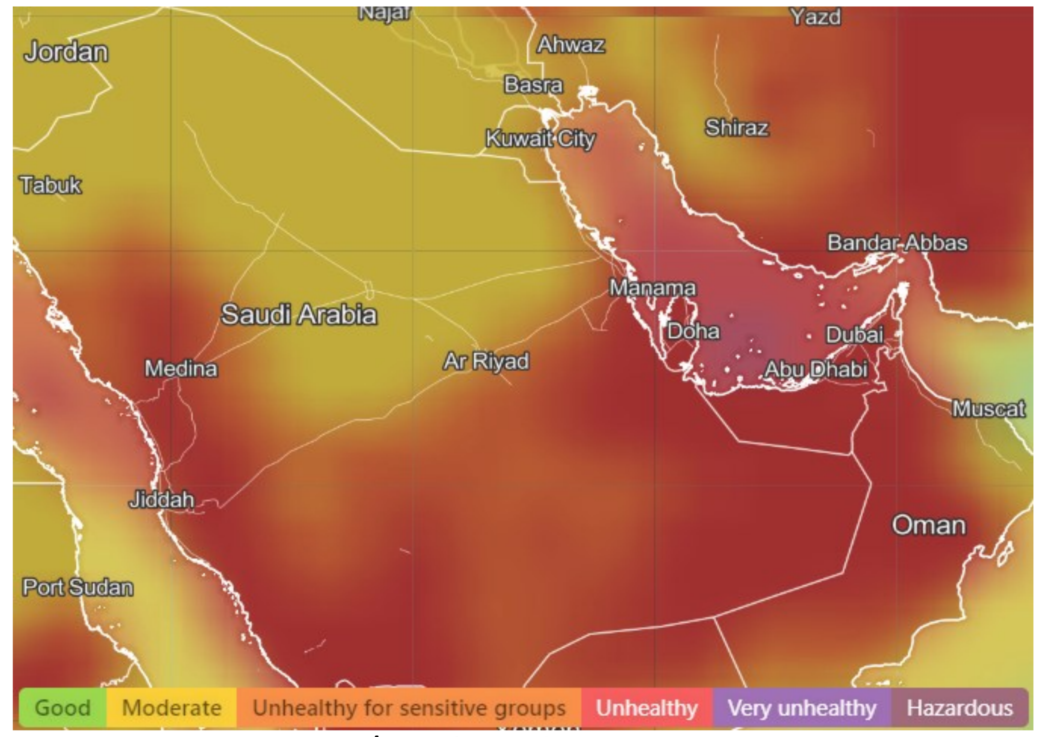
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Statistical Modeling of Extreme PM2.5 Concentration in Saudi Arabia Using Deep Learning

Abdullah Alkathiry (KFUPM, Saudi Arabia)
Dr. Jordan Richards (KAUST, Saudi Arabia)
Dr. Arnab Hazra (IIT Kanpur, India)
Prof. Raphael Huser (KAUST, Saudi Arabia)

Motivation

- Saudi Arabia is ranked 21st in the world in terms of poor air quality.
- PM2.5 is one major air pollutant and extreme PM2.5 concentration in the air is highly unhealthy.
- Identifying regions with extreme PM2.5 concentration is necessary for warning and mitigation.



From www.iqair.com/saudi-arabia

Objectives

- Produce spatial exceedance probability maps and consider temporal variation.
- Model high threshold exceedances of PM2.5 concentration incorporating relevant covariate information.
- Build a deep learning-based spatiotemporal generalized Pareto distribution (GPD) model, that combines the strength of extreme-value theory models with the representation power of artificial neural networks.

PM2.5 data

Monthly ground-level PM2.5 concentration ($\mu g/m^3$) that combines data from a number of sources.
Our study includes 36 predictors. Some of them are:

- Meteorological variables from the ERA5 reanalysis on land surface
- Land cover proportions that are provided by Copernicus, the EU's Earth observation program
- Orographical variables that are derived from Amazon Web Services terrain tiles

Methodology

- A statistically justified model by **extreme-value theory** for the tail of a random variable is the GPD with the following distribution function (CDF)

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}},$$

where $\sigma > 0$ and $\xi \in \mathbb{R}$ are the scale and shape parameters, respectively. In other words, for large u, we have

$$\Pr(Y > u + y \,|Y > u) \approx H(y), \qquad y > 0.$$

- The GPD is fitted to the exceedances $(y_i - u)|y_i > u$, for some large threshold u.
- We assume ξ to be unknown but fixed over space and time.
- σ is assumed to be varying across space and time (denoted by $\sigma(s,t)$), and

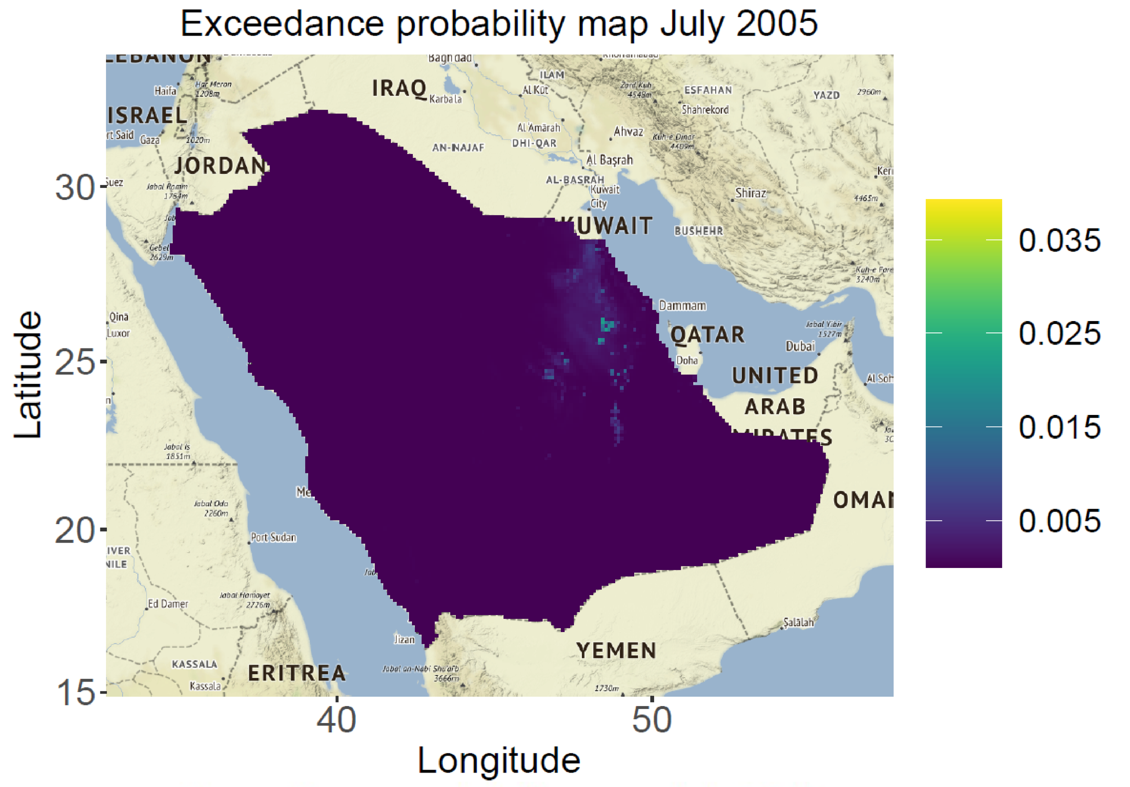
$$\log(\sigma(s,t)) = f(X_1(s,t), \dots, X_{36}(s,t)),$$

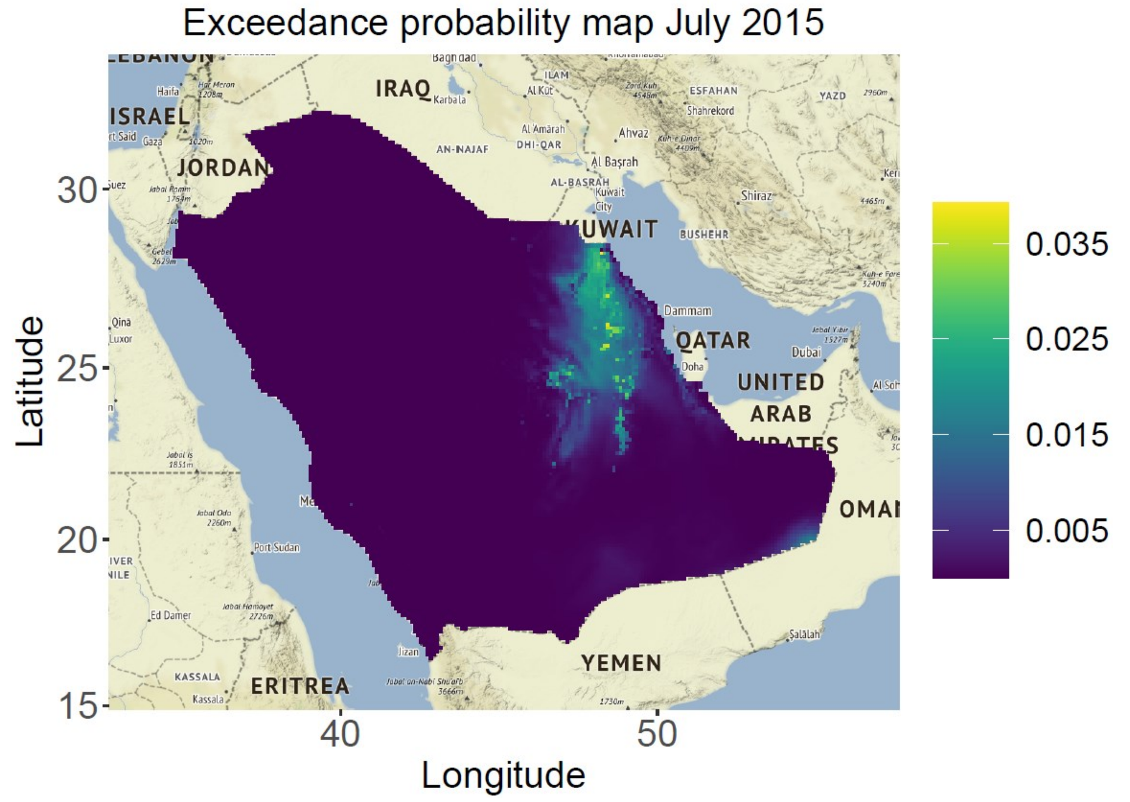
where f is an unknown nonlinear function.
- The parameters $\sigma(s,t)$ and ξ are estimated using **neural networks** with a single negative log-likelihood loss function.
- The dense neural network we used for σ has 4 layers with 87 neurons and 1291 learnable parameters while ξ network has only one layer with only one learnable parameter.
- The threshold level for the GPD is a quantile, estimated by another neural network minimizing the tilted loss, given by

$$e(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} \max\{\tau(y_i - \hat{y}_i), (\tau - 1)(y_i - \hat{y}_i)\}$$

where $\tau \in (0,1)$ is a pre-specified probability.

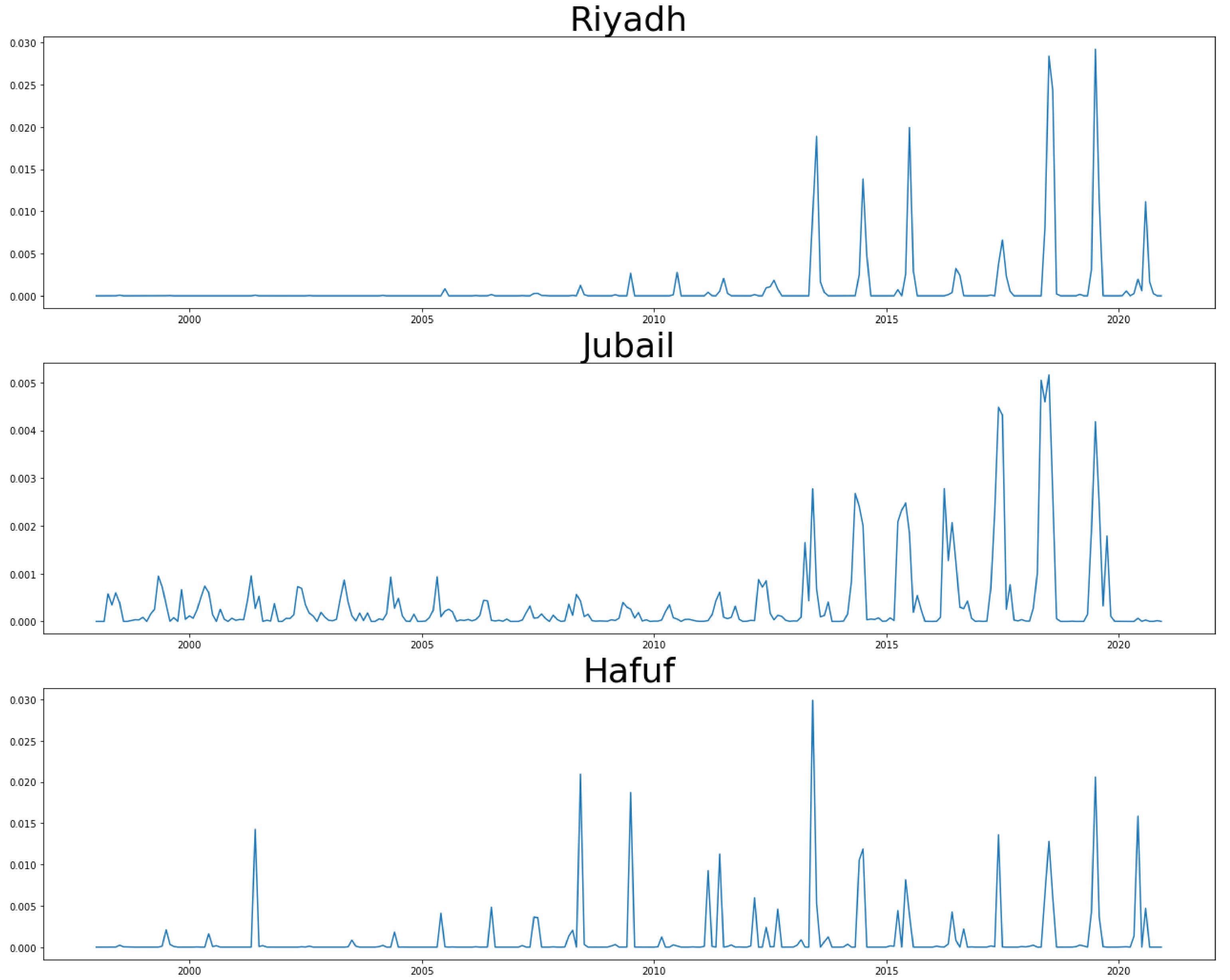
Results





The maps show the monthly probability estimates of exceeding the threshold of **150.4 $\mu g/m^3$**, which is considered unhealthy level by the U.S. Environmental Protection Agency (EPA).

The exceedance probability in major cities



Conclusion

- Our deep learning model can predict the probability of exceeding the unhealthy threshold and learn the complicated relationships between the output and the covariates.
- The results show that the exceedance probability is highly dependent on spatial and temporal location.
- Exceedance probability from different cities reveal an **alarming increase in pollution** over the years.
- In future research, it would be interesting to use convolutional and/or recurrent neural networks to more effectively capture spatio-temporal characteristics.