IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 14, 2022, accepted May 26, 2022, date of publication June 2, 2022, date of current version June 7, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3179697*

# Characterizing Wind Power Forecast Error Using Extreme Value Theory and Copulas

**NDAMULELO MARARAKANYE, AMARIS DALTON, (Member, IEEE), AND BERNARD BEKKER, (Member, IEEE)**
Department of Electrical and Electronic Engineering, Stellenbosch University, Matieland 7602, South Africa

Corresponding author: Ndamulelo Mararakanye (ndamulelo@sun.ac.za)

This work was supported in part by the Centre for Renewable and Sustainable Energy Studies (CRSES), Stellenbosch University; and in part by the Eskom Power Plant Engineering Institute (EPPEI).

**ABSTRACT** Wind energy is one of the fastest-growing renewable energy sources in the world. However, wind power is variable in all timescales. This variability is difficult to predict with perfect certainty, with potentially significant financial implications when rare extreme forecast errors occur. This paper focuses on three key aspects associated with the extreme errors of geographically distributed wind farms: suitable parametric distribution representation, effects of diurnality, seasonality and larger atmospheric circulations, and modeling multivariate distribution. The paper shows that some of the distributions commonly used for modeling forecast errors may be inappropriate in representing extreme errors. As the first contribution, this paper fits a Generalized Pareto distribution (GPD) from extreme value theory to achieve a better estimation of extreme errors. In the second contribution, this paper splits extreme errors by hour, month, and atmospheric states to investigate the statistical regularities of GPD parameters along diurnal and seasonal timescales and larger atmospheric circulations. In the third contribution, this paper uses copula functions to model multivariate extreme error distribution and investigates their effectiveness in providing a regional view of extreme errors. This paper tests the proposed methodology using the forecast error data obtained from 29 wind farms in South Africa. The results show that GPD outperforms commonly used distributions. Extreme errors have strong diurnal and seasonal components and vary significantly between SOM nodes. Copulas can be useful in providing a regional view of extreme errors. This paper improves the estimation of extreme errors, which is an important step toward better operating reserve allocation.

**INDEX TERMS** Atmospheric state, copula, extreme value theory, Generalized Pareto distribution, wind power forecast error.

## NOMENCLATURE

| | |
|---|---|
| $u$ | Threshold |
| $F_u(.)$ | Cumulative distribution function of exceedances over threshold |
| $P(.)$ | Probability operator |
| $F(.)$ | Cumulative distribution function |
| $N$ | Sample size |
| $G_{\xi,\beta}(.)$ | Cumulative distribution function of Generalized Pareto distribution |
| $\xi$ | Shape parameter |
| $\beta$ | Scale parameter |
| $N_u$ | Number of exceedances over threshold |
| $y_i$ | Exceedance over threshold |
| $l(.)$ | Log-likelihood |
| $x_m$ | Return level |
| $m$ | Return period |
| $\zeta_u$ | Proportion of observations that are greater than threshold |
| $P_n(.)$ | Weight vector for each node |
| $\varphi(.)$ | Number of nodes in a neighborhood stretching between nodes $j$ and $i$ for the $t$'th iteration |
| $\rho(.)$ | Temporally decreasing learning function |
| $R(.)$ | Randomly selected feature vector |
| $C(.)$ | Copula |
| $D, S$ | Forecast errors of clusters 1 and 2, respectively |
| $F_D(.), F_S(.)$ | Marginal distribution functions of clusters 1 and 2, respectively |

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaodong Liang.

$F_{D,S}(.)$ Joint distribution function with marginal distribution functions $F_D\ (.)$ and $F_S\ (.)$

$d^{'}, s^{'}$ Forecast error thresholds for clusters 1 and 2, respectively

## I. INTRODUCTION

The use of wind energy for electricity generation is increasing worldwide. This is because there is a need to decarbonize the electricity industry and wind energy is becoming cost-competitive. However, unlike conventional thermal power, wind power varies over time. This is a source of concern for system operators, who must ensure that supply matches demand at all times. One way of mitigating the impacts of wind power variations in power system operations is wind power forecasting. These forecasts provide system operators with an estimate of future wind power generation, but they are rarely perfect. Small forecast errors are not always a concern for system operators since power systems can accommodate a certain level of variability and uncertainty from the load demand. However, during significant wind power ramp events, forecast errors from a day-ahead prediction can be as high as 60-80% of total wind capacity [1], [2]. If there is insufficient operating reserve to deal with these extreme forecast errors, the system operator may have to implement wind generation curtailment or load shedding – scenarios that system operators try to avoid due to the associated financial implications. In deregulated markets, extreme forecast errors can also affect energy traders since inaccurate bids during these events can result in costly penalties. As an example, on February 26, 2008, Electric Reliability Council of Texas reported a high forecast error event, forcing them to declare a system emergency, which is a high-cost system condition [3]. These potential implications justify the need for understanding and characterizing the magnitude and frequency of extreme wind power forecast errors, towards better operational decisions such as dynamic operating reserve allocation to account for wind power uncertainty.

In this paper, we analyze the tails of forecast error distributions of geographically distributed wind farms. We focus on three main aspects associated with extreme forecast errors; 1) suitable parametric distribution representation, 2) effects of diurnality, seasonality, and large atmospheric circulations, and 3) modeling the multivariate distribution.

The studies in the literature frequently used the normal distribution to model wind power forecast errors [4]–[11]. Other distributions considered in the literature include beta [12], Weibull [13], [14], Cauchy [14], [15] and hyperbolic [15], [16]. While these distributions are relatively suitable for representing the body of the forecast error distribution, the same assertion is not valid for the tails of the forecast error distribution. According to the findings in [1], [10]–[12], [17], normal, beta, and Weibull distributions are not fat-tailed enough, and therefore often underestimate the frequency of extreme forecast errors. On the other hand, the study in [15] demonstrated that the Cauchy distribution is overly fat-tailed and over-represents the frequency of extreme forecast errors. According to the findings in [14], [15], the hyperbolic distribution seems to perform better compared to normal, beta, Weibull, and Cauchy distributions in modeling extreme forecast errors. Given the severe financial implications of extreme forecast errors, finding models that best represent these extreme forecast errors remains critical. As a result, other studies in the literature have considered non-parametric approaches for modeling wind power forecast errors [18], [19]. While non-parametric approaches can be accurate, extreme forecast errors often do not occur frequently enough to make accurate non-parametric inferences [15], [20]. This paper will investigate whether extreme forecast errors can be modeled accurately using the Extreme Value Theory (EVT) by fitting the Generalized Pareto distribution (GPD) on the extreme forecast errors. Recent literature has used the same approach to model the tail behavior of wind speed [21], [22] and wind power ramp [23], however, the approach has not been explored for wind power forecast error.

The second aspect of this paper is investigating the influence of diurnality, seasonality, and large atmospheric circulations on extreme forecast errors. Several studies (e.g. [24]–[30]) in the literature have demonstrated that wind power profiles exhibit a high degree of statistical regularity along diurnal and seasonal timescales. In addition, the study in [31] revealed that large atmospheric circulations are the major cause of wind power variations over timescales ranging from hours to days. These variables are the basis used to understand wind power variability and ultimately improve wind power forecasting (e.g. [32]–[34]). However, there is little to no investigation in the literature on how these variables affect forecast errors – towards improved estimation of extreme forecast errors. To investigate the diurnal and seasonal patterns of extreme forecast errors, this paper investigates the tail distribution (or fitted GPD) associated with each hour and month. Additionally, this paper assigns atmospheric states, derived from self-organizing maps (SOMs), to each historical forecast scenario and investigates the tail distribution associated with each state.

The third aspect of this paper is modeling multivariate forecast error distribution. While univariate analysis can be useful in certain applications (e.g. congestion management), other operational decisions such as operating reserve allocation need to consider all wind farms within a region. It is thus important to evaluate the dependence structure of forecast errors from geographically distributed wind farms. Copula theory is widely used to model dependence structure between variables, mostly in financial market analysis, portfolio investments, and risk assessments [35]. In recent years, copula theory has been applied in wind power analyses (e.g. [21], [32], [33], [35]–[38]). The majority of these studies used copula theory to model the spatial dependency of regional wind speeds or power outputs. However, there is little to no investigation on using copula theory to model forecast errors. This paper uses copula functions to model multivariate extreme forecast error distribution and investigates if

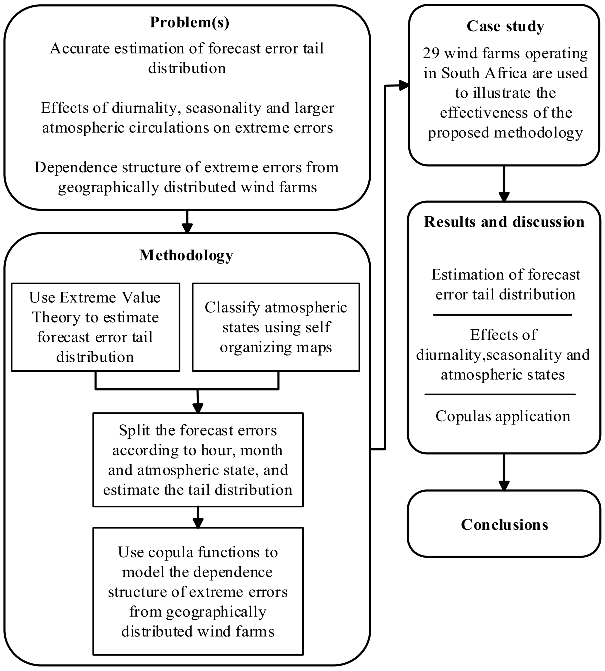
Problem(s)
Accurate estimation of forecast error tail distribution
Effects of diurnality, seasonality and larger atmospheric circulations on extreme errors
Dependence structure of extreme errors from geographically distributed wind farms
Methodology
Use Extreme Value Theory to estimate forecast error tail distribution
Classify atmospheric states using self organizing maps
Split the forecast errors according to hour, month and atmospheric state, and estimate the tail distribution
Use copula functions to model the dependence structure of extreme errors from geographically distributed wind farms
Case study
29 wind farms operating in South Africa are used to illustrate the effectiveness of the proposed methodology
Results and discussion
Estimation of forecast error tail distribution
Effects of diurnality,seasonality and atmospheric states
Copulas application
Conclusions

**FIGURE 1.** Paper overview.

this is effective in providing a region-wide view of extreme forecast errors using numerical examples. Fig. 1 shows an overview of this paper.

The remainder of this paper is organized as follows: section II provides the theoretical framework of the proposed methodology. Section III introduces the case study of 29 wind farms in South Africa used for illustrating the proposed methodology and presents the results and discussion. Section IV summarizes the findings before concluding the paper.

The main contributions of this paper can be summarized as finding a suitable parametric distribution for representing extreme forecast errors, improving understanding on some of the factors that may influence extreme forecast errors, and proposing a suitable model for representing spatial correlations of extreme forecast errors between wind farms. These contributions can assist system operators with a method of deriving conditional extreme forecast error distributions that changes based on various states (i.e. hour, month, atmospheric and spatial configuration of wind farms). These conditional distributions usually contain more probabilistic information (as compared to unconditional distribution), which can improve operating reserve allocation to account for wind power uncertainty. In addition, the classified atmospheric states represent larger atmospheric circulation, allowing inputs into reserve allocation based on physical meteorological phenomena.

## II. METHODOLOGY

### A. FITTING FORECAST ERROR TAIL DISTRIBUTION

In EVT, there are two ways to sample extreme events: block maxima (BM) and peak-over-threshold (POT). The BM method divides the data into equal blocks, extracts the highest observation in each block, then fits a Generalized Extreme Value (GEV) distribution to the block maxima. The BM method has a significant limitation in that it retrieves only one observation from each block, regardless of whether the second-highest observation in a block exceeds the largest observations in adjacent blocks. As a result, adopting BM necessitates a large amount of data [43], [44]. The POT, on the other hand, entails setting a threshold, extracting the excess of observations over the threshold, and fitting a GPD to the exceedances. This is a more flexible technique that typically enables more observations to be retrieved (rather than just one in each block), resulting in reduced uncertainty [43], [44].

As a result, the POT approach is used in this paper to model the distribution of forecast error exceedances over a high threshold $u$. Assuming that $X_1, X_2, \ldots, X_N$ (representing the forecast errors for individual clusters) is an independent and identically distributed sequence of random variables and $N$ is the sample size, the distribution function $F_u(y)$ of exceedances $X$ over a threshold $u$ is defined by:

$$F_u(y) = P(X - u \leq y \,|\, X > u) = \frac{F(y+u) - F(u)}{1 - F(u)} \qquad (1)$$

With high enough $u$, $F_u(y)$ can be approximated by a GPD with the following cumulative distribution function.

$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}} & if\ \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\beta}\right) & if\ \xi = 0 \end{cases}$$

For

$$\begin{cases} \beta > 0\ and\ y \geq 0 & if\ \xi \geq 0 \\ 0 \leq y \leq -\frac{\beta}{\xi} & if\ \xi < 0 \end{cases} \qquad (2)$$

where $\xi$ is the shape parameter and $\beta$ is the scale parameter. To estimate the values of $\xi$ and $\beta$, this paper uses the maximum likelihood estimation (MLE) approach. If $y_1, y_2, \ldots, y_{N_u}$ is a sequence of $N_u$ exceedances over a threshold $u$, the log-likelihood can be derived for $\xi \neq 0$ as:

$$l(\beta, \xi) = -N_u log\beta - (1 + \frac{1}{\xi})\sum_{i=1}^{N_u} log\left(1 + \xi\frac{y_i}{\beta}\right) \qquad (3)$$

Provided $\left(1 + \xi y_i/\beta\right) > 0$ for $i = 1, 2, \ldots, N_u$; otherwise $l(\beta, \xi) = \infty$. When $\xi = 0$, the log-likelihood can be derived as:

$$l(\beta) = -N_u log\beta - (1/\beta)\sum_{i=1}^{N_u} y_i \qquad (4)$$

The maximum likelihood estimates for GPD distributions are achieved by maximizing (3) and (4) with respect to parameters $\beta$ and $\xi$.

After estimating the suitable parameters of the GPD, we can evaluate the forecast error $x_m$ that is expected to be

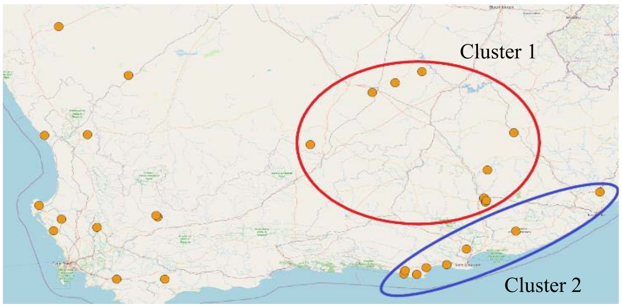

**FIGURE 2.** Geographical locations of operating wind farms in South Africa and clusters used for evaluating the proposed methodology.

exceeded on average once every $m$ observations (with probability $1/m$). The forecast error $x_m$ is also known as the return level, while the $m$ observations or inverse of the probability that the forecast error $x_m$ will be exceeded is also known as the return period. The return level and return period can be useful to system operators to allocate reserves to account for wind power uncertainty. The return level $x_m$ can be derived for $\xi \neq 0$ as:

$$x_m = u + \frac{\beta}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] \tag{5}$$

Provided that $m$ is large to ensure that $x > u$. When $\xi = 0$, the return level can be derived as:

$$x_m = u + \beta \log(m\zeta_u) \tag{6}$$

The parameter $\zeta_u = N_u/N$ is the proportion of observations that are greater than $u$.

### B. CLASSIFICATION OF ATMOSPHERIC STATES

To investigate the relationship between large – or *synoptic*-scale atmospheric circulation and extreme forecast errors, atmospheric circulation was classified into a set of atmospheric states that serve as archetypical representations of weather regimes associated with the climatology of a region. The classification of atmospheric circulation, as a complexity reduction mechanism, is a common and well-established practice in the meteorological community. Classification techniques have evolved from subjective approaches, which are dependent on expert knowledge, toward objective computer-assisted methodologies such as principal component analysis, k-means clustering, and self-organizing maps (SOMs) [39].

This paper makes use of SOMs as a classification technique. A SOM is a class of self-learning artificial neural network that allows for the representation of high dimensional data onto what is typically a 2D lattice (or map), whilst preserving the topology of the higher dimensional data [40]. SOMs are trained using a competitive learning algorithm represented in (7) below. During training, a set of SOM nodes ($n$), established during the initialization process, are continually updated according to which node best matches (based on the Euclidean distance) each randomly selected iterative input vector ($R(s)$), for each step ($t$) of the training process. This most similar node is called the best matching unit (BMU). Subsequently, each BMU along with a number of nodes in a neighborhood ($\varphi$) stretching between nodes $j$ and $i$, are adjusted to increase their similarity to that of the input vector. The size of the neighborhood decreases throughout the training process based on a monotonically decreasing learning coefficient ($\rho$). Thereby weight vector for each node $P_n$ is updated as follows:

$$P_n(t+1) = P_n(t) + \varphi(i,j,t) \cdot \rho(t) \cdot (R(s) - P_n(t)) \tag{7}$$

### C. MULTIVARIATE FORECAST ERROR DISTRIBUTION

Once the forecast error distribution of exceedances over a high threshold for each cluster has been obtained, it is important to link these univariate distributions (to form multivariate distribution) to get a system-wide view of forecast error. The multivariate distribution should account for spatial-temporal correlations in forecast errors between various clusters.

This paper uses copula functions to model the bivariate joint distribution function $F_{D,S}(d,s)$, where $D$ and $S$ are forecast errors for clusters 1 and 2, respectively. According to the Sklar's Theorem, if $F_{D,S}(d,s)$ is a two-dimensional distribution function with marginal distributions $F_D(d)$ and $F_S(s)$, then there exists a copula $C$ such that:

$$F_{D,S}(d,s) = C(F_D(d), F_S(s)) \tag{8}$$

Conversely, if $C$ is a copula with $F_D(d)$ and $F_S(s)$ being the distribution functions, then the function $F_{D,S}(d,s)$ defined by (8) is a joint distribution function with marginal distributions $F_D(d)$ and $F_S(s)$. Section III-D will discuss the selection of an appropriate copula function for this particular application.

The derived copula-based joint forecast error distribution provides some important information about forecast error in a region. For example, the probability that forecast errors from both clusters exceed certain thresholds can be obtained in terms of copulas as follows:

$$P(D \geq d, S \geq s) = 1 - F_D(d) - F_S(s) + C(F_D(d), F_S(s)) \tag{9}$$

In addition, it may be of interest to system operators to evaluate the forecast error distribution of cluster 1 given that the forecast error of cluster 2 exceeds a certain threshold $s'$. This conditional distribution is given by:

$$P(D \leq d, S \geq s') = \frac{F_D(d) - C(F_D(d), F_S(s'))}{1 - F_S(s')} \tag{10}$$

Conversely, the conditional forecast error distribution of Cluster 2 given that the forecast error of Cluster 1 exceeds a certain threshold $d'$ is given by:

$$P(S \leq s, D \geq d') = \frac{F_S(s) - C(F_D(d'), F_S(s))}{1 - F_D(d')} \tag{11}$$

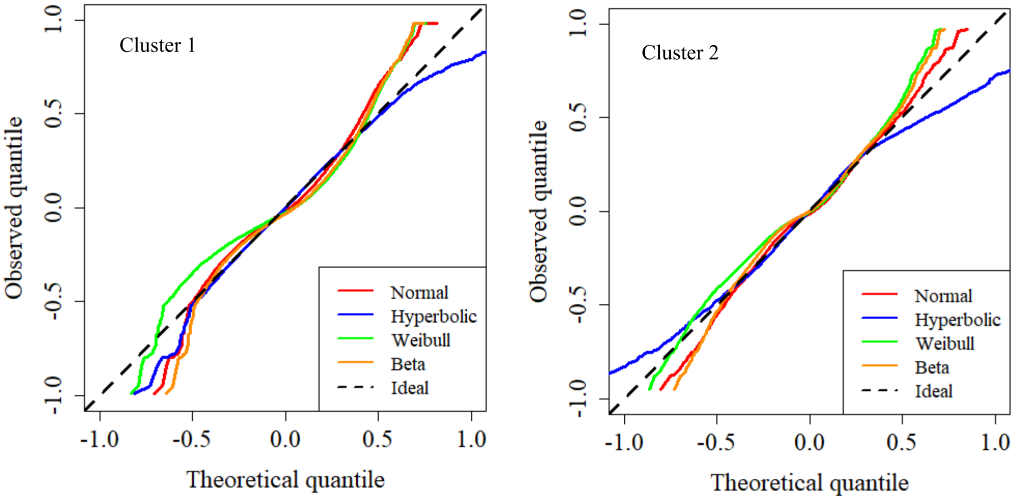

**FIGURE 3.** Q-Q plot of normal, hyperbolic, Weibull and beta distributions for both clusters.

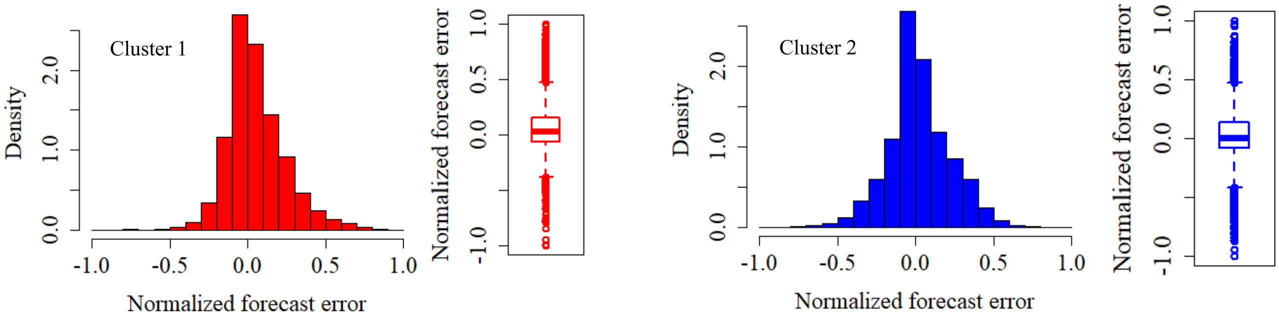

**FIGURE 4.** PDF and boxplot of forecast error for both considered clusters.

## III. CASE STUDY AND RESULTS

### A. DESCRIPTION OF DATA USED

To test the proposed methodology, this paper uses the day-ahead point forecast error data (between 01 January 2018 and 31 March 2021) from 29 wind farms, obtained from Eskom (the power utility company in South Africa). However, due to the confidentiality of individual point forecast error data, Eskom was only able to provide the data in clusters of wind farms summed together. This paper uses two of those clusters (with 18 wind farms) to demonstrate the concepts proposed in this paper. Fig. 2 shows the locations of the wind farms within each of these clusters.

The forecast errors range between −57.04% and 54.14% (of installed wind capacity) in Cluster 1, while the errors range between −63.71% and 67.31% in Cluster 2. To make comparisons between clusters easier, this paper represents the errors that are greater or equal to zero on a scale of [0, 1] and negative errors on a scale of [−1, 0). Fig. 3 shows the main characteristics of the forecast error data – the probability density function (PDF) and boxplot of forecast errors for both considered clusters. As seen in Fig. 3, the observed error distribution from both clusters is positively skewed. In addition, there is a significant amount of observations that are located outside the whiskers of box-plots (or outliers), which can be an indication of heavy-tailed distributions.

### B. PARAMETER ESTIMATION OF THE FORECAST ERROR TAIL DISTRIBUTION

As discussed in section I, common distributions used for representing forecast errors include, normal, hyperbolic, Weibull and beta. To see if these distributions can also represent the forecast error data described in section III-A, we consider the quantile-quantile (Q-Q) plot. This plot shows the quantiles of hypothesized distributions (in this case normal, beta, Weibull and beta) as a function of the observed quantiles. If the observed data is drawn from the hypothesized distribution, then the Q-Q plot is linear with a slope of 45 degrees. Fig. 4 shows the Q-Q plot of normal, hyperbolic, Weibull, and beta distributions for both considered clusters. As seen from Fig. 4, the considered distributions are relatively suitable for representing forecast errors in ranges −0.53 to 0.36 and −0.47 to 0.45 for clusters 1 and 2, respectively. However, the

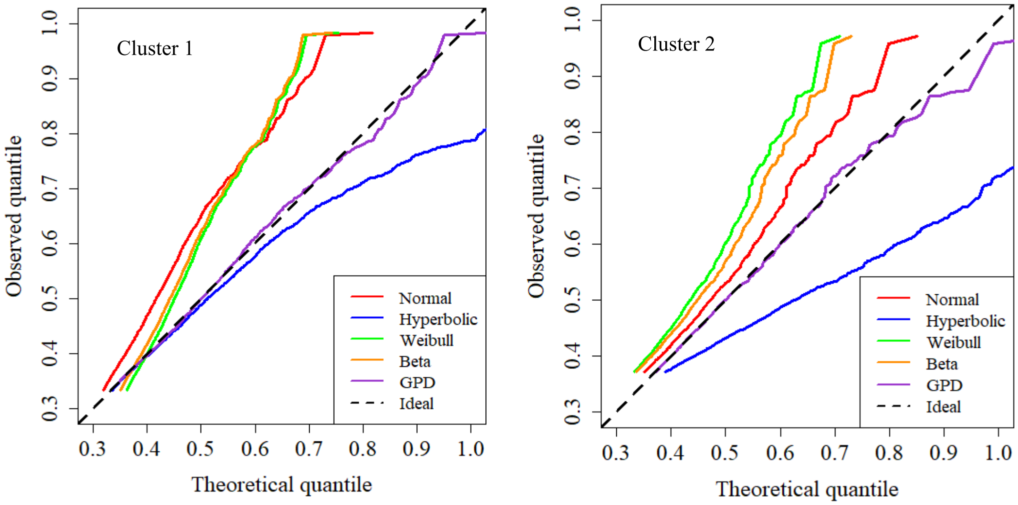

**FIGURE 5.** Q-Q plot of GPD compared with Q-Q plot of considered theoretical distributions for both clusters.

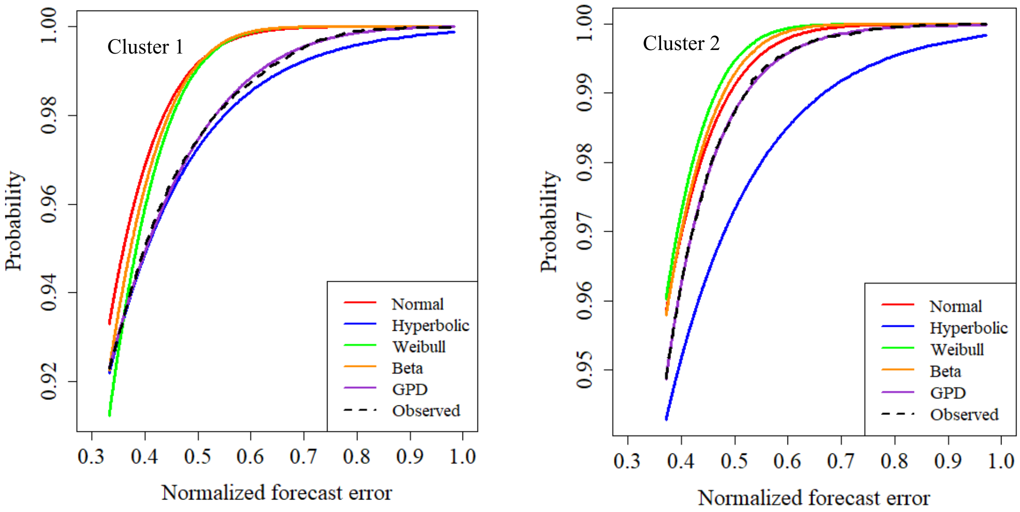

**FIGURE 6.** Comparison of cumulative distribution functions of observed data and considered theoretical distributions for both clusters.

**TABLE 1.** Comparison of MAE and RMSE between cumulative distribution functions for both clusters.

| | Cluster 1 | |
|---|---|---|
| Distribution | MAE | RMSE |
| Normal | 1.38e-2 | 1.45e-2 |
| Hyperbolic | 1.32e-3 | 1.56e-3 |
| Weibull | 2.58e-2 | 3.25e-2 |
| Beta | 2.49e-2 | 3.14e-2 |
| GPD | 6.55e-4 | 8.05e-4 |
| | Cluster 2 | |
| Distribution | MAE | RMSE |
| Normal | 6.72e-3 | 7.31e-3 |
| Hyperbolic | 1.11e-2 | 1.14e-2 |
| Weibull | 1.74e-2 | 2.11e-2 |
| Beta | 1.71e-2 | 2.07e-2 |
| GPD | 1.82e-4 | 2.19e-4 |

normal, Weibull, and beta distributions underestimate the extreme forecast errors in both clusters, while the hyperbolic distribution tends to overestimate the extreme forecast errors (except for negative extreme errors in Cluster 1). In other words, the observed data has heavier tails than estimates from the normal, Weibull, and beta distributions, and lighter tails than estimates from hyperbolic distribution.

In this paper, we propose fitting the tails of the forecast error distribution with a GPD. The forecast error data, as shown in Fig. 3 and Fig. 4, have both left and right tails. This study uses the right tail as an example; however, it is possible to apply the same approach to the left tail. The first step in the proposed approach is to identify the forecast error threshold above which we can fit the GPD. If the threshold is set too high, there will be few observations that exceed it, resulting in a significant variance [43]. If the threshold is too low, data with ordinary values will be included as extremes, making the asymptotic assumption less valid [43]. In this paper, the mean excess and parameter threshold stability plots (also used for example in [43], [46], [47]) are used to identify the right tail thresholds of 0.33 and 0.37 for clusters 1 and 2,

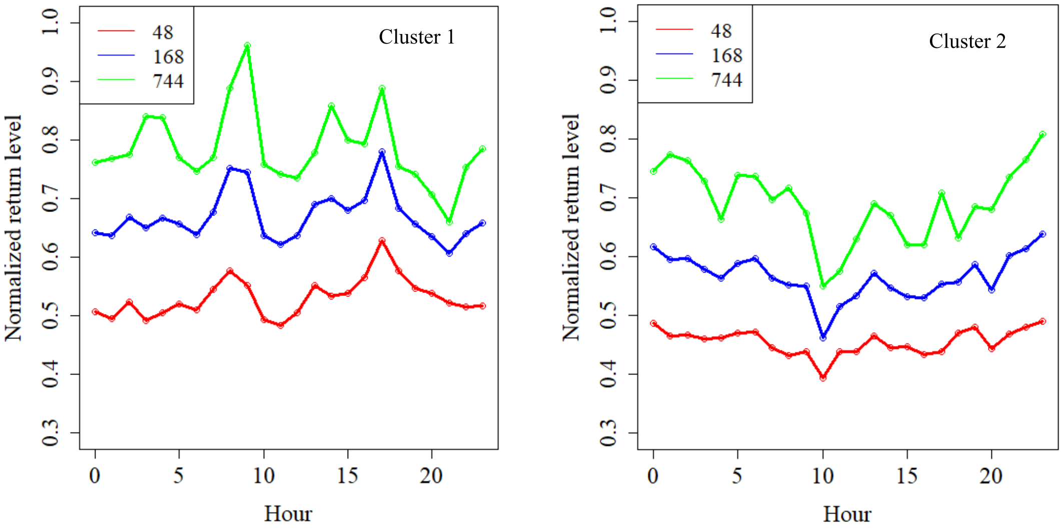

**FIGURE 7.** Conditional return level associated with each hour at different return periods for both considered clusters.

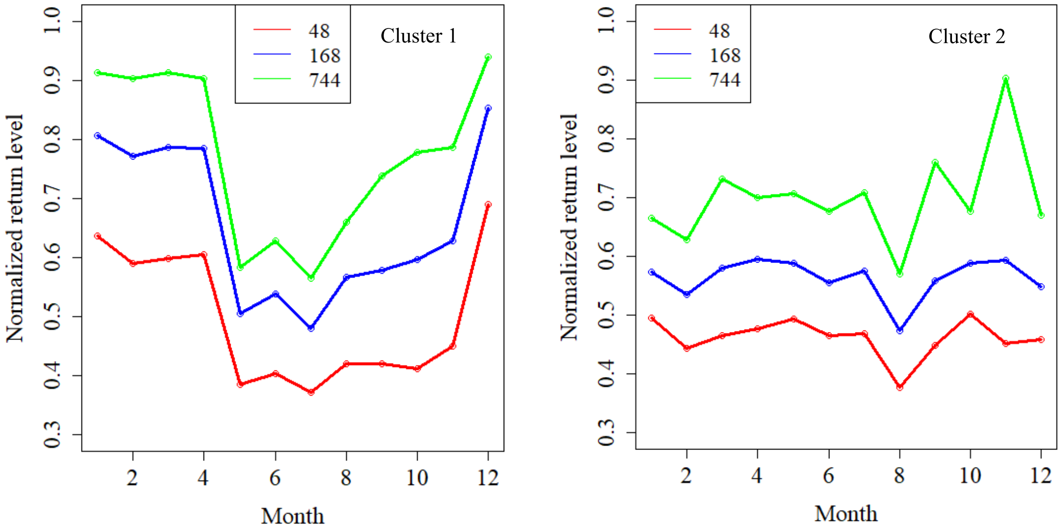

**FIGURE 8.** Conditional return level associated with each month at different return periods for both considered clusters.

respectively. After determining the thresholds, the parameters of the GPD for both clusters are calculated as outlined in section II-A.

Fig. 5 shows the Q-Q plots of the GPD for both clusters. For comparative purposes, Fig. 5 also shows the Q-Q plots of normal, hyperbolic, Weibull, and beta distributions (focused on the tails of the distribution) for both clusters. One can notice a significant improvement in fitting the GPD on the extreme forecast errors compared to the other distributions.

To emphasize this finding, Fig. 6 illustrates the cumulative distribution functions of observed data, normal, hyperbolic, Weibull, beta, and GPD for both clusters. The GPD is noticeably closer to the observed cumulative distribution function, whereas the other distributions significantly under- or overestimate the probabilities of extreme forecast errors as already observed on the Q-Q plots. We also examine the mean absolute error (MAE) and root mean squared error (RMSE) between cumulative distribution functions of normal, hyperbolic, Weibull, beta, and GPD in relation to the observed cumulative distribution function – see Table 1. This can be seen as a numerical confirmation that the GPD is closest to the observed data.

### C. EFFECTS OF DIURNALITY, SEASONALITY, AND ATMOSPHERIC STATES ON TAIL DISTRIBUTION

With the estimated parameters of the GPD in Section III-B, we can calculate the return levels using (5) and (6). For example, the expectation is that the forecast error will exceed 0.53 and 0.45 on overage once every 48 hours (or a probability of 0.021) for clusters 1 and 2, respectively.

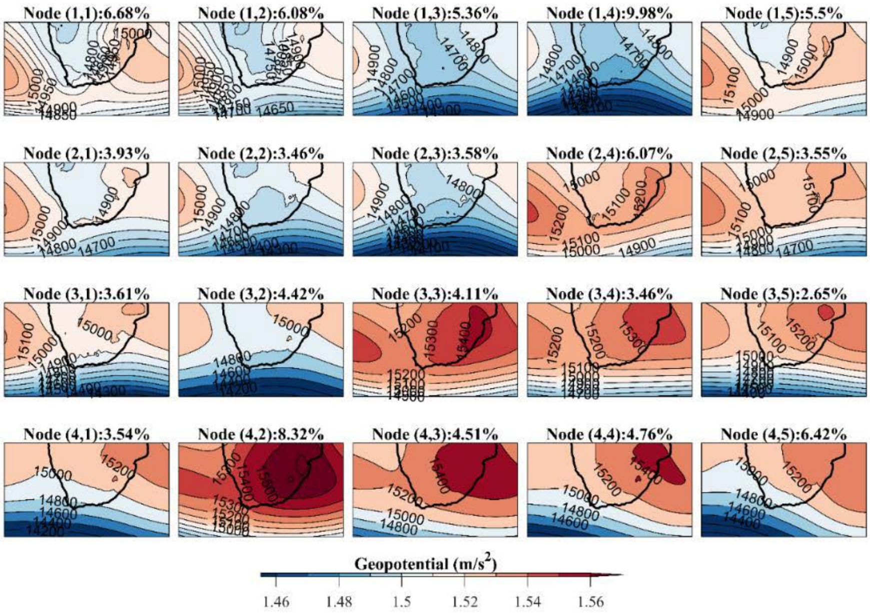

**FIGURE 9.** SOM geospatial heights together with frequency of SOM node occurrence over South Africa.

**TABLE 2.** Conditional return level for each SOM node at different return periods for both clusters.

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| Node | 48 | 168 | 744 | 48 | 168 | 744 |
| (1,1) | 0.61 | 0.72 | 0.80 | 0.44 | 0.57 | 0.76 |
| (1,2) | 0.61 | 0.76 | 0.87 | 0.46 | 0.62 | 0.85 |
| (1,3) | 0.62 | 0.76 | 0.88 | 0.45 | 0.57 | 0.71 |
| (1,4) | 0.59 | 0.72 | 0.81 | 0.45 | 0.55 | 0.65 |
| (1,5) | 0.51 | 0.64 | 0.75 | 0.40 | 0.51 | 0.65 |
| (2,1) | 0.41 | 0.55 | 0.64 | 0.51 | 0.64 | 0.79 |
| (2,2) | 0.43 | 0.55 | 0.65 | 0.44 | 0.50 | 0.52 |
| (2,3) | 0.59 | 0.69 | 0.77 | 0.49 | 0.64 | 0.78 |
| (2,4) | 0.49 | 0.60 | 0.69 | 0.42 | 0.49 | 0.53 |
| (2,5) | 0.34 | 0.45 | 0.55 | 0.45 | 0.54 | 0.64 |
| (3,1) | 0.49 | 0.62 | 0.72 | 0.43 | 0.51 | 0.60 |
| (3,2) | 0.58 | 0.74 | 0.89 | 0.46 | 0.57 | 0.66 |
| (3,3) | 0.44 | 0.56 | 0.67 | 0.48 | 0.54 | 0.57 |
| (3,4) | 0.36 | 0.46 | 0.61 | 0.50 | 0.64 | 0.80 |
| (3,5) | 0.55 | 0.70 | 0.82 | 0.47 | 0.57 | 0.68 |
| (4,1) | 0.49 | 0.66 | 0.85 | 0.49 | 0.64 | 0.77 |
| (4,2) | 0.34 | 0.45 | 0.55 | 0.45 | 0.55 | 0.66 |
| (4,3) | 0.43 | 0.56 | 0.66 | 0.47 | 0.55 | 0.60 |
| (4,4) | 0.51 | 0.67 | 0.82 | 0.45 | 0.56 | 0.70 |
| (4,5) | 0.56 | 0.72 | 0.89 | 0.46 | 0.58 | 0.74 |

Similarly, if we split the data according to hour, month, and atmospheric state, we can estimate the parameters of the conditional GPD (and associated return level) to assess the impact of diurnality, seasonality, and larger atmospheric circulation on tail distribution of forecast errors. Fig. 7 and Fig. 8 show the return level associated with each hour and month at different return periods (48 hours or 2 days, 168 hours or week, and 744 hours or month), respectively. The return level of both clusters fluctuates dramatically across different hours of the day. Cluster 1 has considerable spikes in return levels at 9h and 17h, while Cluster 2 has a dip at 10h. The return level of cluster 1 is often high during the day and low at night, whereas the return level of Cluster 2 is the reverse. The return level of Cluster 1 has a visible seasonal pattern – it drops during the winter months (May to July) and it is at its highest during the summer months (December to February). On the other hand, the return level of Cluster 2 remains relatively flat throughout the months, except for the noticeable dip in October.

To assess the impact of larger atmospheric circulations on extreme forecast errors, this paper selected a 4-by-5 SOM node lattice, resulting in a 20 node SOM. The SOM was trained in two phases using the batch training algorithm on a rectangular lattice – firstly a rough training phase consisting of 1000 iterations which was followed by a fine-tuning phase of 5000 iterations. During the rough training phase, the neighborhood function decreased from 5-to-1 and during the fine-tuning phase, it decreased from 2-to-1. Both training phases use the Epanechnikov neighborhood function, as recommended for small SOMs [45]. Once the training process was completed, and the SOM has been created, each input vector (i.e. each geopotential height time step) was assigned a 'label' based on which SOM node it is most similar to, likewise using the Euclidean distance as with the training phase. Accordingly, this allows firstly each time step in the input geopotential height time series, along with the corresponding wind power prediction error time series, to be clustered based on the atmospheric state that was concurrent to each time step.

Fig. 9 shows the 20 (4 × 5) node SOM representing classified atmospheric states together with the frequency of SOM

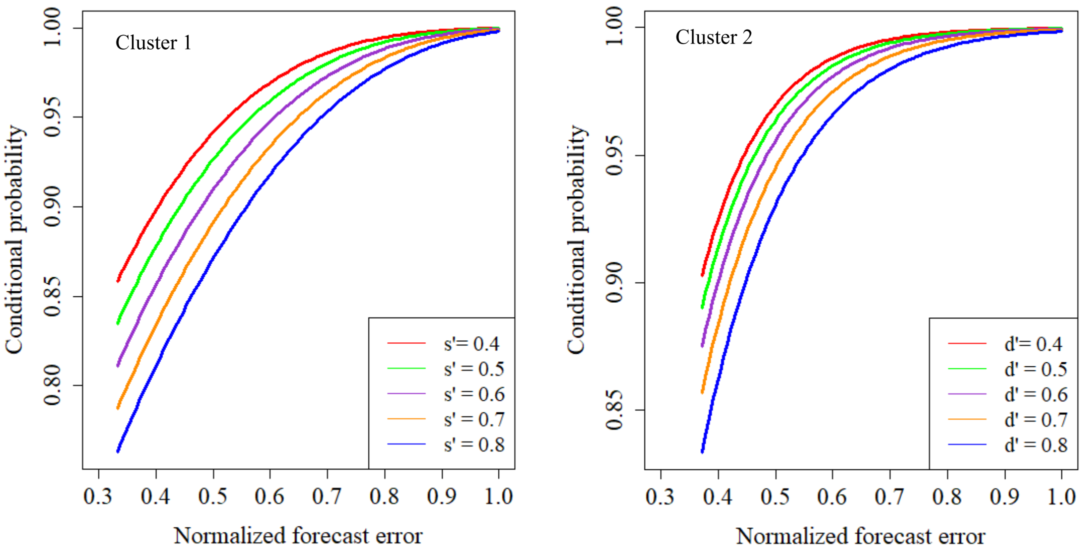

**FIGURE 10.** Conditional forecast error distribution of clusters 1 and 2, given that the forecast error of clusters 2 and 1 exceeds $s'$ and $d'$, respectively.

node occurrence as a percentage above each node. Table 2 shows the return level associated with each SOM node at different return periods. It is evident from Table 2 that the return level can change significantly due to changes in the SOM node. For example, the return level drops significantly (for all return periods) at SOM nodes (2, 5), (3, 4), and (4, 2) for Cluster 1, while the same happens at (1, 5) and (2, 4) for Cluster 2. Each of these nodes illustrates dominant high-pressure circulation over the respective clusters, particularly the ridging of the Indian Ocean High-Pressure System.

### D. COPULAS APPLICATION

As seen in Section III-C, the results from different clusters can be contradictory and it can be challenging for system operators to make system-wide decisions based on the univariate analysis. This paper uses copula functions to model the bivariate joint distribution from the univariate distributions obtained in Section III-B.

To obtain a joint distribution using copulas, the first step is to select the appropriate copula function. The Gaussian copula is widely used due to its simplicity [21], [48]. However, the Gaussian copula lacks the flexibility to model the tail dependence. As a result, this paper uses the t-Student copula, which is a realistic function for modeling tail dependence [21], [48].

Once obtaining the bivariate joint distribution, we can conduct a wide range of probabilistic analyses on system-wide forecast error without losing spatial-temporal correlations between clusters. For example, we can analyze the probability that both clusters simultaneously exceed certain thresholds ($s$ and $d$) using (9). If $d = s = 0.44$, then $F_D(d)$, $F_S(s)$ and $C(F_D(d), F_S(s))$ are equal to 0.96, 0.98 and 0.94, respectively. Therefore, the probability that the forecast error of both clusters will exceed 0.44 is 0.0035.

**TABLE 3.** Average running time of the proposed models.

| Model | Running time (s) |
|---|---|
| Parameter estimation of GPD | 0.11 |
| Classification of atmospheric states per time step | 0.23 |
| Copula application | 61.14 |

With copulas, it is also easy to derive the conditional forecast error distribution of Cluster 1 given that the forecast error of Cluster 2 exceeds a certain threshold, and vice versa. Fig. 10 illustrates the conditional forecast error distribution of clusters 1 and 2, given that clusters 2 and 1 exceed various forecast error thresholds ($s'$ and $d'$), respectively. From Fig. 10 we can deduce, for example, that the probability that the forecast error of Cluster 1 is less than 0.37 given that the forecast error of Cluster 2 exceeds 0.64 is equal to 0.83.

### E. RUNNING TIME

To assess the computational cost, the average running times of the models that are part of the proposed methodology are listed in Table 3. All models were tested on a Windows PC with 2 GHz and 8 GB RAM. The creation of the SOM-map can be a slow process especially when working with large datasets and depends on the SOM set-up, initialization and training parameters (i.e. number of training iterations, number of SOM nodes, size of neighborhood function, training algorithm etc.). It took approximately 48 hours to create the SOM map used in this study. It should however be noted that, in the implementation of the methodology described in this paper, the creation of the initial SOM is a once off procedure. The operational application of the proposed methodology will be in classifying each atmospheric state based on the set of SOM nodes already created, the average running time of which is described in Table 3.

## IV. CONCLUSION

This paper has demonstrated that some of the common distributions (normal, hyperbolic, Weibull, and beta) currently used for modeling wind power forecast errors can be inappropriate in representing extreme forecast errors. The paper then modeled the extreme forecast errors using the EVT by fitting the GPD. The GPD showed a superior representation of extreme forecast errors as compared to the commonly used distributions. The paper also estimated the conditional GPDs associated with each hour of day, month of year, and SOM node. It was found that extreme forecast errors can have strong diurnal and seasonal components depending on the location of wind farms under consideration. Therefore, diurnal and seasonal cycles play an important role in the occurrence of extreme forecast errors and can improve estimation thereof. In addition, extreme forecast errors can also change significantly from one SOM node to the other. The dominant high-pressure circulation, particularly the ridging of the Indian Ocean High-Pressure System is associated with reduced extreme forecast errors. This not only improves the estimation of extreme forecast errors but also allows for the estimation of extreme forecast errors based on physical meteorological phenomena. This paper then used the copula functions to estimate the bivariate joint forecast error distribution of different wind generation clusters. With numerical examples, this paper showed that copulas could be effective in providing a wide range of probabilistic analyses, giving more insight into the characteristics of region-wide extreme forecast errors.

The most significant contribution made by this paper is in improving the estimation and understanding of extreme forecast errors. This is an important step toward better allocation of operating reserves to account for wind power uncertainty. In future, one can test the proposed methodology using individual wind farms' data and not as clusters. This will ensure more spatial-temporal information is extracted from the data, which could further improve the estimation of extreme forecast errors in a region. Furthermore, because the power grid is often made up of a mix of variable renewable sources, it may be valuable to evaluate the applicability of the proposed methodology to other variable renewable sources.

## REFERENCES

[1] B. Hodge, D. Lew, M. Milligan, E. Gómez-Lázaro, D. Flynn, and J. Dobschinski, "Wind power forecasting error distributions: An international comparison," in *Proc. 11th Annu. Int. Workshop Large-Scale Integr. Wind Power Power Syst. Well Transmiss. Netw. Offshore Wind Power Plants Conf.*, 2012, pp. 1–8.

[2] Z.-S. Zhang, Y.-Z. Sun, D. W. Gao, J. Lin, and L. Cheng, "A versatile probability distribution model for wind power forecast errors and its application in economic dispatch," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3114–3125, Aug. 2013.

[3] T. Ouyang, X. Zha, and L. Qin, "A survey of wind power ramp forecasting," *Energy Power Eng.*, vol. 5, no. 4, pp. 368–372, 2013.

[4] M. A. Ortega-Vazquez and D. S. Kirschen, "Estimating the spinning reserve requirements in systems with significant wind power generation penetration," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 114–124, Feb. 2009.

[5] R. Doherty and M. O'Malley, "A new approach to quantify reserve demand in systems with significant installed wind capacity," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 587–595, May 2005.

[6] V. S. Pappala, I. Erlich, K. Rohrig, and J. Dobschinski, "A stochastic model for the optimal operation of a wind-thermal power system," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 940–950, May 2009.

[7] F. Bouffard and F. D. Galiana, "Stochastic security for operations planning with significant wind power generation," in *Proc. IEEE Power Energy Soc. Gen. Meeting-Convers. Del. Electr. Energy 21st Century*, Jul. 2008, pp. 1–11.

[8] K. Methaprayoon, C. Yingvivatanapong, W.-J. Lee, and J. R. Liao, "An integration of ANN wind power estimation into unit commitment considering the forecasting uncertainty," *IEEE Trans. Ind. Appl.*, vol. 43, no. 6, pp. 1441–1448, Nov./Dec. 2007.

[9] E. D. Castronuovo and J. A. P. Lopes, "On the optimization of the daily operation of a wind-hydro power plant," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1599–1606, Aug. 2004.

[10] J. Wu, B. Zhang, H. Li, Z. Li, Y. Chen, and X. Miao, "Statistical distribution for wind power forecast error and its application to determine optimal size of energy storage system," *Int. J. Elect. Power Energy Syst.*, vol. 55, pp. 100–107, Feb. 2014.

[11] S. Tewari, C. J. Geyer, and N. Mohan, "A statistical model for wind power forecast error and its application to the estimation of penalties in liberalized markets," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2031–2039, Nov. 2011.

[12] H. Bludszuweit, J. A. Dominguez-Navarro, and A. Llombart, "Statistical analysis of wind power forecast error," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 983–991, Aug. 2008.

[13] J. M. Lujano-Rojas, G. J. Osorio, J. C. O. Matias, and J. P. S. Catalao, "Wind power forecasting error distributions and probabilistic load dispatch," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2016, pp. 1–5.

[14] B.-M. Hodge and M. Milligan, "Wind power forecasting error distributions over multiple timescales," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2011, pp. 1–8.

[15] B.-M.-S. Hodge, E. G. Ela, and M. Milligan, "Characterizing and modeling wind power forecast errors from operational systems for use in wind integration planning studies," *Wind Eng.*, vol. 36, no. 5, pp. 509–524, Oct. 2012.

[16] B.-M. Hodge, D. Lew, and M. Milligan, "Short-term load forecast error distributions and implications for renewable integration studies," in *Proc. IEEE Green Technol. Conf. (GreenTech)*, Apr. 2013, pp. 435–442.

[17] D. D. Tung and T. Le, "A statistical analysis of short-term wind power forecasting error distribution," *Int. J. Appl. Eng. Res.*, vol. 12, no. 10, pp. 2306–2311, 2017.

[18] P. Pinson, "Estimation of the uncertainty in wind power forecasting," Ph.D. dissertation, Dept. Math. Syst., Mines ParisTech, Paris, France, 2006.

[19] G. Liao, J. Ming, B. Wei, H. Xiang, N. J. P. Ai, C. Dai, X. Xie, and M. Li, "Wind power prediction errors model and algorithm based on non-parametric kernel density estimation," in *Proc. 5th Int. Conf. Electric Utility Deregulation Restructuring Power Technol. (DRPT)*, Nov. 2015, pp. 1864–1868.

[20] A. Z. Zambom and R. Dias, "A review of kernel density estimation with applications to econometrics," *Int. Econ. Rev.*, vol 5, no. 1, pp. 20–42, 2012.

[21] G. D'Amico, F. Petroni, and F. Prattico, "Wind speed prediction for wind farm applications by extreme value theory and copulas," *J. Wind Eng. Ind. Aerodyn.*, vol. 145, pp. 229–236, Oct. 2015.

[22] E. C. Morgan, M. Lackner, R. M. Vogel, and L. G. Baise, "Probability distributions for offshore wind speeds," *Energy Convers. Manage.*, vol. 52, no. 1, pp. 15–26, Jan. 2011.

[23] D. Ganger, "Enhanced power system operation performance with anticipatory control under increased penetration of wind energy," Ph.D. dissertation, School Elect., Comput. Energy Eng., Arizona State Univ., Tempe, AZ, USA, 2016.

[24] M. Fripp and R. H. Wiser, "Effects of temporal wind patterns on the value of wind-generated electricity in California and the Northwest," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 477–485, May 2008.

[25] S. Rehman, "Wind energy resources assessment for Yanbo, Saudi Arabia," *Energy Convers. Manage.*, vol. 45, nos. 13–14, pp. 2019–2032, Aug. 2004.

[26] B. Karki and R. Billinton, "Effects of seasonality and locality on the operating capacity benefits of wind power," in *Proc. IEEE Electr. Power Energy Conf. (EPEC)*, Oct. 2009, pp. 1–6.

[27] K. Knorr, B. Zimmermann, S. Bofinger, A. Gerlach, T. Bischof-Niemz, and C. Mushwana, "Wind and solar PV resource aggregation study for South Africa," Council Sci. Ind. Res., Pretoria, South Africa, RFP Rep. 542-23-02-2015, 2016.

[28] F. M. Mulder, "Implications of diurnal and seasonal variations in renewable energy generation for large scale energy storage," *J. Renew. Sustain. Energy*, vol. 6, no. 3, May 2014, Art. no. 033105.

[29] R. Carapellucci and L. Giordano, "The effect of diurnal profile and seasonal wind regime on sizing grid-connected and off-grid wind power plants," *Appl. Energy*, vol. 107, pp. 364–376, Jul. 2013.

[30] L. Zhou, Y. Tian, S. Baidya Roy, Y. Dai, and H. Chen, "Diurnal and seasonal variations of wind farm impacts on land surface temperature over western Texas," *Climate Dyn.*, vol. 41, no. 2, pp. 307–326, Jul. 2013.

[31] A. Dalton, B. Bekker, and M. J. Koivisto, "Simulation and detection of wind power ramps and identification of their causative atmospheric circulation patterns," *Electr. Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106936.

[32] A. Dalton, B. Bekker, and M. J. Koivisto, "Classified atmospheric states as operating scenarios in probabilistic power flow analysis for networks with high levels of wind power," *Energy Rep.*, vol. 7, pp. 3775–3784, Nov. 2021.

[33] A. Dalton, B. Bekker, and M. J. Koivisto, "Atmospheric circulation archetypes as clustering criteria for wind power inputs into probabilistic power flow analysis," in *Proc. Int. Conf. Probabilistic Methods Appl. Power Syst. (PMAPS)*, Aug. 2020, pp. 1–6.

[34] Y. Zhang and J. Wang, "K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting," *Int. J. Forecasting*, vol. 32, no. 3, pp. 1074–1080, Jul. 2016.

[35] W. Hu, Y. Min, Y. Zhou, and Q. Lu, "Wind power forecasting errors modelling approach considering temporal and spatial dependence," *J. Mod. Power Syst. Clean Energy*, vol. 5, no. 3, pp. 489–498, May 2017.

[36] G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertainty analysis," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 40–49, Feb. 2009.

[37] M. Yang, Y. Lin, S. Zhu, X. Han, and H. Wang, "Multi-dimensional scenario forecast for generation of multiple wind farms," *J. Mod. Power Syst. Clean Energy*, vol. 3, no. 3, pp. 361–370, Sep. 2015.

[38] N. Zhang, C. Kang, Q. Xia, and J. Liang, "Modeling conditional forecast error for wind power in generation scheduling," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1316–1324, May 2014.

[39] R. Huth, C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynová, J. Kyselý, and O. E. Tveito, "Classifications of atmospheric circulation patterns," *Ann. New York Acad. Sci.*, vol. 1146, no. 1, pp. 105–152, Dec. 2008.

[40] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.

[41] Copernicus Climate Change Service, *ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate*, Copernicus Climate Change Service Climate Data Store (CDS), Bonn, Germany, 2017.

[42] R. J. Davy, M. J. Woods, C. J. Russell, and P. A. Coppin, "Statistical downscaling of wind variability from meteorological fields," *Boundary-Layer Meteorol.*, vol. 135, no. 1, pp. 161–175, Apr. 2010.

[43] J. Chen, X. Lei, L. Zhang, and B. Peng, "Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang, China," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0118521.

[44] K. Sharma, V. Chavez-Demoulin, and P. Dillenbourg, "An application of extreme value theory to learning analytics: Predicting collaboration outcome from eye-tracking data," *J. Learn. Anal.*, vol. 4, no. 3, pp. 140–164, Dec. 2017.

[45] Y. Liu, R. H. Weisberg, and C. N. Mooers, "Performance evaluation of the self-organizing map for feature extraction," *J. Geophys. Res., Oceans*, vol. 111, no. C5, 2006, Art. no. C05018.

[46] L. Bhangwandin, "Multivariate extreme value theory with an application to climate data in the Western Cape," M.S. thesis, Dept. Stat. Sci., Univ. Cape Town, Cape Town, South Africa, 2017.

[47] M. Rydman, "Application of the peaks-over-threshold method on insurance data," Dept. Math., Uppsala Univ., Uppsala, Sweden, U.D.D.M. Project Rep. 2018:32, 2018.

[48] A. AghaKouchak, S. Sellers, and S. Sorooshian, "Methods of tail dependence estimation," in *Extremes in a Changing Climate*, 1st ed. Dordrecht, The Netherlands: Springer, 2013, ch. 6, pp. 163–179.

**NDAMULELO MARARAKANYE** received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Cape Town, South Africa, in 2013 and 2017, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with Stellenbosch University. He is a Senior Engineer with the Centre for Renewable and Sustainable Energy Studies. His research interests include variable renewable energy grid integration studies, forecasting, characterizing uncertainty, and power system operations.

**AMARIS DALTON** (Member, IEEE) received the B.Sc. degree (Hons.) in meteorology and the M.Sc. degree in environmental management from the University of Pretoria, in 2014 and 2017, respectively, and the Ph.D. degree in electrical engineering from Stellenbosch University, in 2020, working under the supervision of Dr Bernard Bekker. From 2015 to 2018, he worked for Eskom as an Environmentalist. Since 2021, he has been working as a Postdoctoral Research Fellow with the Centre for Renewable and Sustainable Energy Studies, Stellenbosch University. His research interests include energy meteorology, wind power prediction, and power systems modeling.

**BERNARD BEKKER** (Member, IEEE) received the B.Eng. and M.Eng. degrees in electrical engineering from Stellenbosch University, South Africa, in 1996 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Cape Town, South Africa, in 2010. He is currently the Eskom Research Chair of power system simulation with Stellenbosch University, where he is also an Associate Director with the Centre for Renewable and Sustainable Energy Studies. His research interests include the grid integration of distributed energy resouces, probabilistic modeling, and variable renewable energy forecasting.