

Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model

Thakhani Ravele^{1,*}, Caston Sigauke¹, Lordwell Jhamba²

¹Department of Mathematical and Computational Sciences, University of Venda, South Africa

²Department of Physics, University of Venda, South Africa

Received September 13, 2021; Revised November 2, 2021; Accepted November 21, 2021

Cite This Paper in the following Citation Styles

(a): [1] Thakhani Ravele, Caston Sigauke, Lordwell Jhamba, "Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model," *Mathematics and Statistics*, Vol.10, No.1, pp. 116-133, 2022. DOI: 10.13189/ms.2022.100109

(b): Thakhani Ravele, Caston Sigauke, Lordwell Jhamba, (2022). Estimation of Extreme Quantiles of Global Horizontal Irradiance: A Comparative Analysis Using an Extremal Mixture Model and a Generalised Additive Extreme Value Model. *Mathematics and Statistics*, 10(1), 116-133. DOI: 10.13189/ms.2022.100109

Copyright ©2022 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Solar power poses challenges to the management of grid energy due to its intermittency. To have an optimal integration of solar power on the electricity grid it is important to have accurate forecasts. This study discusses the comparative analysis of semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models for probabilistic forecasting of hourly global horizontal irradiance (GHI) at extremely high quantiles ($\tau = 0.95, 0.97, 0.99, 0.999$ and 0.9999). The data used are from the University of Venda radiometric in South Africa and are from the period 1 January 2020 to 31 December 2020. Empirical results from the study showed that the AQR-2 is the best fitting model and gives the most accurate prediction of quantiles at $\tau = 0.95, 0.97, 0.99$ and 0.999 , while at 0.9999 -quantile the GAEV model has the most accurate predictions. Based on these results it is recommended that the AQR-2 and GAEV models be used for predicting extremely high quantiles of hourly GHI in South Africa. The predictions from this study are valuable to power utility decision-makers and system operators when making high-risk decisions and regulatory frameworks that require high-security levels. This is the first application to conduct a comparative analysis of the proposed models using South African solar irradiance data, to the best of our knowledge.

Keywords Additive Quantile Regression, Global Horizontal Irradiance Forecasting, Penalised Cubic Regression Smoothing Spline, Quantile Regression via Asymmetric Laplace Distribution, Semi-parametric Extremal Mixture

1 Introduction

1.1 Background

The generation of power from clean energy sources makes an important contribution to sustainable development. Energy policy-makers are establishing more penetration of clean energy onto the grid, such as solar and wind power [1]. Variability of accurate

renewable energy forecasts and lack of consistency on renewable power generation are the limitations for using renewable energy sources (RES) to produce power for supplying peak energy demand [1]. Accurate forecasting models are useful to system operators and market players in solving problems created by the integration of RES and also viable to handle their variability and uncertainty [2]. However, this study focuses on modelling and forecasting global horizontal irradiance (GHI) at extremely high quantiles. Solar power is one of the most preferable RES to be integrated onto the grid because its contract price is cheaper since it depends on primary sources [3]. The energy sector needs to take the direction of exploring solar energy as it is one of the cleanest energy sources [4]. Accurate solar power forecasts are useful for an economic operation dispatch, optimal unit commitment and ensuring the stability of a national grid. It also reduces the uncertainties of solar energy sources and results in assuring safety and easier grid management. Lack of accurate forecasts of solar power generation creates problems for grid energy management [3, 5].

Solar irradiance forecasting is required to guide the grid operators to apply the relevant procedure to optimise electricity production and reduce production costs. Solar irradiance forecasts are used in an electrical grid for the integration of photovoltaic (PV) power [5]. Power system operators can have problems because of inaccurate extreme quantiles of GHI forecasts. Accurate extreme forecasts of GHI quantiles help decision-makers in making high-risk decisions and regulatory frameworks that require high-security levels. However, accurate forecasts are important for the effective operation of the electrical grid when solar energy is integrated. It has been shown that solar power forecasting is essential in different areas of planning operations in the energy sector such as unit dispatch and renders schedules of production of power from renewable energy sources for the next hours or days. These forecasts are also used to know in advance the amount of solar power that will be integrated into the grid in the following hours or days. The purpose of producing new forecasting models for solar power is to provide accurate forecasts [6].

1.2 A review of the solar irradiance forecasting literature

Accurate forecasts are very important in the energy utility because operators rely on them to maintain the uncertainty created when the penetration of solar energy increases [5]. According to [7], accurate PV power forecasting is required for electric power system operation during the installation of grid-connected PV generation on a huge scale. Application of PV power forecasting is not only used in decision making for grid dispatch, it is also used to complement the control of other multiple power sources. Potential statistical modelling errors including changing weather patterns pose challenges in forecasting accurately solar power [8]. To reduce the intermittent solar power and operational expenses when integrating it on the electricity grid it is important to have accurate forecasts and also capture the uncertainty surrounding the forecasts [9].

Reliable accurate forecasts are required to maintain the balance of fluctuating PV energy generation and also make the integration to render better quality service [10]. Solar power forecasts are mainly classified periodically, ranging from very short-term, short-term, medium-term and long-term [11]. According to [11], very short-term forecasts consist of 15 minutes to a few hours ahead with a granularity of 30 seconds to 5 minutes, which are used for events ramping and variability related to operations. For short-term forecasts, twenty-four up to 168 hours ahead with a granularity of hourly. These forecasts are useful for transmission scheduling including unit commitment. Medium-term forecasts which are important for planning and asset optimisation range from a week to a year ahead. Long-term forecasts are generated from a year up to several years ahead and are used for planning of the extension of the electricity grid including the building of new power plants [11].

Many researchers have been focusing on short-term forecasting with an hourly cycle [7]. However, the focus of this study is on hourly GHI forecasting at extremely high quantiles. According to references [12] and [13], solar irradiance short-term forecasts are essential for operational planning including scheduling of power systems. PV power forecasting consists of models such as physical, statistical, machine learning, and hybrid integration models [7]. The forecasting techniques of solar irradiance are typically separated into physical and statistical approaches including machine learning algorithms [10, 14]. Physical approaches use numerical weather predictions (NWP) and statistical approaches use historical time solar time series data [3, 4, 10, 14]. There are three types of solar irradiance which are direct normal irradiance (DNI), diffuse horizontal irradiance (DHI) and global horizontal irradiance (GHI), respectively [12]. GHI is the sum of DNI and DHI.

Solar energy plays an important role in electricity markets [5]. Reference [5] compiled a review of solar power forecasting focusing on the current improvements and future projections of solar energy. The idea of the study was to present a precise forecast with its analysis of the economic implications. The study revealed that most of the research has been done focusing on day-ahead forecast horizon [5]. Accurate solar power forecasts are required by utilities and independent system operators, as well as solar power producers and energy traders for core operations [11]. Reference [11] proposed a set of standards for performance evaluation of solar power forecasts. The standards consist of sound methodologies and a lot of field experience, which provide accurate inter-agency comparisons of forecast performance [11]. A hybrid short-term forecasting model was proposed by reference [7] to solve the problem of forecasting accuracy using solar data from Yunnan PV power generating plant. Several forecast evaluation metrics were used which include normalised absolute average error, normalised root-mean-square

error, among others. Results from the study showed that the proposed model was the best fitting model [7].

Solar irradiance forecasting using South African data is discussed in the literature. In a recent study, reference [14] predicted hourly global horizontal irradiance using data from the University of Pretoria radiometric station. A comparative analysis of long short-term memory networks (LSTM), support vector regression (SVR) and feed-forward neural networks (FFNN) models was done with results showing that the FFNN model gives the most accurate forecasts. The forecasts from the individual models were then combined to get an ensemble LSTM-SVR-FFNN model via quantile regression averaging (QRA) and convex combination (CC). The ensemble LSTM-SVR-FFNN model via QRA was the best fitting model with the most accurate forecasts. The study suggested that the sufficient and thorough metrics for evaluation, as well as statistical tests, render more perspective into the proposed forecasting methods [14]. Using data from Tellerie radiometric station in the Northern Cape province of South Africa, reference [3] developed partially linear additive quantile regression models which were then used in forecasting day-ahead GHI. The selection of variables was done using the least absolute shrinkage and selection operator (Lasso) via hierarchical interactions. The forecasts from the individual models were then combined using QRA and CC. The study revealed that QRA yields the most accurate forecasts [3].

Reference [15] used a multiple linear regression model to forecast global solar irradiance in South Africa. Covariates used in the study were weather variables, traditional extraterrestrial irradiance including sunshine hours. The study showed that the solar irradiance models performance and accuracy in other areas improved by the inclusion of weather parameters [15]. Reference [16] solved the issue of long-range dependence ingrained in the solar irradiance data in South Africa using three models which are seasonal autoregressive fractionally integrated moving average (SARFIMA), harmonically Coupled SARIMA and regression model with SARFIMA error terms. In the study, additive quantile regression was used as a benchmark model. Empirical results from the study suggest that long memory is anti-persistent in all the developed models [16].

Reference [2] estimated extremely low (below 0.05-quantiles) and extremely high (above 0.95-quantiles) quantiles of wind power data from a wind power plant located in Galicia, Spain. The study applied conditional extreme value theory estimators by integrating gradient boosting trees and a truncated generalized Pareto distribution. Based on pinball losses, the proposed method outperformed the benchmark models. The study did not include the weather variables as additional covariates to improve the accuracy of the proposed model. However, this study will include temperature variables as a way to improve the precision of models and also use different models for probabilistic forecasting of distribution tails.

1.3 Research highlights and contributions

The main contribution of this study is that it compares semi-parametric extremal mixture (SPEM), generalised additive extreme value (GAEV) or quantile regression via asymmetric Laplace distribution (QR-ALD), additive quantile regression (AQR-1), additive quantile regression with temperature variable (AQR-2) and penalised cubic regression smoothing spline (benchmark) models in predicting extremely high hourly GHI data. To the best of our knowledge, this is the first study to do a comparative analysis of the proposed models using data from South Africa. The following is an overview of the study's highlights and findings:

- The forecasts from SPEM, GAEV, BM, AQR-1 and AQR-2 models were integrated using simple average (AVG) and median (MED) methods to develop AVG and MED forecasts,
- Based on both the continuous rank probability score (CRPS) and pinball loss function (PLF), the AQR-2 model was identified to be the most accurate model for quantiles $\tau = 0.95, 0.97, 0.99$ and 0.999 and the GAEV model was also identified to be the most accurate model for quantile $\tau = 0.9999$,
- The results from this study yield improved forecasts for extremely high quantiles of hourly global horizontal irradiance (GHI),
- This is the first study to compare the SPEM, GAEV, BM, AQR-1 and AQR-2 models in estimating extremely high quantiles of hourly GHI using data from South Africa,
- Providing accurate solar power forecasts which are needed in the integration of solar power in the electricity grid.

The sections of the study are arranged as follows: Section 2 discusses the proposed models. Section 3 presented the empirical results and discussion while conclusion of the study presented Section 4.

2 The proposed models

2.1 Semi-parametric extremal mixture model

The proposed modelling framework is based on the work of [17]. Let $F_t(x) = P(X \leq x | T = t)$, where X is a random variable given at time index $t \in [0, T_{max}]$. The excess distribution function over a sufficiently high threshold μ is then given by [17, 18]:

$$F_{t,\mu}(x) = 1 - \frac{1 - F_t(x)}{1 - F_t(\mu)}, \quad x \in [\mu, \infty), \tag{1}$$

where the cumulative distribution function (CDF) F_t is assumed to be in the domain of attraction of the Fréchet class of distributions. The CDF $F_{t,\mu}(x)$ can be approximated by a Pareto distribution, which is defined by:

$$G_{\mu,\theta}(x) = 1 - \left(\frac{x}{\mu}\right)^{-\frac{1}{\theta}}, \quad x \in [\mu, \infty), \tag{2}$$

where the parameter $\theta > 0$ denotes the conditional tail index (extremal index) and an unknown threshold $\mu \geq x_0$, which depends on t . The distribution function F_t can be approximated by the empirical distribution function. The semi-parametric extremal mixture model (SPEM) is defined by [17, 18]:

$$y_{t,\mu,\theta} = \begin{cases} F_t(x_t), & \text{if } x \in [x_0, \mu], \\ 1 - (1 - F_t(x_t))(1 - G_{\mu,\theta}(x_t)), & \text{if } x > \mu, \end{cases} \tag{3}$$

where $y_{t,\mu,\theta}$ is the global horizontal irradiance (GHI). From equation (3), $F_t(x_t)$ is a bulk model, in which we fit a non-parametric model (kernel density) and $1 - (1 - F_t(x_t))(1 - G_{\mu,\theta}(x_t))$ is the tail model, in which we fit the Pareto distribution. In the study we use the kernel estimator to estimate the parameter θ . A point-wise data driven procedure is used to choose a threshold μ . The bandwidth denoted by h will be determined by minimising the cross-validation function. Let $K(\cdot)$ denote a kernel function which is assumed to be non-negative continuous and symmetric to the real line such that $K(x) \leq 1$. The set of weights for $i = 1, \dots, n$ defined by the kernel function is given by:

$$W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right), \tag{4}$$

where bandwidth parameter $h > 0$.

2.1.1 Threshold selection

In reference, [19], extremal mixture models for estimating sufficiently high thresholds are proposed to overcome the drawback of fixed threshold models which do not capture threshold uncertainty. In their paper, [19] grouped the threshold methods into four classes which are the classical fixed threshold, re-sampling based, tail fraction estimation and extremal mixture models. Reference [20] developed an extremal mixture model which combines a non-parametric kernel density for the bulk distribution and GDP for the tail distribution. Uncertainty around threshold selection and parameters estimation is addressed by Bayesian inference [21]. In a recent study, reference [22] discussed an automatic threshold selection method. Reference [21] developed a time-varying threshold which is coupled with a positive shift factor. In this present study, the selection of the threshold is done by testing a goodness-of-fit test discussed by reference [18], for the parametric part of the model 3. The k upper statistics are used to test the tail adjustment of Pareto distribution. If the test is not rejected, the number k of upper statistics is then increased. The tail adjustment is tested again until it is rejected.

2.1.2 Inference

The maximisation of the weighted quasi-log-likelihood estimator is defined by [17, 18]:

$$\hat{\theta}_{t,h,\mu} = \frac{1}{\hat{n}_{t,h,\mu}} \sum_{i=1}^n W_{t,h}(t_i) I_{\{X_{t_i} > \mu\}} \log\left(\frac{X_{t_i}}{\mu}\right), \tag{5}$$

where $\hat{n}_{t,h,\mu} = \sum_{i=1}^n W_{t,h}(t_i) I_{\{X_{t_i} > \mu\}}$ represents the weighted number of observations above the threshold μ . The $F_t(x_t)$ is estimated by:

$$\hat{y}_{t,h,\mu} = \begin{cases} \hat{F}_{t,h}(x_t), & \text{if } x \in [x_0, \mu], \\ 1 - (1 - \hat{F}_{t,h}(\mu))(1 - G_{\mu,\hat{\theta},h,\mu}(x_t)), & \text{if } x > \mu, \end{cases} \tag{6}$$

For any $\tau \in [0.99, 1)$ the estimator of the τ quantile of X_t is given by:

$$\hat{q}_\tau(t, h) = \begin{cases} \hat{F}_{t,h}^{-1}(\tau), & \text{if } \tau < \hat{\tau}_\mu, \\ \mu \left(\frac{1-\hat{\tau}_\mu}{1-\tau} \right)^{\hat{\theta}_{t,h,\mu}}, & \text{otherwise,} \end{cases} \quad (7)$$

and $\hat{\tau}_\mu = \hat{F}_{t,h}(\mu)$.

2.2 Generalised additive extreme value model

The high threshold $\mu(x)$ is estimated using quantile regression (QR) based on a pre-specified quantile τ . The likelihood-based inference is achieved by comparing the tilted loss function used in QR and the asymmetric Laplace distribution (ALD). Let $Y(x)$ be a random variable indexed by a covariate x [23, 24]. Then:

$$Y(x) \sim ALD(\mu(x), \sigma(x), \tau(x)), \quad (8)$$

and the density function of ALD is given by [23, 24]:

$$f_{ALD,\tau}(y(x), \mu(x), \sigma(x), \tau(x)) = \frac{\tau(x)(1-\tau(x))}{\sigma(x)} \exp \left\{ -\rho_\tau(x) \left(\frac{y(x) - \mu(x)}{\sigma(x)} \right) \right\}, \quad (9)$$

where $\mu(x)$ corresponds to quantile $0 < \tau < 1$ at location x , $\sigma > 0$ and $\rho_\tau(y) = y(x)(\tau(x) - I\{y(x) < 0\})$ denotes the check function, for indicator function I . Assume generalised additive model (GAM) forms in covariate x defined by [23, 24]:

$$y_t = \beta_0 + \sum_{i=1}^I \sum_{j=1}^{J_i} \beta_{ij} b_{ij}(x_t) + \varepsilon_t, \quad (10)$$

where y_t is a GHI, β_{ij} denotes basis coefficients and b_{ij} is the basis functions.

2.2.1 Inference: Restricted maximum likelihood

In order to make inference for β , penalised log-likelihood function is defined by [23, 24]:

$$\ell(\beta_\lambda, \lambda) = \ell(\beta_\lambda) - \frac{1}{2} \beta^T S_\lambda \beta, \quad (11)$$

where $\lambda = (\lambda_1, \dots, \lambda_I)$ represents the smoothing parameters, S_λ denotes a penalty matrix with elements determined by the basis function b_{ij} that has been chosen. And penalty matrix is defined by $S_\lambda = \sum_{i=1}^I \lambda_i S_i$, where S_i denotes rows and columns of the matrix corresponds to $b_{i'j}$, $i' \neq i$ and consists of zeros. The parameters of extreme value distribution are estimated by a restricted maximum likelihood defined by [23, 24]:

$$\ell(\beta_\lambda) = \ell(\hat{\beta}_\lambda, \lambda) + \frac{1}{2} \log |S_\lambda|_+ - \frac{1}{2} \log |H(\beta_\lambda^T)| + \text{constant}, \quad (12)$$

where for given λ , $\hat{\beta}_\lambda$ maximises $\ell(\beta_\lambda, \lambda)$, $H(\beta_\lambda^T) = -\nabla \nabla^T \ell(\beta, \lambda)|_{\beta=\hat{\beta}_\lambda}$ and $|S_\lambda|_+$ is the product of positive eigenvalues of matrix S_λ .

2.2.2 Estimation of the threshold

The log-likelihood for the quantile regression threshold model from reference [24] is used for estimation of the threshold. Log-likelihood based on fixed τ is defined by [24]:

$$\log f(y; \beta) = \sum_{j=1}^J \sum_{t=1}^T \log f(y(s_j, t); u(s_j, t), \sigma(s_j, t)), \quad (13)$$

where $\mathbf{u} = (u(s_1, 1), \dots, u(s_J, T))'$ = $\mathbf{X}_u \beta_u$, $\sigma = (\sigma(s_1, 1), \dots, \sigma(s_J, T))'$, $\log \sigma = \mathbf{X}_\sigma \beta_\sigma$, \mathbf{X}_u and \mathbf{X}_σ denote design matrices. $\mathbf{S}_{u,\lambda}$ and $\mathbf{S}_{\sigma,\lambda}$ are penalty matrices.

2.2.3 Extremal index

The censored log-likelihood model is used for estimation of an extremal index, which is given by [24]:

$$\log f(y; \beta_\theta) = \sum_{j=1}^J \left[\sum_{t=1}^T I(\tilde{w}(s_j, t) > w_u) [\log\{2\theta(s_j, t)\} - 2 \log\{\tilde{w}(s_j, t)\}] - \frac{2\theta(s_j, t)}{\max\{\tilde{w}(s_j, t), w_u\}} \right], \tag{14}$$

where $w_u = \frac{-1}{\log \tau}$.

2.3 Additive quantile regression model

An additive quantile regression (AQR) model is one which combines GAM and QR models. The AQR model is defined by [25, 26, 27]:

$$y_{t,\tau} = \sum_{i=1}^p s_{i,\tau}(x_{ti}) + \varepsilon_{t,\tau}; \quad \tau \in (0, 1), \tag{15}$$

where $y_{t,\tau}$ is a GHI at time $t = 1, \dots, n$ at quantile τ , $s_{i,\tau}$ denote the smooth functions and $\varepsilon_{t,\tau}$ is the error term. In this study we are going to consider two scenarios: one AQR model (16) (AQR-1), which has one linear covariate t and two AQR model (17) (AQR-2) which has one linear covariate t and temperature variable. The two models are written as:

$$q_{Y|X}(\tau) = \sum_{t=1}^n \rho_\tau(y_{t,\tau} - s(t)), \quad \tau \in [0.95, 1), \tag{16}$$

$$q_{Y|X}(\tau) = \sum_{t=1}^n \rho_\tau(y_{t,\tau} - s(t) - s(\text{temp})), \quad \tau \in [0.95, 1), \tag{17}$$

2.4 Benchmark model for estimating extreme conditional quantiles

We first fit a penalised cubic regression smoothing spline given by [21]:

$$\pi_1(x) = \sum_{t=1}^n (y_t - f(x_t))^2 + \lambda \int (f''(x))^2 dx + \varepsilon_t, \tag{18}$$

where y_t denotes our GHI, λ is a smoothing parameter and ε_t is the error term. We then extract residuals $\varepsilon_t = \pi_1(x) - \hat{\pi}_1(x)$ and then estimate the shift factor $\mu_\tau \in \mathbb{R}$, $\tau \in [0.95; 1)$. The shift factor has to be sufficiently large to satisfy the asymptotic conditions when we fit the GPD [21].

$$\pi_2(x) = \sum_{t=1}^n (y_t - f(x_t))^2 + \lambda \int (f''(x))^2 dx + \varepsilon_t + \mu_\tau, \tag{19}$$

To estimate the shift factor μ_τ , threshold stability plots and extremal mixture models will be used. Table 1 presents a summary of some of the advantages and disadvantages of the proposed models, which are SPEM, GAEV, AQR and BM.

2.5 Evaluation of probabilistic forecasting methods and error measures

This section presents the methods for evaluation and comparison of probabilistic forecasts from proposed models of extreme quantiles. In this study two error measures will be used, are the continuous rank probability score (CRPS) and pinball loss function (PLF). When the quantile score is low, the forecasting model is more accurate.

2.5.1 Continuous ranked probability score

Continuous ranked probability score (CRPS) can handle both calibration and sharpness at the same time [28]. The CRPS is given by [2, 28, 29]:

$$\text{CRPS}(y, F) = \int_0^1 \text{QS}_\tau \left(F^{-1}(\tau), y \right) d\tau, \tag{20}$$

where F is the forecast distribution and QS_τ represents the quantile score defined by:

$$\text{QS}_\tau \left(F^{-1}(\tau), y \right) = 2 \left(\mathbf{I} \left[y \leq F^{-1}(\tau) \right] - \tau \right) \left(F^{-1}(\tau) - y \right), \tag{21}$$

where \mathbf{I} denotes an indicator function.

Table 1. Model comparisons.

Models	Strengths	Weaknesses
M1 (SPEM)	<ol style="list-style-type: none"> 1. Semi-parametric extremal mixture model. 2. Based on one covariate which is $t = 1, \dots, n$. 3. It is capable of processing inputs of any length. 4. The model size does not expand as the input grows larger. 5. It is possible to share weights across time steps. 	<ol style="list-style-type: none"> 1. Has limitations on accuracy and stability. 2. Very sensitive to numbers and location of the measured points. 3. The size does not expand due to its recurrence. Computation takes a long time.
M2 (GAEV)	<ol style="list-style-type: none"> 1. A hybrid model which fit QR via ALD. 2. Assume GAM forms in covariate x_t which is $t = 1, \dots, n$. 3. Robust to extreme high quantiles and probabilities of rare events. 	<ol style="list-style-type: none"> 1. Requires a smoothing parameter. 2. Has limitations on accuracy and stability. 3. Very sensitive to numbers and location of the measured points.
M3 (AQR)	<ol style="list-style-type: none"> 1. A hybrid model which combines GAMS with QR. 2. Estimation is distribution-free. 3. Robust to outliers in the response variable. 4. Perform global search and explore the search space using various crossover and mutation operators. 5. It is useful for noisy environments. 6. It is useful on mixed discrete/continuous problem. 	<ol style="list-style-type: none"> 1. Requires a smoothing function of the covariates. 2. Parameters are harder to estimate. 3. Does not give any details about the size of high level of possible exceedences. 4. Computationally expensive i.e. time consuming. 5. Implementation is still an art.
M4 (BM)	<ol style="list-style-type: none"> 1. Inference is done based on large sample approximation. 2. Robust to outliers in the response variable. 3. The transparent process makes it easy to implement and debug. 4. No training period, it is called lazy learner. 5. New data can be added seamlessly. 	<ol style="list-style-type: none"> 1. Requires a smoothing parameter. 2. Outliers only have influence on quantile curves close to them, i.e affect extreme quantiles. 3. Does not work well with large dataset. 4. Need feature scaling. 5. Sensitive to noisy data, missing values and outliers.

2.5.2 Pinball loss function

Even if CRPS is useful for measuring the quality of cumulative distribution function forecast, the study by reference [30] indicated that the CRPS is more applicable to distributions with smaller uncertainty intervals. However, a more suitable scoring rule is the PLF or quantile loss [28]. The PLF is defined by:

$$L(q_\tau, y_t) = \begin{cases} \tau(y_t - q_\tau), & \text{if } y_t > q_\tau, \\ (1 - \tau)(q_\tau - y_t), & \text{if } y_t \leq q_\tau, \end{cases} \tag{22}$$

where q_τ is a quantile forecast and y_t denotes the observed value of hourly GHI.

2.6 Combining the extreme quantiles

In this study, simple average and median methods are used to combine the extreme quantiles from the proposed models. The $100(1-\alpha)\%$ combined forecasts interval for \tilde{x} developed from the k individual intervals with heuristic H is denoted by $[L_H, U_H]$ [31].

2.6.1 Simple average method

A simple average method performs better than other sophisticated methods when combining forecasts. This method calculates a simple average of the endpoints of the intervals. Simple averages are used in combining point forecasts due to their robustness, simplicity and providing accurate forecasts.

The lower and upper averages models are defined by $L_{AV} = (1/k) \sum_{i=1}^k L_i$ and $U_{AV} = (1/k) \sum_{i=1}^k U_i$ respectively [31].

2.6.2 Median method

The median method is a known method used in summarising data because is less sensitive to extreme values than the mean. The median models are given by $L_{Md} = Median\{L_1, \dots, L_k\}$ and $U_{Md} = Median\{U_1, \dots, U_k\}$ [31].

3 Empirical results and discussion

This study used GHI data for the period, 1 January 2020 to 31 December 2020, giving a sample size of 4083 observations. The data is obtained from the Southern African Universities Radiometric Network (SAURAN) website (<http://www.sauran.net>). In this study, the station of focus is the University of Venda radiometric station found in an inland location at Vuwani Science Research Centre, in the Limpopo province of South Africa. It is on latitude -23.1310052, longitude 30.42399979 and elevation 628 m. Figure 1 shows a picture of the pyranometer at the University of Venda radiometric station. The study seeks to predict extremely high quantiles of GHI using R statistical software.

3.1 Exploratory data analysis

The summary statistics of hourly GHI are given in Table 2. Since the mean and median of the GHI data are not equal, the distribution is not normally distributed. The non-normality of the distribution of GHI data is confirmed by the skewness and kurtosis values which are 0.5734277 and -0.9057017, respectively. Figure 2 presents the time series plot of GHI together with

Table 2. Summary statistics for GHI.

Mean	Median	Minimum	Maximum	Standard deviation	Skewness	Kurtosis
355.2654	278.7552	0.0005	1121.8900	316.1626	0.5734	-0.9057

histogram and box plots. These plots suggest that the GHI data does not follow a Gaussian distribution. A plot of hourly GHI superimposed with a non-linear trend is shown in Figure 3 and this best fit was determined based on “mgcv” developed by reference [32].



Figure 1. UNIV – USAid Venda station. Source: <https://sauran.ac.za/>.

3.2 Forecasting results

The models considered in this study are the SPEM, GAEV, BM, AQR-1 and AQR-2. The R-packages such as “extremefit” developed by reference [18] is used for the SPEM, “evgam” developed by reference [23] is used for the GAEV model and “qgam” developed by reference [27] is used for AQR models. The quantiles considered in this study are $\tau = 0.95, 0.97, 0.99, 0.999$ and 0.9999 . All the data is used for predictions. The plots of actual hourly GHI and forecasts of GHI at quantiles ($\tau = 0.95, 0.97, 0.99, 0.999$ and 0.9999) using SPEM, GAEV, BM, AQR-1 and AQR-2 models are given in Figure 4 – 6 respectively. It is shown that the quantile predictions for each model follow the actual hourly GHI data remarkably well at a high quantile. The simple average (AVG) and median (MED) methods are used to combine forecasts for the SPEM, GAEV, BM, AQR-1 and AQR-2 models at each quantile. The plots of actual hourly GHI and forecasts of GHI using AVG and MED methods at quantiles ($\tau = 0.95, 0.97, 0.99, 0.999$ and 0.9999) are given in Figure 7.

3.2.1 Comparative Analysis

Table 3 gives a summary of the values of the evaluation metrics (PLF and CRPS) for the proposed models. At quantile levels, i.e $\tau = 0.95, 0.97, 0.99$ and 0.999 AQR-2 is the most accurate model based on the PLF and at $\tau = 0.9999$ GAEV is the most accurate model based on the PLF and CRPS.

Figure 8 shows the plots of actual hourly GHI and forecasts of GHI from the best-fitting model AQR-2 at quantiles $\tau = 0.95, 0.97, 0.99$ and 0.999 . Figure 9 also shows the plots of actual hourly GHI and forecasts of GHI from the best-fitting model GAEV at quantile $\tau = 0.9999$.

3.3 Discussion of results

This study was motivated by previous studies such as [3, 14], among others. The focus of the study was on a comparative analysis of SPEM, GAEV, BM, AQR-1 and AQR-2 models in the prediction of extremely high hourly GHI using South African data. The forecasts from SPEM, GAEV, BM, AQR-1 and AQR-2 models were combined using AVG and MED methods to develop AVG and MED forecasts. The PLF and CRPS were used to measure the accuracy of models at each quantile including AVG and MED forecasts. Based on both the PLF and CRPS, the best-fitting model for quantiles $\tau = 0.95, 0.97, 0.99$ and 0.999 was found to be AQR-2, while for the 0.9999 -quantile the best model is GAEV. The most accurate model was selected based on the smallest value of PLF and CRPS. It was revealed that fitting a model (i.e AQR-2) including temperature variable improves the forecasts because the temperature is the main predictor of GHI. Prediction of extremely high quantiles of GHI solar power could help system operators to know the possible largest solar power which can be integrated on the national grid. Accurate forecasts of

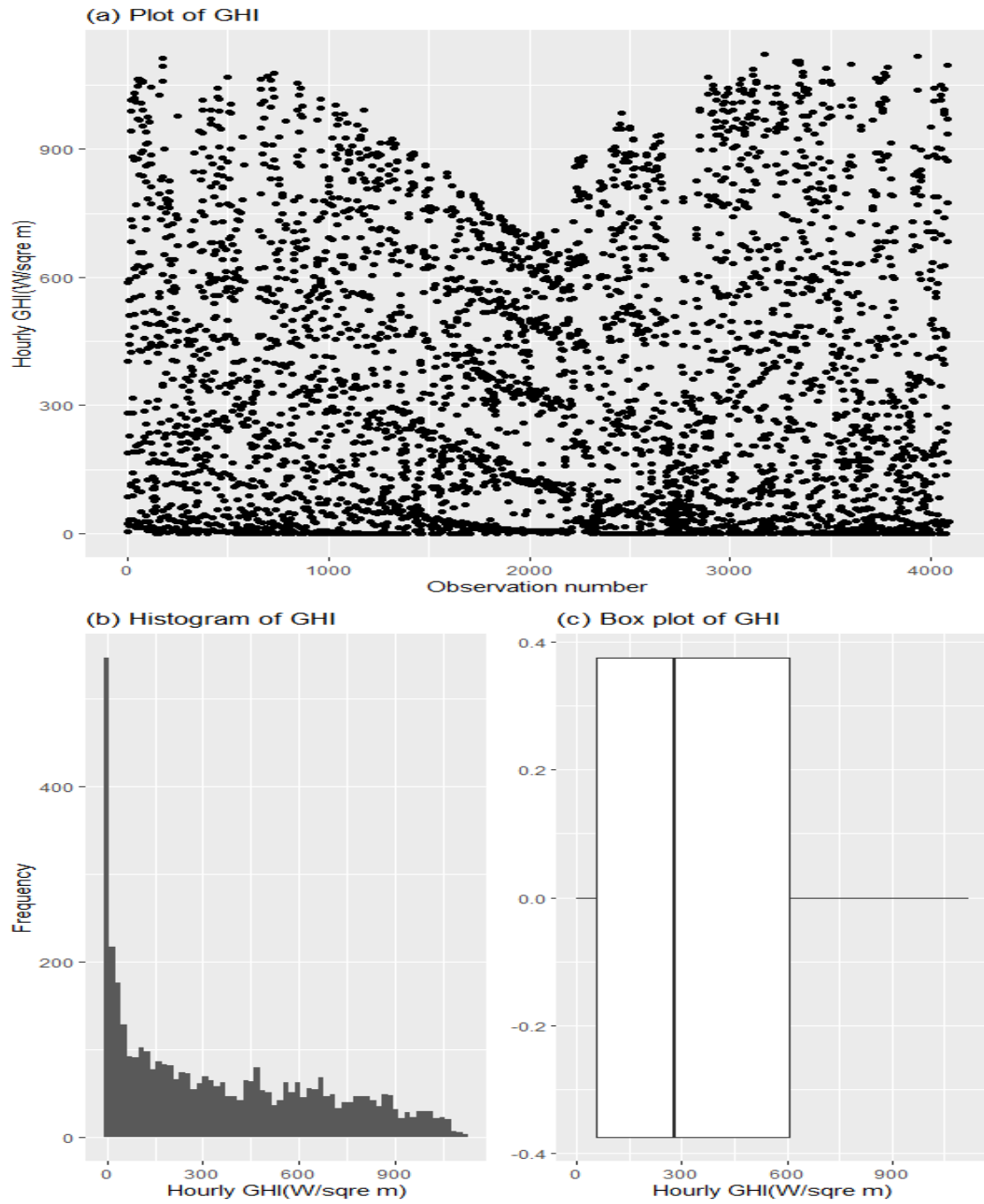


Figure 2. Diagnostic plots for hourly GHI from 01 January 2020 to 31 December 2020.

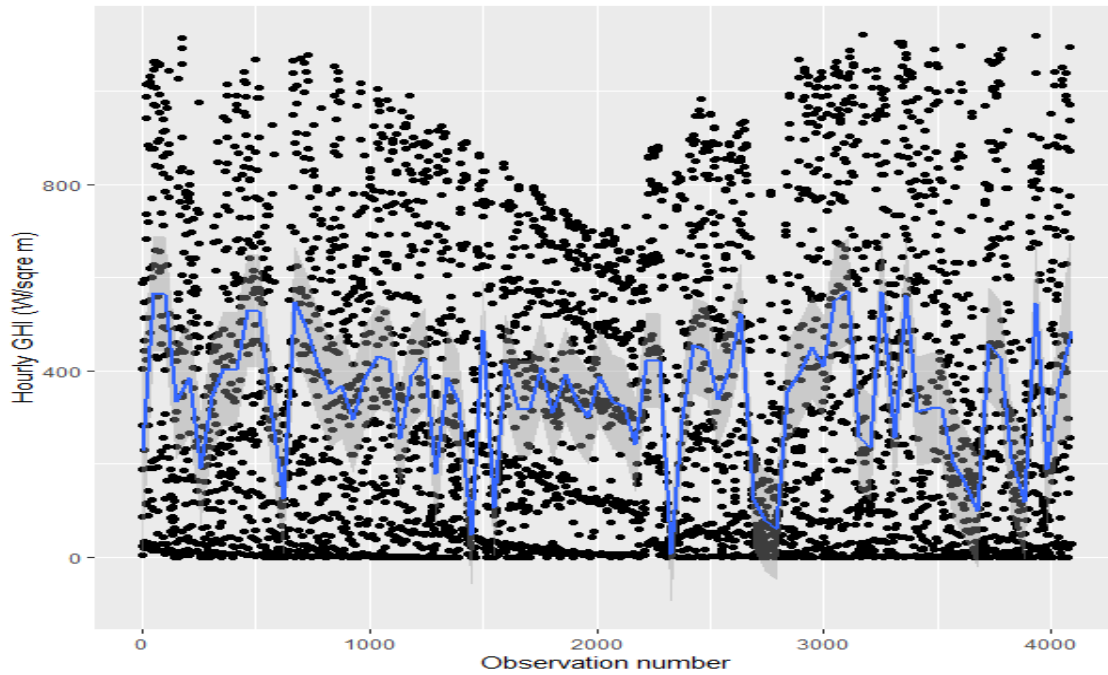


Figure 3. Plots of hourly GHI from 01 January 2020 to 31 December 2020 superimposed with a non-linear trend.

Table 3. Model comparisons using PLF and CRPS.

Accuracy	Model	$\tau = 0.95$	$\tau = 0.97$	$\tau = 0.99$	$\tau = 0.999$	$\tau = 0.9999$
PLF	SPEM	29.4337	18.2303	6.3971	0.7809	0.0976
	GAEV	28.7070	17.6628	5.9598	0.6294	0.0654
	BM	27.5458	17.5389	6.4691	0.7302	0.0744
	AQR-1	29.5307	18.3051	6.5218	0.7615	0.0861
	AQR-2	20.7937	13.6359	5.1719	0.5856	0.0763
	AVG	26.3300	16.5366	5.9588	0.6952	0.0798
	MED	28.3282	17.5617	6.1030	0.7024	0.0769
CPRS	SPEM	27226.53	27255.53	27315.11	27464.02	27658.15
	GAEV	27203.16	27240.34	27264.59	27311.81	27333.55
	BM	27167.32	27207.16	27271.90	27403.37	27426.28
	AQR-1	27235.74	27271.83	27333.23	27444.72	27543.50
	AQR-2	27045.44	27089.42	27177.64	27269.47	27445.83
	AVG	27175.64	27212.85	27272.49	27378.68	27481.46
	MED	27210.31	27245.64	27286.57	27385.88	27452.12

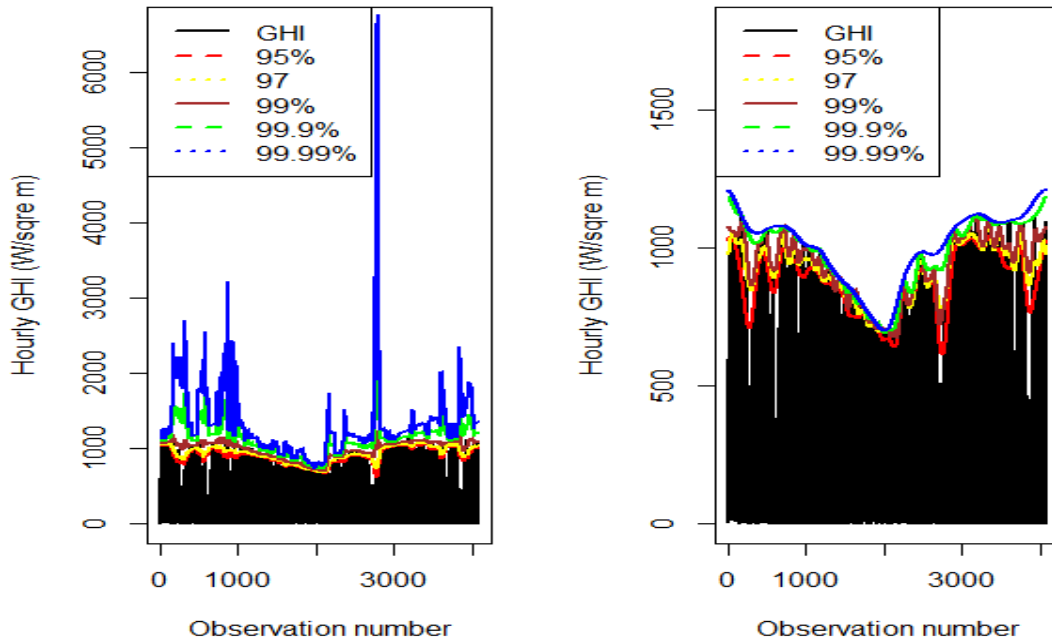


Figure 4. Plots of actual hourly GHI and forecasts of GHI at high quantiles using SPEM (left-side) and GAEV (right-side) models.

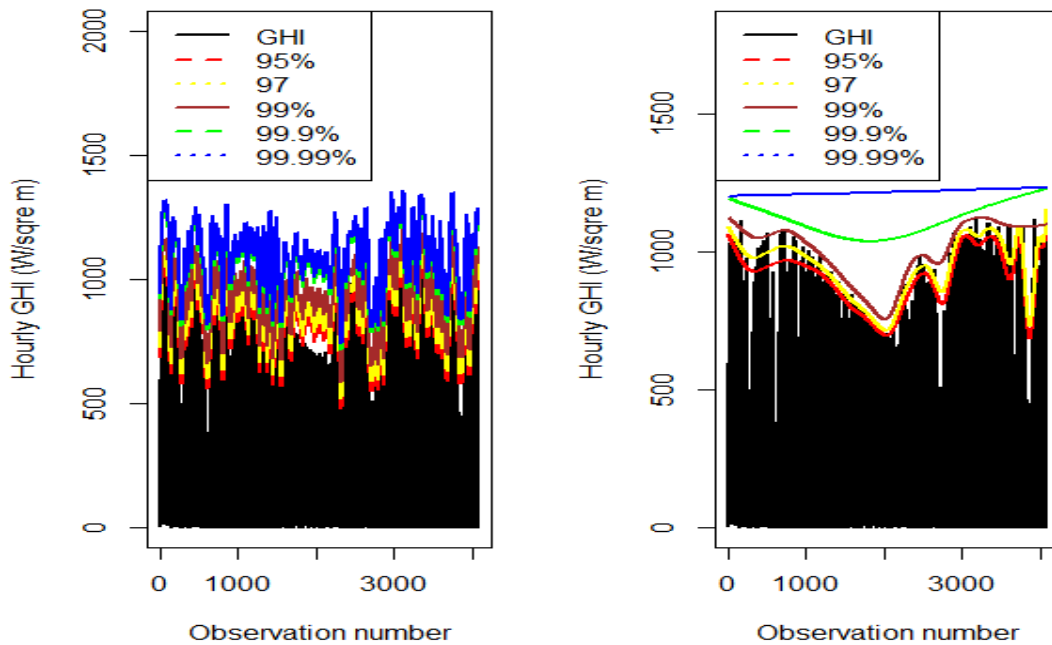


Figure 5. Plots of actual hourly GHI and forecasts of GHI at high quantiles using BM (left-side) and AQR-1 (right-side) models.

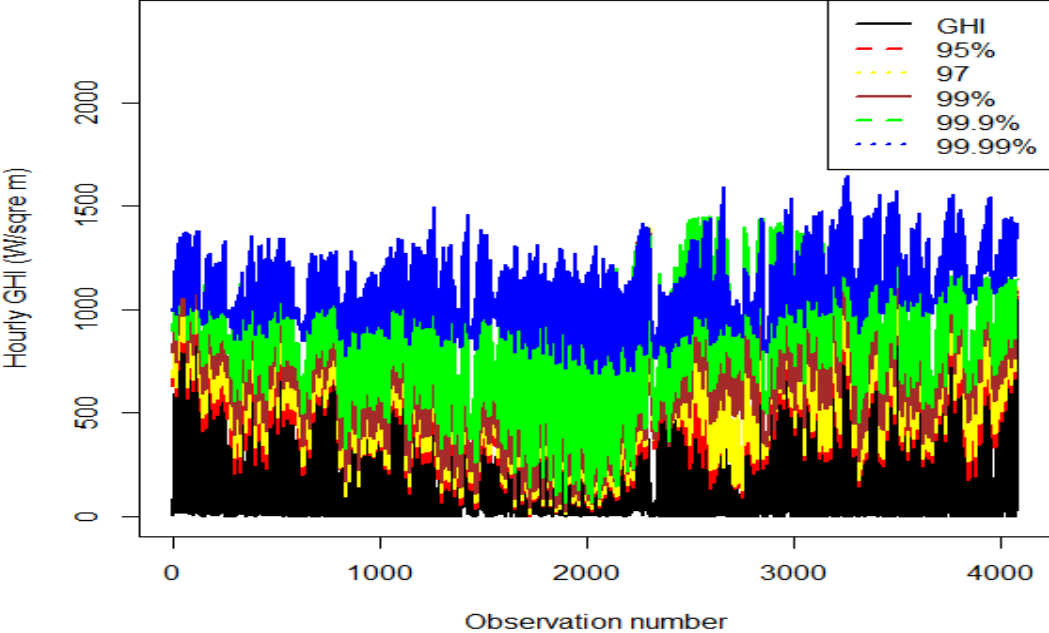


Figure 6. Plots of actual hourly GHI and forecasts of GHI at high quantiles using AQR-2 model.

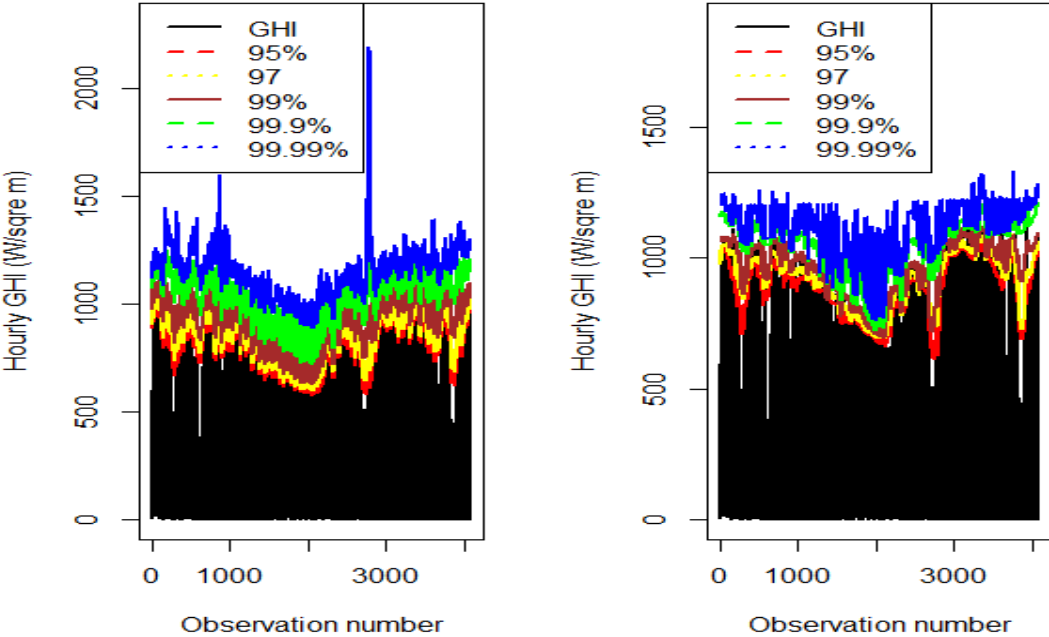


Figure 7. Plots of actual hourly GHI and forecasts of GHI at high quantiles using AVG (left-side) and MED (right-side) model.

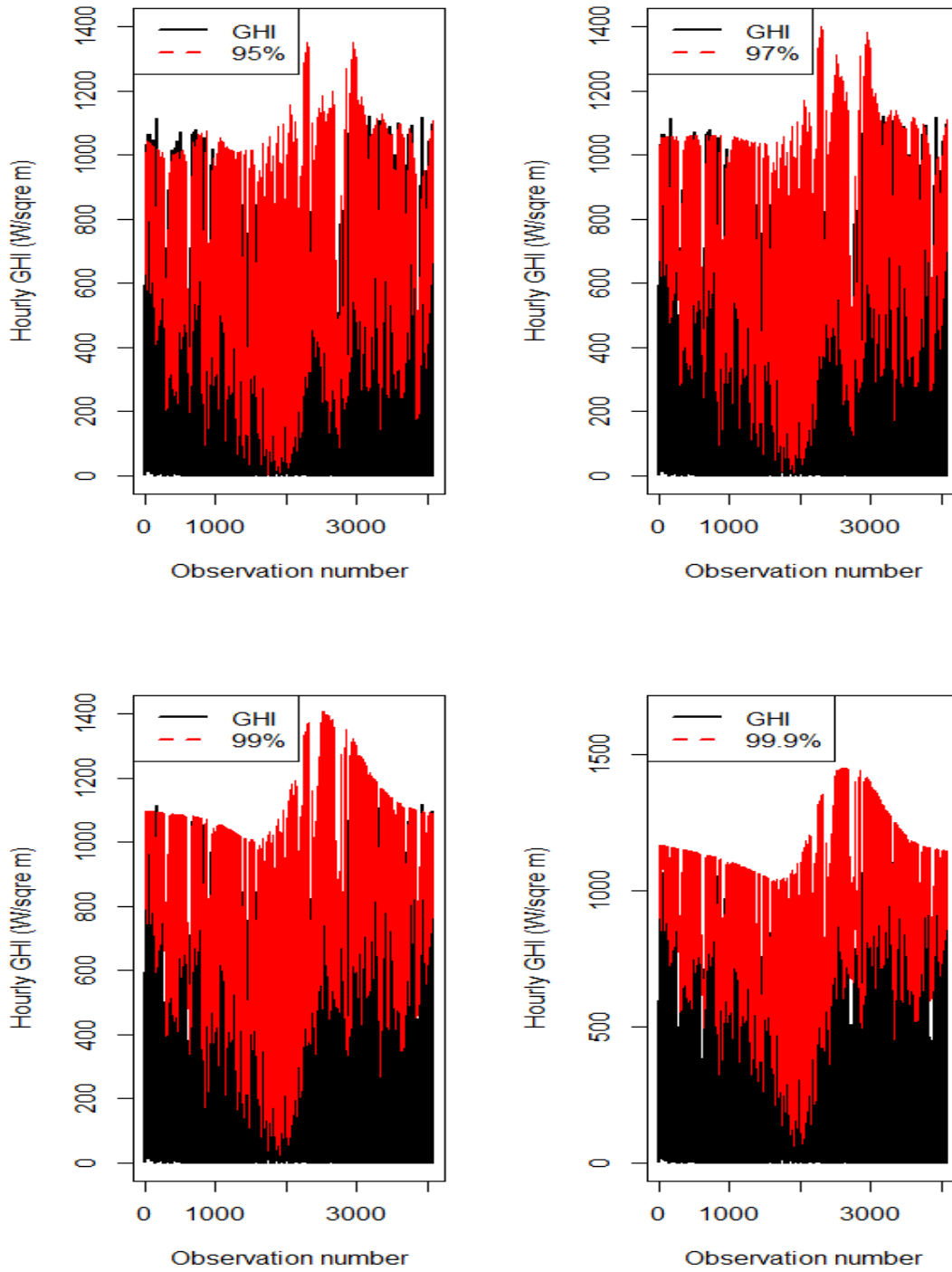


Figure 8. Top panel: Plots of actual hourly GHI and forecasts of GHI using AQR-2 model at $\tau = 0.95, 0.97, 0.99$ and 0.999 .

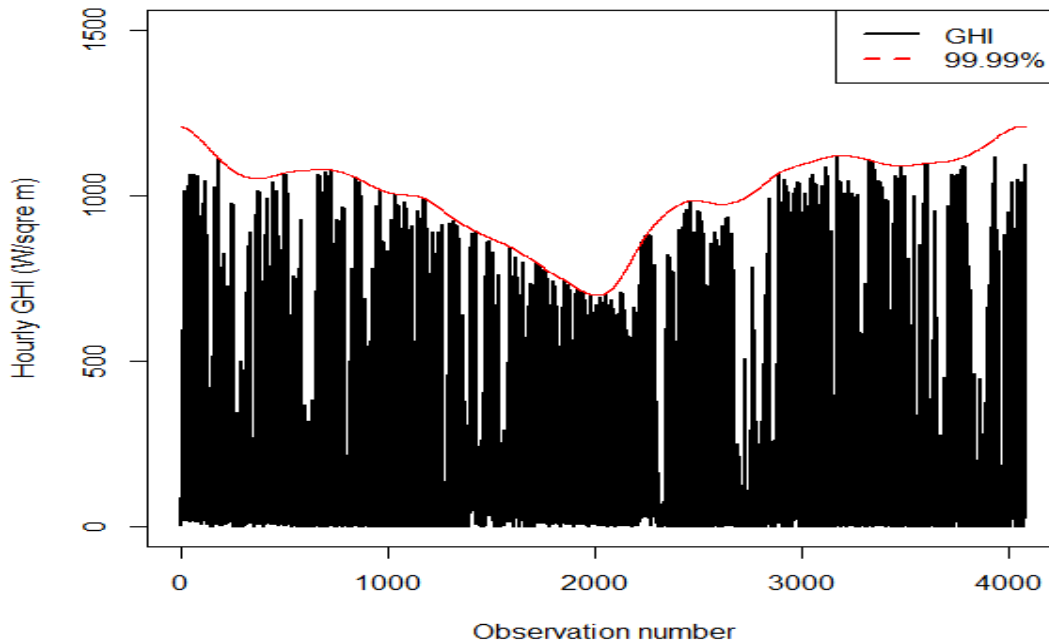


Figure 9. Plots of actual hourly GHI and forecasts of GHI at high quantiles using GAEV model at $\tau = 0.9999$

GHI produced by this study are useful by decision-makers and system operators in effectively balancing demand and supply of electricity that is environmentally friendly and also secures the future economic prosperity of the country.

4 Conclusion

This study carries out a comparative analysis of SPEM, GAEV, BM, AQR-1, AQR-2, AVG and MED models in the prediction of extremely high hourly GHI using data from the University of Venda radiometric station. The thrust of the study was on hourly GHI at extremely high quantiles i.e $\tau = 0.95, 0.97, 0.99, 0.999$ and 0.9999 . The AQR-2 was discovered to be the most accurate model for quantiles ($\tau = 0.95, 0.97, 0.99$ and 0.999) and the GAEV model was discovered to be the most accurate model for quantile ($\tau = 0.9999$). The best-fitting models were also used to forecast the GHI for each quantile. The findings of this study could be useful to decision-makers from power utilities in the optimal integration onto the electricity grid of the energy generated from solar power plants. In future research inclusion of additional covariates such as cloud cover, wind speed, seasons, among others can improve the predictive abilities of the models. Furthermore, it will also be more interesting to use the proposed models on the prediction of other renewable energy sources, such as wind power generation.

Acknowledgements

The authors acknowledge the DSI – CSIR Inter bursary Support (IBS) Programme for financial support and to both SAURAN, Southern African Universities Radiometric Network and the South African Weather Services for providing the data.

Nomenclature

AQR-1	Additive quantile regression
AQR-2	Additive quantile regression with temperature variable
AVG	Simple average method
h	Bandwidth parameter
b	Basis function
CRPS	Continuous rank probability score
DNI	Direct normal irradiance

FFNN	Feed-forward neural networks
GAEV	Generalised additive extreme value
GAM	Generalised additive model
GHI	Global horizontal irradiance (W/m^2)
H	Heuristic
I	Indicator function
L_i	Lower limit extreme quantile forecast
L_{AV}	Lower limit average
L_{Md}	Lower limit median
MED	Median
NWP	Numerical weather predictions
PV	Photovoltaic
PLF	Pinball loss function
QRA	Quantile regression averaging
QR-ALD	Quantile regression via asymmetric Laplace distribution
Q-Q	Quantile to quantile
S_λ	Penalty matrix
SARFIMA	Seasonal autoregressive fractionally integrated moving average
SAURAN	Southern African Universities Radiometric Network
SPEM	Semi-parametric extremal mixture
U_i	Upper limit extreme quantile forecast
U_{AV}	Upper limit average
U_{Md}	Upper limit median

Greek letters

β	Basis coefficient
ρ	Check function
ε	Error term
∞	Infinity
σ	Location
θ	Parameter
π	Penalised cubic regression smoothing spline
ℓ	Penalised log-likelihood function
τ	Quantile
λ	Smoothing function
μ	Threshold

Subscripts

t	time
-----	------

REFERENCES

- [1] Osório G.J., Lujano-Rojas J.M., Matias J.C.O., and Catalão J.P.S., "A probabilistic approach to solve the economic dispatch problem with intermittent renewable energy sources," *Energy*, vol. 82, pp. 949–959, 2015.
- [2] Gonçalves C., Cavalcante L., Brito M., Bessa R.J., and Gama J., "Forecasting conditional extreme quantiles for wind energy," *Electric Power Systems Research*, vol. 190, pp. 106636, 2014.
- [3] Mpfumali P., Sigauke C., Bere A., and Mulaudzi S., "Day ahead hourly global horizontal irradiance forecasting—Application to South African data," *Energies*, vol. 12, no. 18, pp. 3569, 2019.
- [4] Sun S., Wang S., Zhang G., and Zheng J., "A decomposition-clustering-ensemble learning approach for solar radiation forecasting," *Solar Energy*, vol. 163, pp. 189–199, 2018.
- [5] Antonanzas J., Osorio N., Escobar R., Urraca R., Martinez-de-Pison F.J., and Antonanzas-Torres F., "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

- [6] Zieher M., Lange M., and Focken U., “Variable Renewable Energy Forecasting–Integration into Electricity Grids and Markets–A Best Practice Guide,” German Federal Ministry for Economic Cooperation and Development, 2015.
- [7] Xie T., Zhang G., Liu H., Liu F., and Du P., “A hybrid forecasting method for solar output power based on variational mode decomposition, deep belief networks and auto-regressive moving average,” *Applied Sciences*, vol. 8, no. 10, pp. 1901, 2018.
- [8] Azimi R., Ghayekhloo M., and Ghofrani M., “A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting,” *Energy Conversion and Management*, vol. 118, pp. 331–344, 2016.
- [9] Ghayekhloo M., Ghofrani M., Menhaj M.B., and Azimi R., “A novel clustering approach for short-term solar radiation forecasting,” *Solar Energy*, vol. 122, pp. 1371–1383, 2015.
- [10] Diagne M., David M., Lauret P., Boland J., and Schmutz N., “Review of solar irradiance forecasting methods and a proposition for small-scale insular grids,” *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.
- [11] Kostylev V., and Pavlovski A., “Solar power forecasting performance–towards industry standards,” 1st international workshop on the integration of solar power into power systems, Aarhus, Denmark, 2011.
- [12] Lago J., De Brabandere K., De Ridder F., and De Schutter B., “Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data,” *Solar Energy*, vol. 173, pp. 566–577, 2018.
- [13] Voyant C., Notton G., Kalogirou S., Nivet M.L., Paoli C., Motte F., and Fouilloy A., “Machine learning methods for solar radiation forecasting: A review,” *Renewable Energy*, vol. 105, pp. 569–582, 2017.
- [14] Mutavhatsindi T., Sigauke C., and Mbuva R., “Forecasting Hourly Global Horizontal Solar Irradiance in South Africa Using Machine Learning Models,” *IEEE Access*, vol. 8, pp. 198872–198885, 2020.
- [15] Adeala A.A., Huan Z., and Enweremadu C.C., “Evaluation of global solar radiation using multiple weather parameters as predictors for South Africa provinces,” *Thermal Science*, vol. 91, no. suppl. 2, pp. 495–509, 2015.
- [16] Ranganai E., and Sigauke C., “Capturing long-range dependence and harmonic phenomena in 24-hour solar irradiance forecasting: A quantile regression robustification via forecasts combination approach,” *IEEE Access*, vol. 8, pp. 172204–172218, 2020.
- [17] Durrieu G., Grama I., Pham Q.K., and Tricot J.M., “Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles,” *Extremes*, vol. 18, no. 3, pp. 437–478, 2015.
- [18] Durrieu G., Grama I., Jaunatre K., Pham Q.K., and Tricot J.M., “*extremefit: An R Package for Extreme Quantiles*,” 2016.
- [19] Scarrott C., and MacDonald A., “A review of extreme value threshold estimation and uncertainty quantification,” *REVSTAT–Statistical Journal*, vol. 10, no. 1, pp. 33–60, 2012.
- [20] MacDonald A., Scarrott C.J., Lee D., Darlow B., Reale M., and Russell G., “A flexible extreme value mixture model,” *Computational Statistics & Data Analysis*, vol. 55, no. 6, pp. 2137–2157, 2011.
- [21] Sigauke C., and Bere A., “Modelling non-stationary time series using a peaks over threshold distribution with time varying covariates and threshold: An application to peak electricity demand,” *Energy*, vol. 119, pp. 152–166, 2017.
- [22] Fukutome S., Liniger M.A., and Süveges M., “Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland,” *Theoretical and Applied Climatology*, vol. 120, no. 3, pp. 403–416, 2015.
- [23] Youngman B.D., “*evgam: An R package for generalized additive extreme value models*,” arXiv preprint arXiv:2003.04067, 2020.
- [24] Youngman B.D., “Generalized additive models for exceedances of high thresholds with an application to return level estimation for US wind gusts,” *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1865–1879, 2019.
- [25] Gaillard P., Goude Y., and Nedellec R., “Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting,” *International Journal of forecasting*, vol. 32, no. 3, pp. 1038–1050, 2016.
- [26] Fasiolo M., Wood S.N., Zaffran M., Nedellec R., and Goude Y., “Fast calibrated additive quantile regression,” *Journal of the American Statistical Associations*, pp. 1–11, 2020.
- [27] Fasiolo M., Wood S.N., Zaffran M., Nedellec R., and Goude Y., “*qgam: Bayesian non-parametric quantile regression modelling in R*,” arXiv preprint arXiv:2007.03303, 2020.

- [28] Friederichs P., and Thorarinsdottir T.L., “Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction,” *Environmetrics*, vol. 23, no. 7, pp. 579–594, 2012.
- [29] Ghofrani M., and Suherli A., “Time series and renewable energy forecasting,” *Time series analysis and applications*, pp. 77–92, 2017.
- [30] Taillardat M., Fougères A.L., Naveau P., and de Fondeville R., “Extreme events evaluation using CRPS distributions,” *Solar Energy*, arXiv preprint arXiv:1905.04022, 2019.
- [31] Gaba A., Tsetlin I., and Winkler R.L., “Combining interval forecasts,” *Decision Analysis*, vol. 14, no. 1, pp. 1–20, 2017.
- [32] Wood S.N., “Generalized additive models: an introduction with R,” CRC press, 2017.