Extreme-Value Analysis (EVA 2019) Conference: Data Mining Competition

Raphaël Huser (raphael.huser@kaust.edu.sa), KAUST, Saudi Arabia

1 Motivation

Global warming is affecting the Earth climate year by year, the biggest difference being observable in increasing temperatures in the World Ocean. In particular, coral reefs are increasingly threatened worldwide as they are sensitive to modest increases in background seawater temperature (Cantin *et al.*, 2010). Studies have shown that persistent high sea temperatures can result in substantial coral bleaching and some cases coral mortality; see, e.g., McClanahan *et al.* (2007).

In this data mining competition, the goal is to analyse and predict the joint tail behavior of extreme sea surface temperature (SST) anomalies for the whole Red Sea, a warm semi-enclosed sea which hosts one of the largest reef systems in the world (Chaidez *et al.*, 2017).

2 Data

2.1 Original Data and Preprocessing

Daily gridded data at a spatial resolution of $1/20^\circ$ (i.e., the internodal grid length is approximately 5.5km) are produced for the period 1985–2015 by the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. This data product is based on satellite data provided by international agencies, as well as in situ data from ships and buoys, in order to produce accurate SST estimates; see Donlon *et al.* (2012) for more details.

Figure 1 shows temperature time series for three locations, while Figure 2 displays the spatial variability of the data for August 5, 2000. As expected, the data show a clear seasonal pattern and a North–South temperature gradient. There are many ways to deal with this non-stationary behavior. Let $Y(s,t)$ denote the Red Sea surface temperature observed at location $s \in \mathcal{S} \subset \mathbb{R}^2$ and time $t \in \mathcal{T} = \{1, \ldots, T\}$. The Red Sea $\mathcal{S}$ is discretized into $S = 16703$ grid cells, and there are $T = 11315$ days in total, giving about 188 million data points. Here, we decompose $Y(s,t)$ into a mean effect $\mu(s,t)$ and the anomaly (or residual component) $A(s,t)$, i.e.,

$$Y(s,t) = \mu(s,t) + A(s,t).$$

As the data appear to be roughly stationary year-by-year during the period 1985–2015, we simply estimate $\mu(s,t)$ by computing the temperature average for each specific grid cell and each day of the year (by pooling together the 31 years). We then smooth the estimated mean by computing, for each grid cell separately, a moving average over windows of size one week. This yields the estimated mean effect $\widehat{\mu}(s,t)$, and the estimated anomalies $\widehat{A}(s,t)$ are obtained as

$$\widehat{A}(s,t) = Y(s,t) - \widehat{\mu}(s,t).$$

⇒ *The different teams participating to this competition are directly working with (a subset of) the anomalies $\widehat{A}(s,t)$, and do not have access to the original data $Y(s,t)$.*

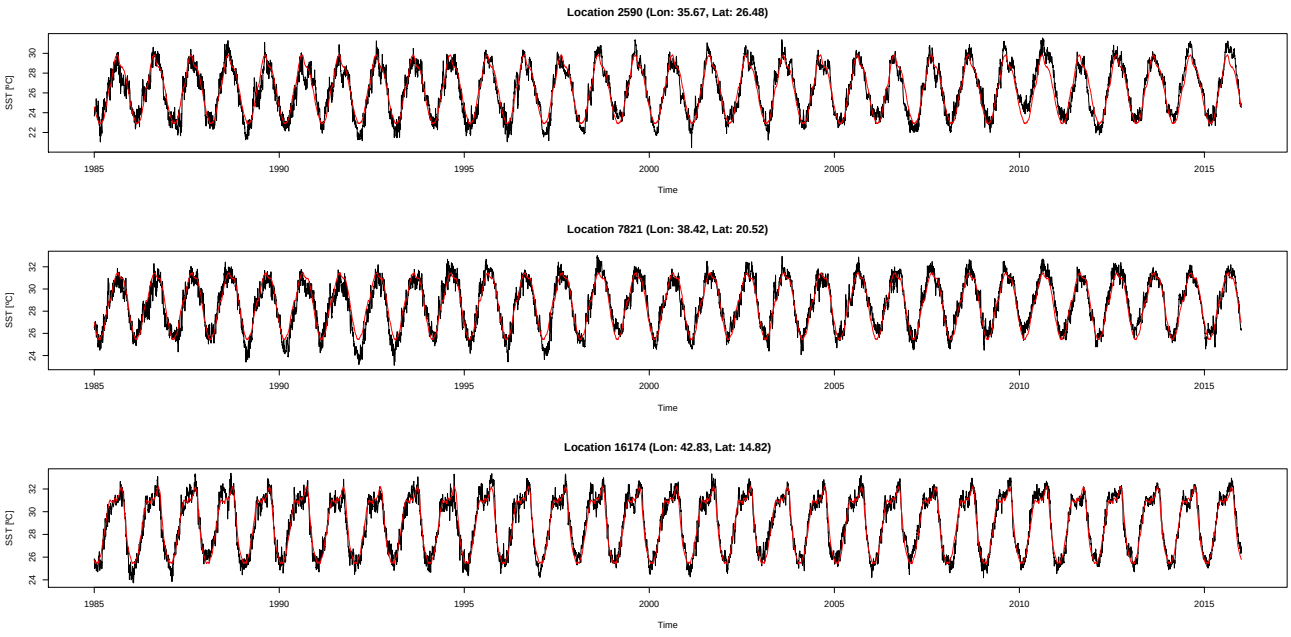

Figure 1: Sea surface temperature time series during the entire period 1985–2015 (black) for three locations in the Red Sea (top to bottom panels correspond to North to South locations); see Figure 2 for the exact locations. The estimated temperature mean is shown in red.

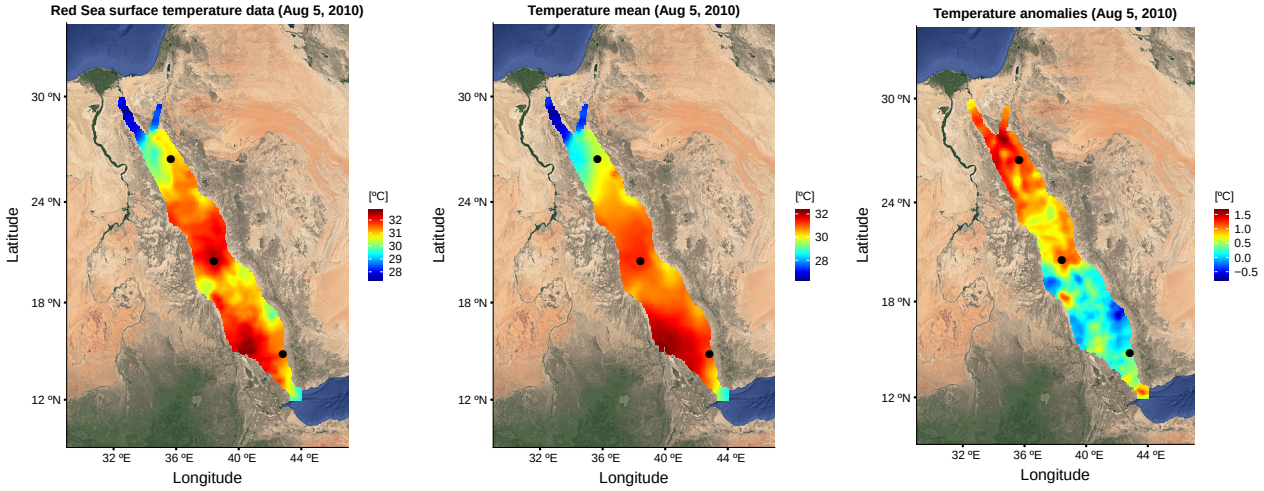

Figure 2: Sea surface temperature data $Y(s,t)$ for the whole Red Sea (left), its estimated mean $\mu(s,t)$ (middle), and the resulting temperature anomaly $A(s,t)$ (right) for August 5, 2000. Complete time series at three highlighted locations (black dots) are shown in Figure 1.

2.2 Training and Validation Datasets

For this competition, about 31.6% of the original anomaly data are artificially masked (at various places in space and time) by introducing missing values (`NA`s in `R`). As illustrated in Figure 3, data may be missing over fairly large spatial areas for a whole month or more. The missing data mechanism is independent of the observable variables.

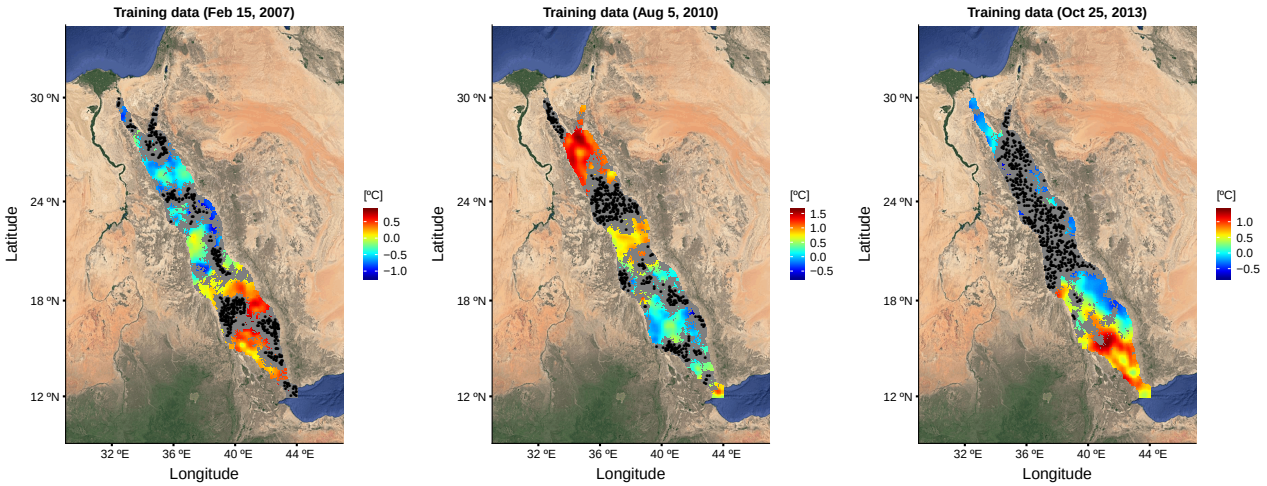

Figure 3: Training data, consisting of sea surface temperature anomalies, here shown for three specific days: February 15, 2007 (left), August 5, 2010 (middle), and October 25, 2013 (right). Grey areas correspond to missing values, and (overlaid) black dots are validation locations.

⇒ *The training dataset (available to the teams) consists of all non-missing temperature anomaly values. It is indexed by the subset* $\mathcal{X}_T \subset \mathcal{X} = \mathcal{S} \times \mathcal{T}$.

⇒ *The validation dataset (not available to the teams) consists of a subset of the missing values. It is indexed by the subset* $\mathcal{X}_V \subset \mathcal{X} \setminus \mathcal{X}_T \subset \mathcal{X} = \mathcal{S} \times \mathcal{T}$, *which comprises* 500 *randomly selected spatial locations for the* 5*th,* 15*th and* 25*th of each month from 2007 to 2015. Its size is* $|\mathcal{X}_V| = 500$ *(locations)* $\times$ 3 *(days per month)* $\times$ 12 *(months)* $\times$ 9 *(years)* $= 162000$. *The full index set* $\mathcal{X}_V$ *(of spatio-temporal locations) is provided to the teams as an* `R` *vector.*

The intersection between the training and validation sets is empty. That is, $\mathcal{X}_T \cap \mathcal{X}_V = \emptyset$.

3 Goal, Evaluation Criterion and Benchmark

3.1 Setting and General Objective

The most devastating ecological and environmental degradations are often caused by large-scale extreme temperature events, which are persistently hotter than their usual level and can simultaneously affect an entire region over a period of time.

For each space-time location $(s,t) \in \mathcal{X} = \mathcal{S} \times \mathcal{T}$, consider a local neighborhood $\mathcal{N}(s,t) \subset \mathcal{X}$. Here we take $\mathcal{N}(s,t)$ to be a 'vertical space-time cylinder', i.e.,

$$\mathcal{N}(s,t) = \{\mathcal{B}(s,r) \times \{t-3, t-2, t-1, t, t+1, t+2, t+3\}\} \cap \mathcal{X},$$

where $\mathcal{B}(s,r)$ is a *ball* centered at location s of radius $r = 50$km. Note that data are missing three days before and after each point $(s,t) \in \mathcal{X}_V$ in the validation set. We define spatio-temporal extremes as events such that

$$X(s,t) = \min_{(\tilde{s},\tilde{t}) \in \mathcal{N}(s,t)} \widehat{A}(\tilde{s},\tilde{t}) > u, \tag{1}$$

for some large threshold u, where $\widehat{A}(s,t)$ denotes the estimated temperature anomalies.

$\Rightarrow$ *In other words, an event is extreme if the sea surface temperature is simultaneously larger than its mean by* u°C *for at least one week over a (circular) area of radius* 50*km.*

The general objective of this competition is to accurately predict the distribution of $X(s,t)$ defined in (1) for all space-time locations in the validation set, i.e., for all $(s,t) \in \mathcal{X}_V \subset \mathcal{X} = \mathcal{S} \times \mathcal{T}$.

3.2 Evaluation Criterion

Let $\widehat{F}_{s,t}(x)$ denote the predicted distribution (or 'probabilistic forecast') for $X(s,t)$, assumed to have a finite first moment. In order to verify the calibration and sharpness of the predicted distribution $\widehat{F}_{s,t}(x)$, while focusing on the upper tail, we use the threshold-weighted continuous ranked probability score (twCRPS) defined as

$$\text{twCRPS}(\widehat{F}_{s,t}, x_{s,t}) = \int_{-\infty}^{\infty} \{\widehat{F}_{s,t}(x) - \mathbb{I}(x_{s,t} \leq x)\}^2 w(x) \mathrm{d}x, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $x_{s,t}$ is the observed (realized) value of $X(s,t)$, $w(x) = \Phi\{(x-1.5)/0.4\}$ and $\Phi(\cdot)$ denotes the standard normal distribution. The chosen weight function $w(x)$ is depicted in Figure 4.

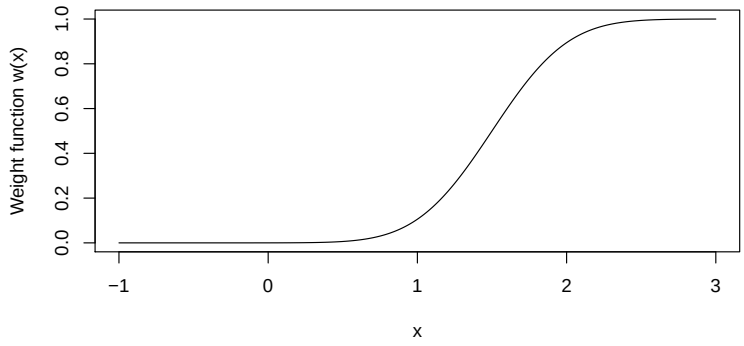

Figure 4: Weight function $w(x) = \Phi\{(x-1.5)/0.4\}$ used for computing the twCRPS.

Although the twCRPS requires the full distribution $\widehat{F}_{s,t}(x)$, it puts the emphasis on temperature anomaly values greater than $u \approx 1$°C. The twCRPS is a proper scoring rule with our choice of weight function $w(x)$; see Gneiting and Raftery (2007), Gneiting and Ranjan (2011), Lerch and Thorarinsdottir (2013), and Lerch *et al.* (2017).

To compute the twCRPS in practice, we restrict the integral in (2) to the interval $[-1, 3]$ and make the following approximation:

$$\text{twCRPS}(\widehat{F}_{s,t}, x_{s,t}) \approx \widehat{\text{twCRPS}}(\widehat{F}_{s,t}, x_{s,t}) = \frac{1}{100} \sum_{k=1}^{400} \{\widehat{F}_{s,t}(x^k) - \mathbb{I}(x_{s,t} \leq x^k)\}^2 w(x^k),$$

where the 'design points' are set to $x^k = -1 + k/100$, $k = 1, \ldots, 400$.

The overall prediction accuracy is then assessed by averaging the $\widehat{\text{twCRPS}}$ values over the validation set $\mathcal{X}_V \subset \mathcal{X} = \mathcal{S} \times \mathcal{T}$, i.e.,

$$\overline{\text{twCRPS}} = \frac{1}{|\mathcal{X}_V|} \sum_{(s,t)\in\mathcal{X}_V} \widehat{\text{twCRPS}}(\widehat{F}_{s,t}, x_{s,t}).$$

The final ranking of the different teams will be based on $\overline{\text{twCRPS}}$. Lower values are better.

3.3 Benchmark

We construct a benchmark prediction by following two basis steps:

(i) Spatio-temporal minimum: From the training dataset of anomalies $\widehat{A}(s,t)$, we compute the spatio-temporal minimum $x_{s,t} = \min_{(\tilde{s},\tilde{t})\in\mathcal{N}(s,t)} \widehat{A}(\tilde{s},\tilde{t})$ for all points $(s,t) \in \mathcal{X}$ that have complete neighborhoods $\mathcal{N}(s,t)$ (i.e., with no missing values).

(ii) Benchmark prediction: Then, assuming spatio-temporal stationarity, the benchmark prediction $\widehat{F}^{\text{ben}}_{s,t}$ is defined for each location $(s,t) \in \mathcal{X}_V$ as the empirical distribution function obtained by pooling together all available data $x_{s,t}$ over space and time.

4 Deliverables

Each team has to provide a matrix (of class "`matrix`" in `R`) with the following properties:

1. The matrix should be called `prediction`, and saved into an `R` object called `prediction_name-of-team.RData` (using the `R` function `save(...)`);
2. The matrix should be of dimension $|\mathcal{X}_V| \times 400 = 162000 \times 400$;
3. The (j,k)th entry of the matrix should contain

$$\widehat{F}_{s_j,t_j}(x^k), \qquad (s_j,t_j) \in \mathcal{X}_V.$$

As mentioned above, we use $x^k = -1 + k/100$, $k = 1,\ldots,400$; In other words, each row of the matrix should contain the predicted distribution, evaluated at each of the 400 design points x^k, for the j-th space-time location (s_j,t_j) in the validation set $\mathcal{X}_V$.

In addition, each team has to provide a clean and commented `R` code to be able to reproduce the results if needed.

Each team will be provided with a dropbox folder to submit the results and the code.

5 Timeline and Final Deadline

1. Preliminary prediction 1 (optional): March 31, 2019 (Sunday)
2. Preliminary prediction 2 (optional): May 12, 2019 (Sunday)

3. **Final prediction: June 9, 2019 (Sunday) at 23:59 UTC**

4. *EVA 2019 Conference: July 1–5, 2019.*

Each team can submit up to two preliminary predictions to verify their approach and improve their model. The final ranking, however, will be based on the final prediction only. Preliminary and final rankings (along with twCRPS values) will be posted on the conference webpage.

6 Rules

1. There is no limit to the number of teams or team members.

2. Only the final submission will be taken into account to rank the teams.

3. Submission of preliminary predictions is not mandatory, but highly encouraged.

4. Results must be submitted as specified above in §4, and the submitted `R` code must be clean and properly commented to be able to reproduce the results if needed.

5. Reverse engineering, or the use of other data sources (such as extra covariates, etc.), is strictly prohibited.

6. Late submissions will not be considered.

7. Failure to comply with the above rules may result in disqualification.

7 Rewards

1. The best-ranked teams will be invited to present their work in an invited session at the EVA 2019 conference organized in Zagreb, Croatia, during the week of July 1–5, 2019.

2. After the EVA 2019 conference, *all* the teams will be invited to submit a paper describing their approach for publication in the journal *Extremes*. The submitted papers will undergo a usual peer-reviewed process (with the same quality standards and acceptance criteria).

8 Getting Started

1. Register your team (and specify a team name) by sending an email to Raphaël Huser at raphael.huser@kaust.edu.sa. You will then be sent the data, the main `R` script to load the data, as well as the dropbox folder to submit your predictions.

2. Open an `R` terminal, open the script `Competition.R`, and read the instructions.

References

Cantin, N. E., Cohen, A. L., Karnauskas, K. B., Tarrant, A. M. and McCorkle, D. C. (2010) Ocean warming slows coral growth in the central Red Sea. *Science* **329**(5989), 322–325.

Chaidez, V., Dreano, D., Agusti, S., Duarte, C. M. and Hoteit, I. (2017) Decadal trends in Red Sea maximum surface temperature. *Scientific Reports* **7**(8144), 1–8.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E. and Wimmer, W. (2012) The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sensing of Environment* **116**, 140–158.

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

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**(3), 411–422.

Lerch, S. and Thorarinsdottir, T. L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography* **65**(1), 21206.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017) Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science* **32**(1), 106–127.

McClanahan, T. R., Ateweberhan, M., Muhando, C. A., Maina, J. and Mohammed, M. S. (2007) Effects of climate and seawater temperature variation on coral bleaching and mortality. *Ecological Monographs* **77**(4), 503–525.

Raphaël Huser (email: raphael.huser@kaust.edu.sa) – King Abdullah University of Science and Technology (extstat.kaust.edu.sa)