Creating Extreme Weather Time Series through a Quantile Regression Ensemble[☆]

Manuel Herrera[a,*], Alfonso P. Ramallo-González[b], Matthew Eames[c], Aida A. Ferreira[d], David A. Coley[a]

[a] *EDEn - Dept. of Architecture and Civil Eng., University of Bath, United Kingdom*
[b] *Faculty of Computer Science, University of Murcia, Spain*
[c] *CEE - School of Physics, University of Exeter, United Kingdom*
[d] *DASE - Federal Institute of Pernambuco, Brazil*

Abstract

Heat waves give rise to order of magnitude higher mortality rates than other weather-related natural disasters. Unfortunately both the severity and amplitude of heat waves are predicted to increase worldwide as a consequence of climate change. Hence, meteorological services have a growing need to identify such periods in order to set alerts, whilst researchers and industry need representative future heat waves to study risk. This paper introduces a new location-specific mortality risk focused definition of heat waves and a new mathematical framework for the creation of time series that represents them. It focuses on identifying periods when temperatures are high during the day and night, as this coincidence is strongly linked to mortality. The approach is tested using observed data from Brazil and the UK. Comparisons with previous methods demonstrate that this new approach represents a major advance that can be adopted worldwide by governments, researchers and industry.

Keywords: Quantile regression, models ensemble, weather files, built environment, heat waves

[☆] This research has been performed in the project COLBE on The creation of localized current and future weather for the built environment, EPSRC [grant code EP/K002724/1]. Ramallo-González would like to acknowledge Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia, CARM. All data created during this research are openly available from the University of Bath data archive at: https://doi.org/10.15125/BATH-XXXX.

[*] Corresponding author

Email address: amhf20@bath.ac.uk (Manuel Herrera)

1. Introduction

The hurricanes which hit the Caribbean in August-September 2017 caused around 300 deaths [1]. Floods, such that experienced by South India in 2015, gave rise to around 500 deaths [2], and mudslides, such the one that affected Sierra Leone in 2017, caused around 1,000 deaths [3]. In contrast, the 2003 European heat wave is believed to have killed more than 70,000 [4], the Russian heat wave of 2010, 55,000 [5] and the UK heat wave of 1976, 6,000 [6]. Taken together with the predictions of heat wave frequency and amplitude increasing as the climate warms [7, 8] implies that understanding the risk heat waves present to human populations is of critical interest [9, 10, 11].

In part simply because people spend most of their time inside buildings, but also because buildings can amplify external temperatures [11], the impact of heat waves on the conditions inside of buildings are of particular interest. Researchers and practising engineers calculate the likely temperatures within a pre-existing or proposed building, or its energy use, by simulation using a dynamic model, one input of which is a statement of the likely (reference) weather at the location, and contained in a computer file [12]. These reference years attempt to summarise, in just one year, weather variable relationships found in the past (usually in the last 20-30 years). There are two common ways to construct a reference weather year. The first is by finding the most typical (in some statistical way) January from the 20 or 30 years, then the most typical February, etc. and assembling these into a composite year. This is the method used to create the test reference year (TRY) used by researchers and industry to model typical conditions. The second is when studying the performance of the building under more extreme years, where months or years are ranked by mean temperature. This approach is used in the creation of the design summer year (DSY). Although this is useful in calculating annual cooling energy use during warmer years, it is unsatisfactory for studying mortality or morbidity as (i) it is based on mean monthly temperature, which might well not be excessive even during a month containing a heatwave; (ii) it tends to identify constantly warm periods, not peaks (i.e. heat waves) and, (iii) it is the combination of unusually high daytime and night time temperatures which kills [13].

Our new method is based on upper and lower quantiles (quantile regression), instead of the ordinary regression method which is highly conditioned by the mean [14, 15]. The advantage of this new method is that it can focus on finding patterns for both higher temperatures during the day and warmer temperatures during the night. This is of critical importance in characterizing a dangerous heatwave, as if temperatures remain high during the night, the building and occupants cannot cool down via night purging [11] and occupants will fail to maintain the biologically required core temperature without intense physical stress [16]. In extreme cases, this leads to a large peak in mortality, such as witnessed during the European heatwave of 2003, where almost 70,000 people died [17].

Quantile regression (QR) [18] is well suited to finding relationships among meteorological variables during time periods of *steadily* high temperatures; we propose a new methodology related to QR but targeted at shorter extremes, i.e. heat waves. The aim of the work presented here is to apply (quantile) regression to many years of summer data and to eventually combine them using a weighted average ensemble to produce a representative summer year that contains the events seen in the original data sets. The method is of equal utility when analysing historic, observed, data or future weather generated by meteorological models.

As Section 2 shows, QR encompasses an optimization process that is computationally intensive. Dealing with this is important to any attempt to use the method for a single weather station but critical for extending its use to high resolution climate models. For instance, a 5 km^2 grid resolution for the UK sufficient to resolve major variability in climate arising from topography and ocean coasts makes it necessary to compute more than 10,000 QR models to cover all of the UK. Each QR is associated with several meteorological variables and it is repeated for each of the available 20-40 years of hourly data. Consequently, the problem needs to be parallelized for running on a high performance computer.

In Section 3, the paper develops a double case-study with meteorological data collected from João Pessoa (Paraíba, Brazil) and London (UK) using 40 years of data. As these locations have two very different climates, we are able to test the validity of the approach for fundamentally different conditions. The result is a novel and tested weather file methodology, *the quantile regression ensemble summer year* (QRESY), representing extremely warm, potentially dangerous, temperatures. QRESY is then tested against the current weather files used to represent warmer conditions in Brazil and the UK.

2. Quantile regression ensemble summer year: QRESY

Extreme value theory (EVT) [19, 20] is an important topic in applied Statistics dealing with extreme deviations from the median of probability distributions [21]. Arising from EVT, several approaches can be applied to weather and climate extreme analysis [22, 23]. This work is focused on pattern extraction of extreme data but aims to tackle steady periods of high temperatures which are not necessarily associated with peak values, but which potentially cause heat waves. In comparison to EVT, quantile regression (QR) [18], and even a quantile regression ensemble, are much simpler approaches as they use many distributions that are non-standard and that are defined only through their quantile functions instead of the probability density functions. This means that QR models can deal with many complicated data structures that may not be dealt easily by conventional models [24]. In addition, QR does not need to meet the rigorous assumptions of the conditional-mean models i.e. normality in shape, and most importantly, homogeneity of variance. QR has the additional advantage of being able to be specifically tailored to analyse non-central locations (for example daily maximums and minimums), which are of great interest for extreme events analysis in both summer and winter. Previous works on QR applied to meteorology can be found at [25] where probabilistic precipitation forecasts have been made through quantile regression methods. Friederichs and Hense [26] applied QR to statistical downscaling for extreme precipitation events. Also, Lee et al. [27] successfully investigated QR models with extreme temperatures. Taillardat et al. [28] used quantile regression forests, a generalization of random forests for quantile regression, for short-term forecasting of temperature and wind speed. The aim of the work presented here is to apply regression to many years of summer data and to eventually combine them using a weighted average ensemble to produce a representative summer year that contains the events seen in the original data sets.

2.1. Quantile regression

As previously mentioned, extreme temperatures can cause serious overheating of humans especially where high maximum temperatures during the day coincide with high minima at night. QR provides a response to this concern by estimating the added effect of a set of weather inputs, x, on the quantiles of the temperature, y.

QR considers the relationship between the input variables and the output, by performing a conditional regression in a similar way to the conditional mean function used for OLS linear regression. Similarly to OLS, the conditional median function, $Q_q(y|x)$, would be applied; where the median is the 50^{th} percentile. The expression can be straightforwardly extended to any quantile, q, of the empirical distribution. The quantile $q \in (0, 1)$ for y splits the data into proportions q below and $1-q$ above: $F(y_q) = q$ and $y_q = F^{-1}(q)$. Following the parallelism, while OLS minimizes $\sum_i e_i^2$, QR minimizes a sum that gives asymmetric penalties: $(1 - q)|ei|$ for over-prediction and $q|ei|$ for under-prediction. The quantile regression estimator for quantile q minimizes the objective function of Equation (1).

$$Q(\beta_q) = \sum_{i:y_i \geq x_i'\beta}^{N} q|y_i - x_i'\beta_q| + \sum_{i:y_i \leq x_i'\beta}^{N} (1-q)|y_i - x_i'\beta_q| \tag{1}$$

Equation (1) is based on non-differentiable functions and requires linear programming methods for its minimization. Commonly used approaches, such as the Simplex method for moderately sized data sets or the Interior Point method for larger databases, guarantee to yield a solution in a finite number of iterations. Bootstrap methods are often preferable because they make no assumption about the distribution of the response [29] hence able to generalize QR for any residual distribution. Then, bootstrap standard errors are often used instead of analytic standard errors.

2.2. Quantile regression ensemble

Ensemble learning is a process that uses a set of models to study a common problem. The set of models is integrated to obtain a more robust and accurate approach for temperature predictions, in addition to helping to maintain a suitable uncertainty level [30]. The final model is generated from combining the single models or by selecting the best models in terms of accuracy [31]. For regression problems, ensemble integration is done using a linear combination of the predictions. For QR, this is given by Equation (2),

$$Q_{Tq}(y|x) = \sum_{i=1}^{K} h_{q,i}(y|x) \cdot Q_i(y|x) \tag{2}$$

where K is the number of single QRs making up the ensemble, q represents a specified quantile for QR, and $h_{q,i}(y|x)$ are weighted functions; $i = 1, \cdots, K$. In this case, the interest is focused on ensemble regressions with weights proportional to the distance of single QRs to the QR for the median, Q_{50}. Thus, $h_{q,i}(y|x)$ is given by the expression of Equation (3),

$$h_{q,i}(y|x) = \begin{cases} \frac{d_{q,i}}{\sum d_{q,i}} & \text{for } q \in \text{upper QR set} \\ \frac{d_{q,i}^{-1}}{\sum d_{q,i}^{-1}} & \text{for } q \in \text{lower QR set,} \end{cases} \tag{3}$$

where $d_{q,i} = |Q_{q,i}(y|x) - Q_{50,i}(y|x)|$.

2.3. QRESY methodology

The aim is to create a weather time series by combining observed or simulated summer extreme temperatures. This is done by endowing higher weights to quantiles away from Q_{50} for ensembles within upper quantiles. At the same time, it increases the importance of quantiles near to Q_{50} for combining lower quantiles. The idea being to focus on explaining critical phases of summer temperatures. Each ensemble is thereby made over the predictors of a number of regression models corresponding to each of the years in the database. The ensemble parameters can be tuned by cross-validation over random partitions of the data into *training* and *test* summer periods. This QR ensemble is key to the final aim of a weather file representing extreme summer temperatures - the so-called quantile regression ensemble summer year (QRESY). Its construction follows the flowchart of Figure 1.

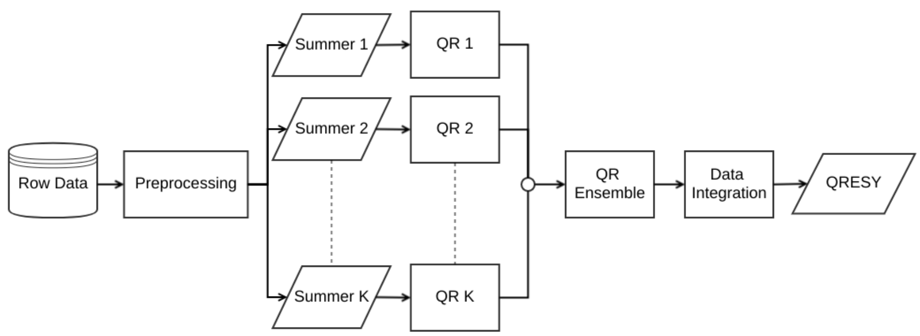


Figure 1: Phases for creating the new QRESY weather file

The QRESY creation process starts by collecting hourly weather data over a long period. Typically, weather files attempt to be representative of periods around 20-40 years (Appendix A), and here we do the same, however much longer periods could be used. The existence, variables and quality of hourly weather data varies depending on the location. The variables usually include temperature, atmospheric pressure, cloud cover, wind speed and wind direction, precipitation, etc. For the QRESY process, preprocessing of this data is required to ensure it contains no long sequences of missing data. If large amount of data are missing in any of the variables, the whole year is removed from the analysis. At this point it is also necessary to decide the target level of extreme weather to work with. That is, to fix the quantile level for the subsequent construction of the QR models depending on the distance to Q_{50}. Running a QR model for every year under analysis is an "embarrassingly parallel" problem, as it is straightforward to separate the problem into a number of parallel tasks and the code run on a parallel machine. The set of regressions is combined in a unique year of hourly data as it is above described.

The last phase in creating a QRESY involves a data integration process (Figure 1). This is necessary as all the variables required to create a weather file are often not included in the regression ensemble. This can be due to using data from different sources and/or at various resolution levels. Or because some data that was not directly measured for the whole range of analysed years needs to be estimated from the values of other variables that were measured. For example, solar radiation, which can be derived from variables such as temperature and sky cloudiness. Finally there is the need to ensure the QRESY contains all the variables required to be of use in an agricultural, building or other model. For example an industry standard building simulation package [12], and be in the required format.

3. QRESY for building performance simulation

This section applies the method to the UK and Brazil.

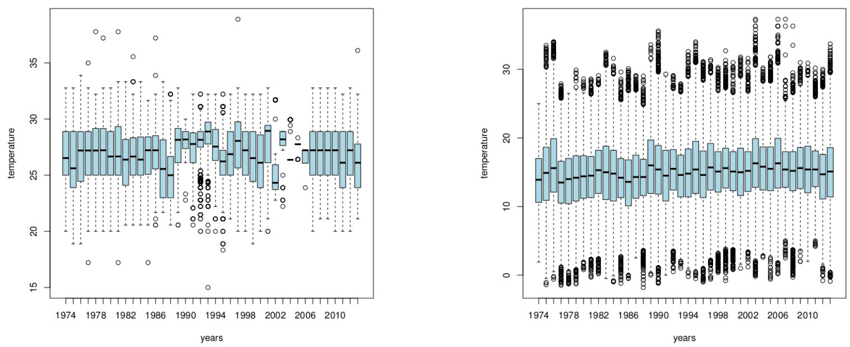
3.1. Data

The meteorological data used was collected every hour in João Pessoa (Paraíba, Brazil) and London (UK), Table 1, over 40 years (1974 - 2013). Whilst João Pessoa has a Tropical Climate, type A [32]; London has an Oceanic climate, type C. Having two distinct climates allows us to test the consistency of any findings.

October to March (of the following year) were used to represent the summer period in João Pessoa and April to September for London. The databases used are available upon request at the NOAA's National Centre for Environmental Information (NCEI) and at the British Atmospheric Data Centre (BADC) [33],

Table 1: Coordinates of the 2 cities selected, with latitude, longitude and elevation of weather stations

City	Country	Weather station	Latitude	Longitude	Elevation	Years
London	UK	Heathrow	51.479	-0.449	25	1974 - 2013
João Pessoa	Brazil	João Pessoa	-7.155	-34.792	40	1974 - 2013



(a) Hourly summer temperatures grouped by years. João Pessoa, 1974 - 2013.

(b) Hourly summer temperatures grouped by years. London, 1974 - 2013.

Figure 2: Comparison of temperatures over the 40 years under study

respectively. The data available is: wind direction (wdir) in North azimuth degrees, wind speed (wspeed) in knots, cloud cover (cloud) in scale 1-8 from minimum to maximum cloud cover, air pressure (airp) in millibars, air temperature (airt) in degrees Celsius, and dew point temperature (dpt) in degrees Celsius. The hourly air temperature distribution for each summer from 1974 to 2013 is summarised in Figure 2. The figure uses a box-plot representation for every year. The years when observations are predominantly missing (as it is the case of the years 2003 - 2005 for João Pessoa) have been removed.

3.2. Reference weather files

Two well accepted, reference, weather files for the same cities were used as comparison weather files to study the benefits of the QRESY approach.

In the case of João Pessoa, the standard IWEC weather file (Appendix A) was the starting point. This represents typical weather conditions, rather than a particularly warm summer, and was used as no warm summer type file has been produced for the location - a common problem in much of the world. In order to maintain homogeneity in the comparisons, a warm summer type file termed a probabilistic design summer year, or PDSY (Appendix A) was also created using the data from João Pessoa and the method described in [34]. PDSY replaced design summer years (DSYs) attempting to better describe overheating events, their relative severity and their expected frequency [35]. The basis years 1974 - 2013 were used. Under this method 1991 was found to be the year best representing a moderately warm year, while 2005 contained an intense hot event and 2010 a long period of extreme summer temperatures. The creation of a PDSY for Brazil can be considered a new result in itself. It is worth mentioning that the variability for the associated distribution found when creating the João Pessoa PDSY is smaller than that of the London PDSY.

In the case of London, the Test Reference Year, or TRY, was used [41]. The TRY is similar in construction to the TMY (Appendix A) used in many parts of the world. For London 2013 is the current PDSY representing a moderately warm year, 1976 a long period of extreme temperatures, and 2003 an extreme hot event for a short period.

3.3. Quantile methods for weather files

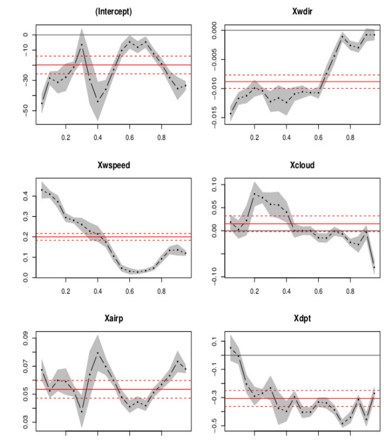
We can check the performance of the QR approach by evaluating the statistical characteristics of the summers of 1991/92 in João Pessoa and 2013 in London - these years being selected because they are the warm summer selected by PDSY. Table 2 shows the difference between the coefficients for OLS regression and QR for quantiles 0.05, 0.5, 0.95; see Equation (1). The coefficients in Table 2 represent the difference in the predicted temperature for a one-unit difference in the input associated with the coefficient, if the rest of inputs remain constant. This means that if the input varies by one unit, and the rest of the inputs do not vary, the temperature will differ on average by the quantity given by the associated coefficient. In short, we refer to the added effect represented by each coefficient as an "impact" of the input variable on the temperature.

Table 2: Coefficients of OLS and QR at 0.05, 05, and 0.95 quantiles. João Pessoa - summer 1991 and London, summer - 2013. The coefficients are marked by '*' when they are significantly different from zero.

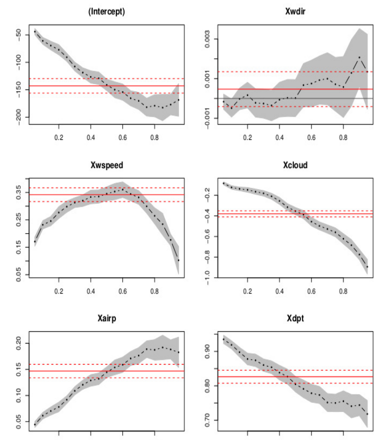
Weather station	Input	OLS	QR 0.05	QR 0.50	QR 0.95
João Pessoa	wdir	-0.01	-0.01	-0.01	-0.00
	wspeed	0.20*	0.43*	0.11*	0.12*
	cloud	0.02	0.00	-0.30*	-0.08*
	airp	0.05*	0.07*	0.06*	0.07*
	dpt	-0.31*	0.05	-0.40*	-0.27*
London - Heathrow	wdir	0.00	0.00	0.00	0.00
	wspeed	0.34*	0.17*	0.34*	0.10*
	cloud	-0.38*	-0.08	-0.35*	-0.89*
	airp	0.15*	0.04*	0.14*	0.18*
	dpt	0.83*	0.93*	0.80*	0.72*

Figure 3 shows prediction intervals for every quantile for each explanatory variable. The X-axis gives the quantile and the Y-axis the estimated value for the regression coefficients. The red lines are the result of the OLS regression based on the mean. We can see how the relationships in the weather database significantly change with the quantile, as there is no intersection between the OLS confidence interval (marked by a red dot-line) and the QR prediction intervals. This suggests that those added effects for each regression coefficient might not be constant across the conditional distribution of the temperature, as OLS supposes.

Wind direction does not seem to have a contribution to temperatures, with a significant difference regarding OLS in most of the quantiles in London. Nor in João Pessoa, where the impact of wind direction is close to zero. The parameter values of Equation (1) for cloud cover and dew point temperature decrease when the QR quantile value increases. Checking the behaviour of these parameters in Figure 3, we find that warmer temperatures in London 2013 are associated



(a) QR for João Pessoa summer 1991/92



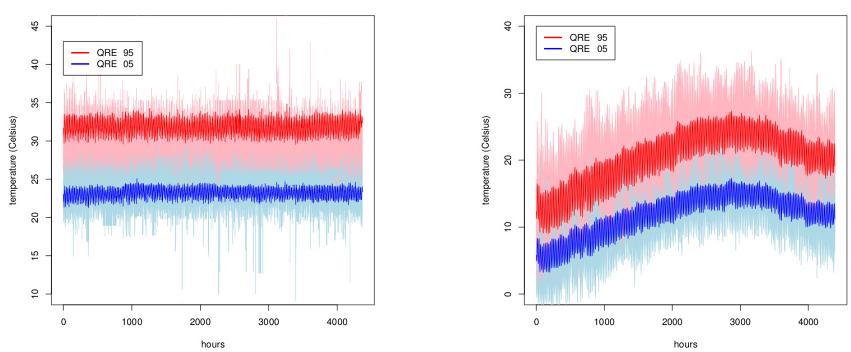
(b) QR for London summer 2013

Figure 3: Impact of QR explanatory variables depending on quantile. Hourly summer temperatures.

with lower than usual humidity and also with clearer skies. Also for London, the air pressure coefficients increase when QR has the highest values and the wind speed drops.

In the case of João Pessoa, the relative impact of cloud cover is not as straightforward as it is for London. The rest of the variables have an impact on the temperature values similar to the results obtained through OLS. This is because the confidence intervals for the QR regression contain the OLS results (marked in red in Figure 3a) for most of the quantiles. These relationships can be explained by the low variability in the temperature values. As Figure 2 suggested, the interquartile range for temperatures in Brazil varies around 5^{o}C while it is 10^{o}C for the UK. Still, Figure 3 shows how the impact of the variables changes from that predicted by just using the mean. The role of wind speed is clear - as this decreases as the temperature quantiles increase. This is a critical finding, as knowledge of the wind speed is fundamental for the correct analysis of overheating in naturally ventilated buildings, as wind generated ventilation is an essential cooling strategy. Worryingly, this observed reduction in wind speed during warmer periods is not currently included in other methods of creating warm years.

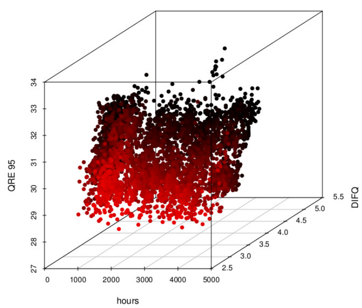
Figure 4 shows the temperature which results from the prediction ensemble of quantile regressions for Q95 and Q05. The individual predictions from it also make the ensemble represented in Figure 4, shown using light red and blue. Firstly, it highlights the difference in how temperatures evolve during the summer in the two locations. While João Pessoa (Figure 4a) presents a steady trend in the quantile regression ensemble (QRE) for Q95 and Q05 (red and blue colours in Figure 4, respectively), London (Figure 4b) has a growing trend until usually reaching maxima temperatures in July which then start to drop in August. Another important difference between the 2 locations is in the variability shown by the individual 40 quantile regressions for each ensemble. These individual predictors are represented in light red and blue for each ensemble level, Q95 and Q05 respectively.



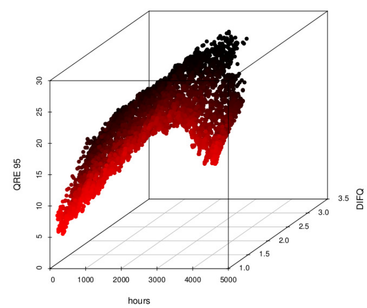
(a) QR ensemble in João Pessoa (b) QR ensemble in London

Figure 4: Ensembles of QRs at 0.95 and 0.05 quantiles for the summer temperatures of João Pessoa and London for the summers 1974 - 2013

Figure 5 shows the distribution in time of the result of QRE for the 0.95 quantile and the difference of the ensembles at 0.05 and 0.50 levels. This provides a visualization of potential periods of the year in which a heat wave could happen. These are periods of higher values of QRE at 0.95 and smaller difference between a lower quantile and the median that can lead to have warmer nights in hotter days (DIFQ = QRE-50 - QRE-05, in this example). The plot corresponding to João Pessoa (Figure 5a) shows a quite homogeneous distribution as it is expected from the previous analysis. In London (Figure 5b), it is possible to better distinguish a section of the 3D scatter plot that is on the front of the cube (warmer nights) and also has high values associated with QRE-95 on the Z-axis (hot days).



(a) QRE analysis for João Pessoa



(b) QRE analysis for London

Figure 5: Visual analysis of quantile regression ensemble as a tool for detecting warmer nights (low values on Y-axis) in also hot days (high values on Z-axis)

3.4. Comparison with existing reference years

Here a QRESY is compared to the corresponding standard typical and warm summer years all assembled from the same observed data sets, both in terms of warm spells and the metrics used to size building systems.

Plant sizing requires identifying near extreme external temperatures (the design temperatures) and then sizing the plant to combat these temperatures. The 99 % and the 99.6 % temperatures are normally used as design temperatures for heating and the 0.4, 1 and 2 percentile for cooling, i.e. the near minimum and maximum temperatures extracted from the cumulative distribution. The results are shown in Figures 6 and 7.

Figure 6 shows that for London the PDSY and the QRESY weather files give almost the same design temperature for heating, but the TRY a rather different

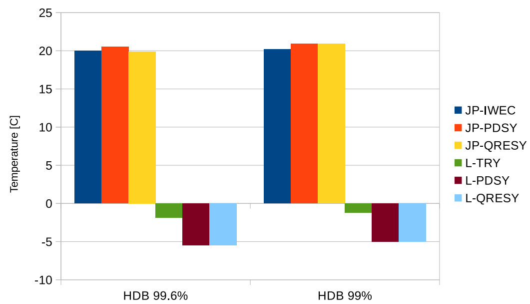


Figure 6: Heating dry bulb design temperature for the 99.6 and 99 % range

value. This suggests that the PDSY and the QRESY are consistent with each other, which is encouraging. However as these are both designed as extreme summer years, it is unlikely that they would be used for heating plant sizing. For João Pessoa this difference is less marked for, however the temperatures are all so high that a heating system is unlikely to be needed.

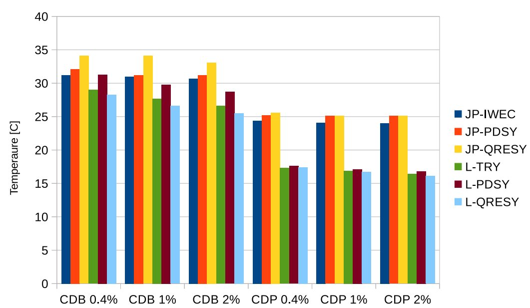


Figure 7: Cooling dry bulb temperature and dew point temperatures for the 0.4 %, 1 % and 2 % thresholds

Figure 7 shows that the size of cooling system selected for a building depends greatly on which weather file type is used. This is most noticeable for João Pessoa. Here QRESY is the most conservative, and suggests a design temperature clearly above all the others for the three different thresholds. This is would lead to the specifying of more powerful cooling systems, which is as one might expect, capable of maintain suitable indoor temperatures even during the heatwave events contained in the QRESY.

When sizing cooling plant, the dew point temperature is also important. This is the temperature at which the air will no longer be able hold all the moisture it contains and condensation will occur. The difference between the results given by the methodologies is less noticeable. However, the difference between the London TRY and the two others still clearly stands out. This further validates QRESY as a summer weather file, as it presents similar results to the PDSY. In addition, the difference between the QRESY and the TRY is in the same direction and magnitude as the difference between the PDSY and the TRY.

With respect to the timing of the peak temperature (Figure 8). We can see that the extremes appear at different dates dependent on the methodology used to assemble the weather file, except for the lowest extreme for London. In this case, there is a coincidence between PDSY and QRESY.

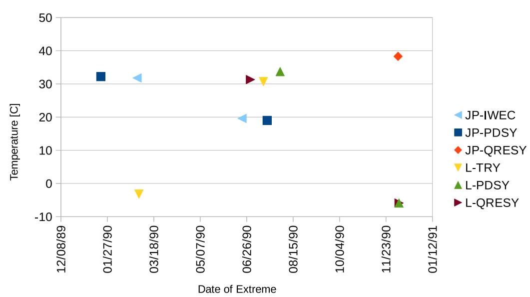


Figure 8: Value and location of extremes. The missing orange diamond is under the blue square close to the 20^{o}C line

In the above we have discussed differences in the peak severity of the weather given by the three methodologies. Heating Degree Days (HDD) and the Cooling Degree Days (CDD) are well known measures of weather severity over the whole winter or summer period. HDD is given by the sum of the product of the time and temperature below a fixed temperature (which represents the external temperature when heating might be commonly used, in this case 15.5^{o}C) for the whole year. CDD makes the same calculation for the time and temperature above a fixed temperature (i.e. when cooling might be used, in this case 18.3^{o}C).

Figure 9 shows the HDD and the CDD for the two locations and the three types of weather files.

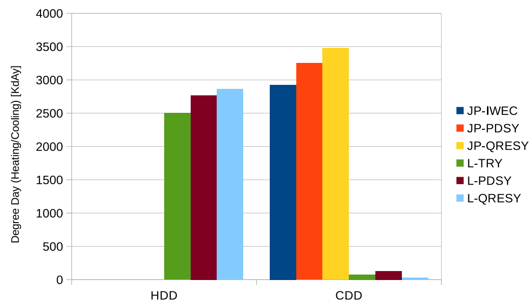


Figure 9: Heating and Cooling Degree Days

Figure 9 shows how each one of the weather files types present different values of HDD and CDD. We can see how the heating and cooling loads are always consistently higher with PDSY and even more with QRESY. This implies that these weather files are more extreme, showing on average hotter summers and colder winters. PDSY and QRESY are therefore weather files that will tend to suggest larger energy consumption in the buildings being modelled.

4. Conclusions

The techniques presented in this paper are based on QR estimates of rates of change for functions along or near the upper or lower boundary of the conditional distribution of temperatures. In disciplines such as building modelling where knowledge of severe events is critical it seems a powerful approach. We found QR models to be useful in understanding the rate of changes in extreme events along with the ability of the method to identify the links between weather variables at extreme temperatures and preserve these coincidences in the final assembled weather file. For example, wind speed which takes its lowest values during high temperature periods. This increases the human-relevant risks of a heat wave in naturally ventilated buildings, as the cooling strategy relies on bringing in fresh night-time air into the building. QR also accounts for the relationship between high temperatures and cloud cover, as cloud cover varies differently when temperatures are far from the average. This is of critical importance as cloud cover has a direct effect on how much solar radiation reaches the Earth's surface, or enters buildings.

In this work, the use of an ensemble of QR predictions based on the distance of the estimated model values from the median is proposed. This ensemble provides a better representation of when the highest daytime temperatures coincide with warmer nights. As the literature points to the critical need for a definition of heat waves that heavily emphasizes nocturnal temperatures (due to the risk to human health such events represent) this is clearly important, and will allow researchers to study when excess mortality is likely to peak. A better real-time early warning system for heatwave events can be created in this way.

Our methods allow meteorological agencies to create a quantile regression ensemble summer year (QRESY) for any location. The QRESY provides a far better representation of extremes than the common reference weather files, particularly the coincidences between weather variables. The work has a strong focus on ensuring that the extreme events are represented. However, the evaluation of HDD and CDD that has been performed demonstrates that this new way of synthesizing weather information is also valid for evaluating year-long metrics such as heating and cooling demands. This makes the newly introduced QRESY as good as the PDSY for all traditional building simulation applications and superior for the evaluation of extreme events.

The QR approach is far more mathematical complex than current methods of assembling reference years. It is also computational intensive. However the methodology we have developed makes the problem tractable and demonstrates a highly effective way to create QRESYs for widely different climates worldwide. In comparison with other standard weather files, QRESY is more statistically robust in dealing with missing information or outliers in the database, because quantiles are based on order statistics and therefore not directly affected by the exact values of the raw data.

The robustness increases even more if QR is computed by a bootstrapping process. The final ensemble phase for creating QRESY also ensures the production of reliable estimates. The methodology we have developed is computationally efficient, therefore making the problem tractable even if the desire is to create extreme weather years with high spacial resolution, for example, every 5km across the land surface. The approach can also be mirrored to deal with minimum winter temperatures.

Using the QR ensemble for analysing summer temperatures opens up a new paradigm for the real-time identification of heat waves and the creation of reference climate files that focus on the impact of extreme temperatures on human morbidity and mortality.

References

[1] M. Lichtveld, Disasters through the lens of disparities: elevate community resilience as an essential public health service, American Journal of Public Health 108 (1) (2018) 28-30.

[2] N. Pandey, S. Natarajan, How social media can contribute during disaster events? Case study of Chennai floods 2015, in: Proceedings of Advances in Computing, Communications and Informatics (ICACCI) IEEE, 2016, pp. 1352–1356.

[3] D. Harris, A. Wurie, F. Baingana, S. Sevalie, F. Beynon, Mental health nurses and disaster response in Sierra Leone, The Lancet Global Health 6 (2) (2018) e146-e147.

[4] J. M. Robine, S. L. Cheung, S. Le Roy, H. van Oyen, C. Griffiths, J. P. Michel, F. R. Herrmann, Death toll exceeded 70,000 in Europe during the summer of 2003, Comptes Rendus Biologies 331 (2) (2008) 171–178.

[5] F. E. L. Otto, N. Massey, G. J. van Oldenborgh, R. G. Jones, M. R. Allen, Reconciling two approaches to attribution of the 2010 Russian heat wave, Geophysical Research Letters 39 (4) (2012) L04702.

[6] R. S. Kovats, S. Hajat, P. Wilkinson, Contrasting patterns of mortality and hospital admissions during hot weather and heat waves in Greater London UK, Occupational and Environmental Medicine 61 (11) (2004) 893–898.

[7] M. Beniston, D. B. Stephenson, Extreme climatic events and their evolution under changing climatic conditions, Global and Planetary Change 44 (1) (2004) 1–9.

[8] IPCC, Climate Change 2014: Mitigation of Climate Change, Vol. 3, Cambridge University Press, 2015.

[9] R. S. Teegavarapu, Modeling climate change uncertainties in water resources management models, Environmental Modelling & Software 25 (10) (2010) 1261–1265.

[10] V. Mishra, 2.12 - food security implications of climate variability and climate change, in: R. A. Pielke (Ed.), Climate Vulnerability, Academic Press, 2013, pp. 117–128.

[11] D. Coley, T. Kershaw, Changes in internal temperatures within the built environment as a response to a changing climate, Building and Environment 45 (1) (2010) 89–93.

[12] M. Herrera, S. Natarajan, D. Coley, T. Kershaw, A.P. Ramallo-González, M. Eames, D. Fosas, M. Wood, A review of current and future weather data for building simulation, Building Services Engineering Research and Technology 38 (5) (2017) 602–627.

[13] J.K. Vanos, L.S. Kalkstein, T.J.Sanford, Detecting synoptic warming trends across the US Midwest and implications to human health and heatrelated mortality, International Journal of Climatology, 35 (1) (2015) 85–96.

[14] M. Herrera, M. Eames, A. Ramallo-González, C. Liu, D. Coley, Quantile regression ensemble for summer temperatures time series and its impact

on built environment studies, in: Proceedings of the 8th International Environmental Modelling and Software Society (IEMSS), Vol. 3, 2016, pp. 602–609.

[15] M. Herrera, A. A. Ferreira, D. A. Coley, R. R. B. de Aquino, SAX-Quantile based multiresolution approach for finding heatwave events in summer temperature time series, AI Communications 29 (6) (2016) 725–732.

[16] P. de Wilde, D. Coley, The implications of a changing climate for buildings, Building and Environment 55 (2012) 1–7.

[17] R. S. Kovats, L. E. Kristie, Heatwaves and public health in Europe, The European Journal of Public Health 16 (6) (2006) 592–599.

[18] R. Koenker, G. Bassett Jr, Regression quantiles, Econometrica: journal of the Econometric Society (1978) 33–50.

[19] S. Coles, J. Bawa, L. Trenner, P. Dorazio, An introduction to statistical modeling of extreme values, Vol. 208, Springer, 2001.

[20] S. Sippel, D. Mitchell, M. T. Black, A. J. Dittus, L. Harrington, N. Schaller, F. E. Otto, Combining large model ensembles with extreme value statistics to improve attribution statements of rare events, Weather and Climate Extremes 9 (2015) 25–35.

[21] M. Ghil, P. Yiou, S. Hallegatte, B. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, et al., Extreme events: dynamics, statistics and prediction, Nonlinear Processes in Geophysics 18 (3) (2011) 295–350.

[22] D. Cooley, Extreme value analysis and the study of climate change, Climatic Change 97 (1) (2009) 77–83.

[23] R. Mínguez, F. Méndez, C. Izaguirre, M. Menéndez, I. J. Losada, Pseudo-optimal parameter selection of non-stationary generalized extreme value

models for environmental variables, Environmental Modelling & Software 25 (12) (2010) 1592–1607.

[24] Y. Cai, D. E. Reeve, Extreme value prediction via a quantile function model, Coastal Engineering 77 (2013) 91–98.

[25] J. Bjørnar Bremnes, Probabilistic forecasts of precipitation in terms of quantiles using nwp model output, Monthly Weather Review 132 (1) (2004) 338–347.

[26] P. Friederichs, A. Hense, Statistical downscaling of extreme precipitation events using censored quantile regression, Monthly Weather Review 135 (6) (2007) 2365–2378.

[27] K. Lee, H.-J. Baek, C. Cho, Analysis of changes in extreme temperatures using quantile regression, Asia-Pacific Journal of Atmospheric Sciences 49 (3) (2013) 313–323.

[28] M. Taillardat, O. Mestre, M. Zamo, P. Naveau, Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, Monthly Weather Review 144 (2016) 2375–2393.

[29] M. Kocherginsky, X. He, Y. Mu, Practical confidence intervals for regression quantiles, Journal of Computational and Graphical Statistics 14 (1) (2012) 41–55.

[30] J. Mendes-Moreira, C. Soares, A. M. Jorge, J. F. D. Sousa, Ensemble approaches for regression: A survey, ACM Computing Surveys (CSUR) 45 (1) (2012) 10.

[31] L. I. Kuncheva, Switching between selection and fusion in combining classifiers: an experiment, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 32 (2) (2002) 146–156.

[32] W. Köppen, Versuch einer klassifikation der klimate, vorzugsweise nach ihren beziehungen zur pflanzenwelt, Geographische Zeitschrift 6 (11. H) (1900) 593–611.

[33] UK Meteorological Office, Met Office Integrated Data Archive System (Midas) land and marine surface stations data (1853-current), NCAS British Atmospheric Data Centre (accessed 01/10/16).

[34] CIBSE, Design Summer Years for London – CIBSE TM49 (2014).

[35] M. Eames, An update of the UK's design summer years: Probabilistic design summer years for enhanced overheating risk analysis in building design, Building Services Engineering Research and Technology 37 (5) (2016) 503–522.

[36] D. B. Crawley, Creating weather files for climate change and urbanization impacts analysis, in: Building Simulation, IBPSA Beijing, 2007, pp. 1075–1082.

[37] J. M. Finkelstein, R. E. Schafer, Improved goodness-of-fit tests, Biometrika 58 (3) (1971) 641–645.

[38] S. Wilcox, W. Marion, Users manual for TMY3 data sets, National Renewable Energy Laboratory Golden, CO, 2008.

[39] Y. J. Huang, Development of 3012 IWEC2 weather files for international locations (rp-1477), ASHRAE Transactions 120 (2014) 340.

[40] CIBSE, Weather, solar and illuminance data, Chartered Institution of Building Services Engineers.

[41] G. Levermore, J. Parkinson, Analyses and algorithms for new Test Reference Years and Design Summer Years for the UK, Building Services Engineering Research and Technology 27 (4) (2006) 311–325.

[42] M. F. Jentsch, G. J. Levermore, J. B. Parkinson, M. E. Eames, Limitations of the CIBSE design summer year approach for delivering representative near-extreme summer weather conditions, Building Services Engineering Research and Technology 35 (2) (2014) 155–169.

[43] J. F. Nicol, J. Hacker, B. Spires, H. Davies, Suggestion for new approach to overheating diagnostics, Building Research & Information 37 (4) (2009) 348–357.

Appendix A. Brief introduction to weather files

This Appendix introduces the concept of weather files, how they are used for decision-making in building design and operation, and shows their limitations at handling extreme weather conditions.

Appendix A.1. Files for typical weather conditions

Buildings should provide occupants with comfort and safe shelter from the external environment. In addition, improvement in the energy efficiency of buildings is of importance to building designers in order to mitigate climate change. To approximate the prevailing weather found at a location, building simulation programs use a weather file of one year of typical hourly data based on the weather observed over 20 or more (basis) years. The typical year is usually created by selecting 12 individual months, each one being the most representative month over the basis years; that is, the January which is closest to the average of all Januaries, etc. [36]. Closest being defined by considering distances to each monthly average of several weather variables using the Finkelstein-Schafer statistic [37]. This is the case for both the typical meteorological year (TMY) with its latest version called TMY3 [38] or the International Weather for Energy Calculation (IWEC) weather file which is associated with the IWEC2 dataset [12, 39].

Appendix A.2. Files for extreme weather conditions

Weather files based on typical years, as described above, are not well suited to represent extreme conditions as they directly work with average data. This is the reason why it is necessary to count on alternatives that allow one to consider for example hot temperatures in summer. One of the alternatives for approaching these extreme weather files is the called Design Summer Year (DSY) proposed by the Chartered Institution of Building Services Engineers (CIBSE, UK) [40]. The DSY is created by selecting an entire year which contains the third hottest mean summer dry bulb temperature [41] (this is April to September for the northern hemisphere and October to March for the southern hemisphere) from observed data (this is 1984-2013 in the case of the UK).

The DSY has proven poor when analysing overheating in buildings [42]. The DSY has since evolved into the so-called probabilistic DSY (PDSY) [34] which is currently used in the UK as reference of warm summers. The PDSY is based upon a different overheating metric to the DSY and uses the number of hours in which the temperature is above a certain threshold when a building is occupied. This basis has been used to propose a new metric, weighted cooling degree hours (WCDH), which takes into account indoor temperature in buildings and also the length of higher temperature sequences. This metric provides a better and more realistic characterization than the DSY regarding extreme temperatures [43]. The methodology for selecting PDSY is based on the ascending order of WCDH [34] using data from 1984 to 2013. PDSY has now been updated taking into account resampling and EVT methods [35].

Appendix A.3. Limitations in the creation of weather files

As just mentioned, weather files used for building simulation represent a compendium of weather data drawn from a number of basis years [36]. As the comparisons are monthly, the weather file is subsequently constrained to consider monthly averages, minima, and maxima. Thus ignoring several scenarios in which heat waves occur in a month of otherwise mild temperature. In addition, given the way in which the months are selected it is possible that the DSY represents a slightly warmer than average summer, but contains no heat waves [42] and thus no weather that might cause excess mortality. Given such limitations, there is an urgent need for new methods to support building performance simulation. Quantile regression allows the modelling of hot temperatures and also considers potential relationships with other meteorological variables.

Thereby it promises the creation of realistic extreme years.