D \) factors forming the MCD factorisation. Under the Cal+Ren model, the search is restricted to the effects appearing in the formula

\[ \eta_{ij} = \bar{\eta}_{ij} + g_{j1}(t_i) + g_{j2}(t_i^2) + g_{j3}(dow_{ij}) + f_{j1}^{10}(doy_i) + f_{j2}^{10}(tod_i) + \\
+ wcap_i \times f_{j4}^{55}(wsp_{ij}^{100}) + f_{j5}^{55}(irm_{ij}) , \]

for \( j = 15, \ldots, 119 \) while, for the Cal model, the search is further restricted to the effects in

\[ \eta_{ij} = \bar{\eta}_{ij} + g_{j1}(t_i) + g_{j2}(t_i^2) + g_{j3}(dow_{ij}) + f_{j1}^{10}(doy_i) + f_{j2}^{10}(tod_i) , \]

for \( j = 15, \ldots, 119 \). Hence, in the Full, Cal+Ren and Cal models, all the MCD elements are considered but the number of candidate effects is increasingly restricted.
Under the Cal+Ren Diag model, the pool of candidate effects is the same as for Cal+Ren, that is those appearing in the formula
\[
\eta_{ij} = \bar{\eta}_{ij} + g_{j1}(t_i) + g_{j2}(t^2_i) + g_{j3}(dow_i) + f^{10}_{j1}(doy_i) + f^{10}_{j2}(tod_i) + \]
\[+ wcap_i \times f^5_{j3}(wsp_{100}^i) + f^5_{j4}(irr^i),
\]
but \(j\) is restricted to take value in 15, 16, \ldots, 28. That is, Algorithm 1 evaluates the gain obtained by adding an effect to a diagonal element of the \(D\) factor, while the elements of the \(T\) factor are not allowed to vary with the covariates (i.e., each of the corresponding linear predictors contains only an intercept).

### B.3 Details on the Implementation of the Copula-Based Models

The gaulss+cop, shash+cop and shash+gpd+cop GAMLSS models (Rigby and Stasinopoulos, 2005) from Section 3.4 model the marginal distribution of regional net-demand separately from the correlation structure. Here we give more details on the structure of these models and on how they are fitted to data.

Consider the shash+cop model. Under this model, the net-demand from \(j\)-th region follows a conditional sinh-arcsinh distribution (Jones and Pewsey, 2009), controlled by the four-dimensional parameter vector \(\theta_{ij}\). Each element of \(\theta_{ij}\) could potentially be modelled additively but, for the reasons put forward in Section 3.4, we let only the log-scale parameter depend on the covariates, and control the location, asymmetry and tail parameters only via intercepts. The gaulss+cop model uses a two-parameter conditional Gaussian model, where the location parameter is kept constant while the log-scale parameter is modelled additively, as explained in Section 3.4.

The predictions of the shash+gpd+cop model are based on the same conditional sinh-arcsinh distribution of shash+cop between quantiles 0.05 and 0.95, and on two separate models based on the generalised Pareto distribution (GPD) beyond them. In particular, let \(F_j(\cdot|x_i)\) be the conditional c.d.f. of the sinh-arcsinh model corresponding to the \(i\)-th observation and the \(j\)-th GSP group. Then \(q^{95}_{ij} = F_j^{-1}(0.95|x_i)\) is an estimate of the 95th conditional net-demand percentile under this model. A two-parameter (scale and location) conditional GPD GAMLSS model is fitted to the observed net-demand values that fall above \(q^{95}_{ij}\) and a separate GPD model is fitted to those falling below \(q^{5}_{ij}\). Given that each of these models is fitted to only around 5% of the data, we model only the log-scale parameter using a smooth effect of the time of day, while the shape parameter is controlled only by an intercept. Having fitted the sinh-arcsinh and the two GPD models, we build a composite model where the sinh-arcsinh model is used to produce predictions between quantiles 0.05 and 0.95, while the GPD models are used to produce tail predictions.

When fitting the GPD models we impose (via a simple reparametrisation) the constraint that the shape parameter of the distribution, \(\xi\), must be larger than 0.001. The reason
is that, without this constraint, the estimated shape parameter would in some cases go negative, with disastrous consequences on the test-set performance of the GPD. To see this, recall that on the upper tail we use the GPD to model the distribution of $\epsilon_{hj}^{95} = y_{hj} - q_{hj}^{95}$, where $h \in S_{hj}^{95}$ and $S_{hj}^{95}$ is a subset of \{1, \ldots, n\} such that $y_{hj} \geq q_{hj}^{95}$. When the shape parameter $\xi$ of the GPD goes negative, the support of the GPD distribution for $\epsilon_{hj}^{95}$ becomes bounded to $[0, -\sigma/\xi]$, where $\sigma$ is the scale parameter. On the upper tail, some very high out-of-sample net-demand values lead to values of $\epsilon$ that fall outside the support of the distribution and are deemed impossible under this model (the lower tail is affected by the same problem when very low net-demand values occur). Hence the corresponding out-of-sample log-likelihood is undefined, which prevents the computation of the first two scoring rules in Table 2. More importantly, having a model under which some observed net-demand values are impossible is undesirable. Hence our decision to limit the parameter range. One might worry that the goodness of fit of the GPD-based model might have been compromised by forcing $\xi$ to stay positive. But the goodness of fit diagnostics reported in SM B.4 show that, even with this constraint, the marginal fit of the $\text{shash+gpd+cop}$ model is excellent on the in-sample data.

We fit the 14 conditional Gaussian, sinh-arcsinh and (pairs of) GPD models separately to the net-demand of each region, using the fitting methods of Wood et al. (2016) and the rolling forecasting approach described in Section 3.4. In a second step, the correlation of net-demand across the GSP groups is modelled as follows. Let $F_j(\cdot|x_i)$ be the conditional c.d.f. of the Gaussian, sinh-arcsinh or of the composition of the sinh-arcsinh and the GPD models corresponding to the $i$-th observation and the $j$-th GSP group. If a marginal net-demand model is well specified, $u_{ij} = F_j(y_{ij}|x_i)$ should approximately follow a uniform distribution $U(0,1)$ and $z_{ij} = \Phi^{-1}(u_{ij})$, where $\Phi^{-1}(\cdot)$ is the inverse standard Gaussian c.d.f., should follow a standard Gaussian distribution, $N(0,1)$. We then adopt a Gaussian copula model by assuming that the $z_{ij}$’s follow a joint multivariate Gaussian distribution, that is

$$
\begin{pmatrix}
z_{i1} \\
z_{i2} \\
z_{i3} \\
\vdots \\
z_{i14}
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
0 \\
0 \\
0 \\
\vdots \\
0
\end{pmatrix},
\begin{pmatrix}
1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,14} \\
\rho_{1,2} & 1 & \rho_{2,3} & \cdots & \rho_{2,14} \\
\rho_{1,3} & \rho_{2,3} & 1 & \cdots & \rho_{3,14} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \rho_{1,14} & \rho_{2,14} & \cdots & 1
\end{pmatrix}
\right).
$$

The parameters, $\rho_{1,2}, \ldots, \rho_{13,14}$, of the copula model are estimated simply by computing the empirical correlation matrix of the $z_i$ vectors on the training data, using the same rolling forecasting origin used for the multivariate Gaussian models based on the MCD parametrisation, which ensures that all the models are fitted exactly to the same training data.

The composite models with GAMLSS margins and a static Gaussian copula described above are closely related to the model class discussed in Kock and Klein (2023). In particu-
ular, Kock and Klein (2023) consider models with GAMLSS margins coupled with a static Gaussian copula for the correlation structure, but fit all parameters jointly via Markov chain Monte Carlo (MCMC) methods, rather than in two steps as done here.

B.4 Additional Results on UK Regional Net-Demand Forecasting

Figure 6 shows the bootstrapped differences between the performance scores of several pairs of models, under the scores shown in columns two to seven of Table 2. The interpretation of the plots is analogous to that of Figure 5a.

As mentioned in Section 3.5, it is possible to linearly transform the joint regional forecasts to obtain marginal probabilistic forecasts of differences in net-demand between regions or macro-regions. Here, we consider the difference in net-demand between the Scottish or South macro-regions and the rest of the country, or between London and its neighbouring regions. These boundaries are of particular interest to the network operator due to the strong influence of wind/solar generation on power flows and to the constraints related to network capacity and stability. Table 4 reports the performance of each model, when forecasting the marginal distribution of each net-demand difference. As for Table 2, the log scores of the models based on the sinh-arcsinh distribution are not reported because they are not readily computable.

Due to the importance of the Beast from the East cold wave on the performance scores (see Figure 5b), in Tables 5, 6, and 7 we report the same scores as in Tables 2, 3, and 4, but after having excluded the Beast from the East period (which is included between the two black marks at the bottom of Figure 5b) when fitting the models and evaluating their performance. Looking at Table 5, note that now the multivariate Gaussian Cal+Ren and Full models are better than the non-Gaussian models on the marginal log score as well.

Table 8 reports the shape parameters of the GPD models for the lower and upper tail, estimated on the whole data as part of the shash+gpd+cop model, with and without the observations corresponding to the Beast from the East cold wave. Note that for many GSP groups this parameter is equal to 0.001. This is because we impose a lower bound of 0.001 on this parameter, as explained in SM B.3.

The QQ-plots shown in Figure 7 and 8 are useful to assess the marginal goodness of fit of three of the models considered in this work. They are based on quantile residuals which, following Dunn and Smyth (1996), have been computed by evaluating the marginal c.d.f. of each model to obtain uniform residuals and then transforming them via the inverse standard normal c.d.f.. The left column in each figure shows the QQ-plots of the in-sample quantile residuals on 2018 data, obtained after fitting the models to the whole data (2014-18). There is a clear progression, that is the marginal fit gets better and better as the model for the margins becomes more flexible. The right column in each plot shows QQ-plots of
Figure 6: Bootstrapped differences between the scores of several pairs of models, for the performance score shown in columns two to seven of Table 2. Negative values mean that the first method is better than the second (e.g., Cal+Ren does better than Cal under all the scores). Note: The numbering of the figures shown here follows from the main text.
Table 4: Performance scores on 2018 test data, when forecasting the marginal distribution of differences in net-demand between macro-regions. The best score in each column is underlined. **Note:** The numbering of the tables shown here follows from the main text.

| Model          | Log CRPS | Pin 001 | Pin 999 | Log CRPS | Pin 001 | Pin 999 |
|----------------|----------|---------|---------|----------|---------|---------|
| Cal            | 2614     | 1305    | 20.75   | 9.31     | 2192    | 945.1   |
| Cal+Ren        | 2593     | 1301    | 18.97   | 8.89     | 2168    | 944.1   |
| Full           | 2591     | 1299    | 19.32   | 8.36     | 2161    | 943.8   |
| Cal+Ren Diag   | 2610     | 1304    | 21.01   | 9.52     | 2183    | 946.3   |
| gaulss+cop     | 2597     | 1304    | 19.49   | 9.65     | 2169    | 944.2   |
| shash+cop      | 1305     | 19.15   | 9.35    | 941.7    |
| shash+gpd+cop  | 1305     | 19.57   | 9.74    | 941.7    |

| Model          | Log CRPS | Pin 001 | Pin 999 |
|----------------|----------|---------|---------|
| Cal            | 932.4    | 367.8   | 8.73    | 2.13     |
| Cal+Ren        | 899.5    | 365.9   | 7.54    | 2.00     |
| Full           | 906.3    | 365.7   | 7.84    | 1.93     |
| Cal+Ren Diag   | 913.6    | 367.1   | 7.64    | 2.23     |
| gaulss+cop     | 896.8    | 366.7   | 6.96    | 2.45     |
| shash+cop      | 366.7    | 6.62    | 2.44    |
| shash+gpd+cop  | 366.7    | 6.81    | 2.32    |

Table 5: Performance scores on 2018 test data, when forecasting the joint distribution of net-demand across the 14 GSP groups, after removing the Beast from the East cold wave from the data. The best score in each column is underlined.

| Model          | Log CRPS | Pin 001 | Pin 999 | Var 0.5 | Var 1.0 |
|----------------|----------|---------|---------|---------|---------|
| Cal            | -4170    | -1421.3 | 2415    | 20.69   | 22.82   | 8978    | 9421    |
| Cal+Ren        | -4440    | -1646.6 | 2408    | 19.61   | 19.56   | 8802    | 9203    |
| Full           | -4444    | -1694.0 | 2407    | 18.53   | 19.36   | 8770    | 9159    |
| Cal+Ren Diag   | -4258    | -1535.9 | 2410    | 20.08   | 20.79   | 8804    | 9207    |
| gaulss+cop     | -4161    | -1541.2 | 2412    | 20.20   | 19.46   | 8860    | 9277    |
| shash+cop      | -4205    | -1589.8 | 2411    | 20.00   | 18.52   | 8847    | 9269    |
| shash+gpd+cop  | -4275    | -1606.6 | 2412    | 19.74   | 18.51   | 8848    | 9279    |
| Model          | Log CRPS | Log Ind | Pin 001 | Pin 999 | Var 0.5 | Var 1.0 |
|---------------|----------|---------|---------|---------|---------|---------|
| Cal           | 2807     | 3799.2  | 1849    | 14.13   | 18.95   | 2076    | 4429    |
| Cal+Ren       | 2724     | 3714.2  | 1844    | 13.09   | 17.72   | 2050    | 4363    |
| Full          | 2684     | 3689.3  | 1843    | 12.35   | 17.15   | 2040    | 4342    |
| Cal+Ren Diag  | 2818     | 3796.3  | 1847    | 13.63   | 19.89   | 2059    | 4386    |
| gaulss+cop    | 2810     | 3752.8  | 1847    | 14.15   | 19.52   | 2053    | 4375    |
| shash+cop     |          |         | 1847    | 13.64   | 18.58   | 2052    | 4375    |
| shash+gpd+cop |          |         | 1847    | 13.32   | 19.51   | 2053    | 4379    |

Table 6: Performance scores on 2018 test data, when forecasting the joint distribution of net-demand across the five GSP macro-regions, after removing the Beast from the East cold wave from the data. The best score in each column is underlined.

| Scotland - Rest | South - Rest |
|----------------|-------------|
| Log CRPS Pin 001 Pin 999 | Log CRPS Pin 001 Pin 999 |
| Cal | 2439 | 1198 | 11.47 | 9.16 | 2034 | 872.2 | 5.00 | 5.62 |
| Cal+Ren | 2431 | 1196 | 11.46 | 9.02 | 2038 | 872.9 | 5.38 | 5.55 |
| Full | 2427 | 1194 | 11.24 | 8.49 | 2033 | 872.3 | 5.23 | 5.34 |
| Cal+Ren Diag | 2449 | 1199 | 13.62 | 9.45 | 2053 | 875.5 | 6.06 | 6.44 |
| gaulss+cop | 2442 | 1199 | 12.79 | 9.89 | 2035 | 873.0 | 5.46 | 5.89 |
| shash+cop | 1200 | 12.78 | 9.59 | 871.0 | 5.28 | 5.72 |
| shash+gpd+cop | 1200 | 13.22 | 9.84 | 871.1 | 5.40 | 5.95 |

| London - Neighbours |
|---------------------|
| Log CRPS Pin 001 Pin 999 |
| Cal | 792.3 | 335.3 | 5.40 | 1.92 |
| Cal+Ren | 769.8 | 333.9 | 4.92 | 1.89 |
| Full | 770.2 | 333.5 | 4.92 | 1.84 |
| Cal+Ren Diag | 784.1 | 334.8 | 5.10 | 2.07 |
| gaulss+cop | 777.5 | 335.0 | 5.05 | 2.38 |
| shash+cop | 334.7 | 4.98 | 2.33 |
| shash+gpd+cop | 334.9 | 4.75 | 2.23 |

Table 7: Performance scores on 2018 test data, when forecasting the marginal distribution of differences in net-demand between macro-regions, after removing the Beast from the East cold wave from the data. The best score in each column is underlined.
Table 8: Estimated shape parameters of the GPD models for the lower and upper tail of the net-demand distribution, for each GSP group.

| Area                        | With the Beast from the East cold wave | Without the Beast from the East cold wave |
|-----------------------------|----------------------------------------|------------------------------------------|
|                             | Lower tail    | Upper tail    | Lower tail    | Upper tail    |
| N Scotland (P)              | 0.127         | 0.142         | 0.132         | 0.068         |
| S Scotland (N)              | 0.133         | 0.145         | 0.132         | 0.135         |
| NE England (F)              | 0.001         | 0.001         | 0.001         | 0.001         |
| Yorkshire (M)               | 0.007         | 0.001         | 0.018         | 0.001         |
| NW England (G)              | 0.001         | 0.001         | 0.001         | 0.001         |
| Merseyside & N Wales (D)    | 0.001         | 0.114         | 0.001         | 0.015         |
| S Wales (K)                 | 0.001         | 0.001         | 0.001         | 0.001         |
| W Midlands (E)              | 0.044         | 0.092         | 0.043         | 0.030         |
| E Midlands (B)              | 0.001         | 0.084         | 0.001         | 0.001         |
| E England (A)               | 0.001         | 0.039         | 0.001         | 0.001         |
| London (C)                  | 0.064         | 0.205         | 0.063         | 0.001         |
| SE England (J)              | 0.001         | 0.139         | 0.001         | 0.083         |
| S England (H)               | 0.001         | 0.103         | 0.001         | 0.001         |
| SW England (L)              | 0.001         | 0.177         | 0.001         | 0.013         |

The out-of-sample (one day ahead) quantile residuals on 2018 data, obtained under the monthly model updating scheme described in Section 3.4. We can still see a progression as we move to more flexible marginal models, but this is much less clear than when looking at in-sample residuals. Hence, the goodness of fit improvements brought about by adopting more flexible marginal models are more limited in the test set, which is not surprising considering the extraordinary nature of events such as the Beast from the East cold wave.
Figure 7: Left: QQ-plots of the in-sample quantile residuals for each GSP group, for the Cal+Ren, shash+cop and shash+gpd+cop model. Right: QQ-plots of the out-of-sample (one day ahead) quantile residuals under the same three models.
Figure 8: Left: QQ-plots of the in-sample quantile residuals for each GSP group, for the Cal+Ren, shash+cop and shash+gpd+cop model. Right: QQ-plots of the out-of-sample (one day ahead) quantile residuals under the same three models. Here the Beast from the East cold wave has been removed from the data.
References

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4), 1059–1086.

Capezza, C., B. Palumbo, Y. Goude, S. N. Wood, and M. Fasiolo (2021). Additive stacking for disaggregate electricity demand forecasting. *The Annals of Applied Statistics* 15(2), 727–746.

Dunn, P. K. and G. K. Smyth (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5(3), 236–244.

Hofner, B., T. Hothorn, T. Kneib, and M. Schmid (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics* 20(4), 956–971.

Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. *Biometrika* 96(4), 761–780.

Kock, L. and N. Klein (2023). Truly multivariate structured additive distributional regression. *arXiv preprint arXiv:2306.02711*.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3), 507–554.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111(516), 1548–1563.