

## Background

The Ramganga Water Quality Fusion project is developing methods for fusing water quality data from multiple sources to improve our understanding of water quality in rivers with sparse in-river measurements.

The motivating application is the Ramganga river (see Figure 1), a major tributary of the Ganga in northern India, which serves as a vital source of water for household, agricultural and industrial purposes. The river is highly polluted due to agricultural runoff, chemicals from industries and untreated sewage from settlements along its banks all entering the river. It is therefore vital to gain an understanding of spatial and temporal patterns of water quality parameters within the river.



Figure 1: Images of the Ramganga river around the city of Moradabad in February 2020.

In-river measurements are highly accurate due to laboratory analysis of samples, but they are highly expensive to collect, requiring field trips over long distances. These data usually therefore provide limited spatial and temporal coverage of the river water quality. Satellite data have recently become a useful new source of data for rivers, with the European Space Agency's Sentinel-2 programme now providing appropriate reflectance data on a high spatial resolution grid that allows for retrieval of water quality parameters for rivers. Drone data provide a solution to fill in the gaps in satellite data coverage during dry seasons, over short stretches of river of particular interest. These three data sources provide complimentary information, which if combined can provide high accuracy and comprehensive spatial and temporal coverage. The project is therefore developing methodology to allow this data fusion, with clear assessments of uncertainties provided.

The project team is made up of scientists from the University of Glasgow, the University of Stirling, the James Hutton Institute, the UK Centre for Ecology and Hydrology and the Indian Institute of Technology, Kanpur.



## Data

Data are currently being collected by the Indian Institute of Technology, Kanpur, team. These data will consist of in-river samples that have been analysed in the laboratory, along with measures of reflectance using a "WISP" spectrometer and additional data from a sonde



monitoring device. A drone will also be flown during these field trips to provide high spatial resolution data, initially over several kilometres around the city of Moradabad. The data will be collected from around 15 sites along the 350km length of the river from the Kalagarh Dam to the confluence with the Ganga, as shown in Figure 2.



Figure 2: Modified normalised water difference index of the Ramganga river between the Kalagarh Dam and the confluence with the Ganga.

Legend:

- Site on Ramganga river
- Site on Ramganga tributary
- ▲ Site on Ganga river

Flow direction is northwest to southeast.

In-river data will be collected at each of the 15 sampling sites shown. Drone data will be collected at a small number of sites.

Satellite images have been obtained from the European Space Agency's Sentinel-2 mission. Four tiles cover the river at a spatial resolution of 10m by 10m. Data are available from 2015 to present. The University of Stirling team are converting the reflectance data to the scale of the water quality variables of interest.

Expected spatial and temporal resolutions of the three data sources are given in Table 1.

Table 1: Expected spatial and temporal resolutions of the three data sources. \*Some months of satellite data may be missing due to narrowness of river in pre-monsoon season.

| Dataset   | Spatial resolution                   | Temporal resolution |
|-----------|--------------------------------------|---------------------|
| In situ   | ~15 sites                            | every few months    |
| Satellite | 10m × 10m grid                       | monthly*            |
| Drone     | 2m × 2m grid up to several km length | every few months    |

Initial variables of interest are chlorophyll-a and total suspended matter, since both of these parameters relate directly to water quality and can be identified from remote sensing reflectance imagery.



## Data simulation

The data are not yet available in sufficient quantity for model fitting. Data simulation is therefore required. Data are simulated along a 50km by 50m rectangle that represents a stretch of river, with values representing some water quality variable.

A trend along the length of the river is generated using sine curves of different periods, giving the nonlinear pattern shown in Figure 3(a), with a general decrease in value with increasing x-coordinate before changing to a general increase in value with x-coordinate. The data values are generated to be larger closer to the river banks, as illustrated in Figure 3(b). A seasonal pattern of one peak per year in data value is simulated, with an additional long-term nonlinear trend, as shown in Figure 3(c). Five years of data are simulated in total, on a monthly timescale and on the 2m by 2m grid required for the drone data. Data values are generated by adding the appropriate values of the trend in the x, y and time components.

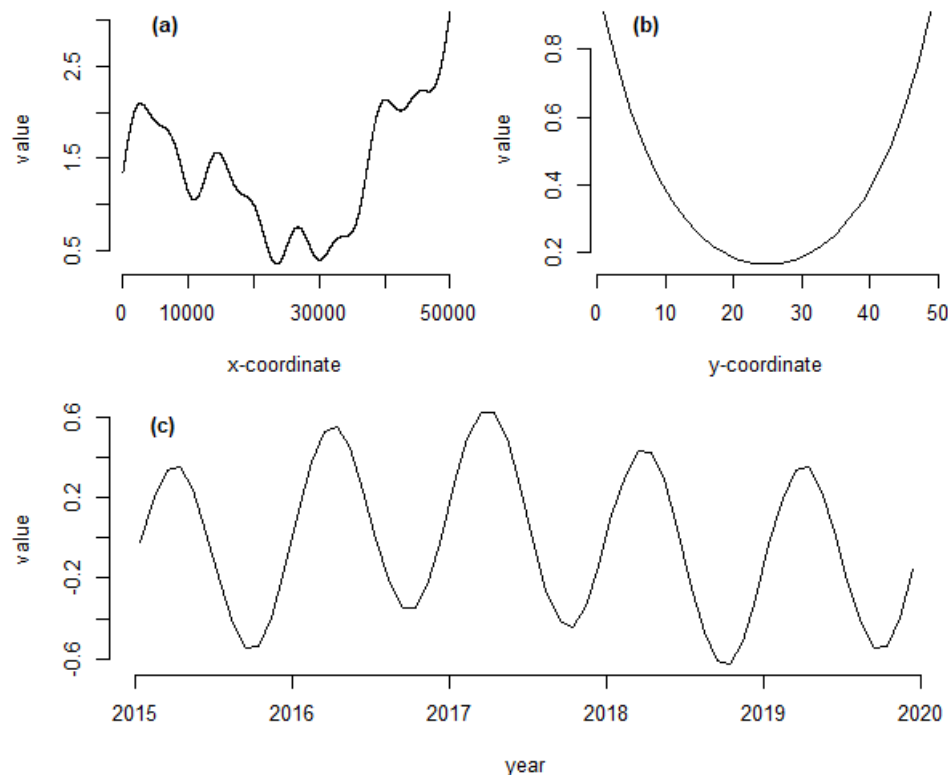


Figure 3: (a) Simulated trend along x-coordinate; (b) Simulated trend along y-coordinate; (c) Simulated trend over time.

Two point sources are also generated, which affect only the river downstream of each source, with a polygon generated to define this downstream area in each case. The effect of the point source is generated to decline with distance from source according to an exponential spatial covariance function.

## Statistical modelling

There are many different methods for data fusion in the statistical literature. For example, Fuentes and Raftery (2005) propose Bayesian melding, which deals with change of support between data types through fitting integrals and is therefore computationally intensive.



Nguyen et al (2012) propose a method for spatial statistical data fusion based on fixed rank Kriging, but this method is not suitable for addressing the temporal change of support. Brynjarsdóttir and Berliner (2014), and Yarali and Rivaz (2020), both propose methods of dimension-reduced statistical downscaling, where the spatial processes are modelled using low dimension bases. Their method can represent large spatial regions, but assumes discrete time. Ludwig et al (2017) propose roving sensor data fusion, which allows for moving sensor locations in continuous space and time, but may be computationally intensive over the large regions required for assessment of water quality in rivers. Another set of methods is made up of those based on spatially-varying coefficients regression, such as Sahu et al (2010), Berrocal et al (2010) and Wilkie et al (2019). Nonparametric statistical downscaling is in the latter group of methods and will be described here.

Let  $\mathbf{y}_i$  be a vector of point-location, point-time data observed at location  $i$  ( $i = 1, \dots, n$ ) with length  $q_i$ . This response vector is assumed to be accurate up to measurement error. Let  $\mathbf{x}_i$  be a  $p_i$ -length vector of grid-scale, possibly temporally-averaged data, for the grid cell that contains point location  $i$ . Then the data at each location are modelled as observations of smooth functions using basis functions:

$$\mathbf{y}_i | \mathbf{c}_i, \sigma_y^2 \sim N(\boldsymbol{\phi}_i \mathbf{c}_i, \sigma_y^2 \mathbf{I}),$$

$$\mathbf{x}_i | \mathbf{d}_i, \sigma_x^2 \sim N(\boldsymbol{\psi}_i \mathbf{d}_i, \sigma_x^2 \mathbf{I}),$$

where  $\boldsymbol{\phi}_i$  and  $\boldsymbol{\psi}_i$  are matrices of basis functions of dimensions  $(q_i \times m)$  and  $(p_i \times m)$ , respectively, with  $\mathbf{c}_i$  and  $\mathbf{d}_i$  being their corresponding  $m$ -length vectors of basis coefficients. The basis coefficients are then related via regression, with  $i = 1, \dots, n$  and  $j = 1, \dots, m$ :

$$c_{ij} | \alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 \sim N(\alpha_{ij} + \beta_{ij} d_{ij}, \sigma_c^2),$$

where the intercept and slope parameters vary smoothly over space according to an exponential spatial covariance:

$$\boldsymbol{\alpha}_j | \sigma_\alpha^2 \sim N(\mathbf{0}, \sigma_\alpha^2 \exp(-\boldsymbol{\phi}_\alpha \mathbf{D})),$$

$$\boldsymbol{\beta}_j | \sigma_\beta^2 \sim N(\mathbf{1}, \sigma_\beta^2 \exp(-\boldsymbol{\phi}_\beta \mathbf{D})),$$

with spatial variances  $\sigma_\alpha^2$  and  $\sigma_\beta^2$ , and spatial decay parameters  $\boldsymbol{\phi}_\alpha$  and  $\boldsymbol{\phi}_\beta$  controlling the speed of decline of correlation to zero as distance increases, with  $\mathbf{D}$  being the  $(n \times n)$  matrix of distances between response data locations. All variance parameters are given inverse-gamma(2,1) prior distributions following Sahu et al (2010), while the spatial decay parameters are best chosen outwith the model, for example through a cross validation procedure. To complete the model, the basis coefficients  $\mathbf{d}_i$  are given a Normal prior distribution:

$$\mathbf{d}_i \sim N(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d),$$

where  $\boldsymbol{\mu}_d$  and  $\boldsymbol{\Sigma}_d$  are chosen a priori, with suggested values being  $\mathbf{0}$  and some multiple of  $\mathbf{I}$ , respectively, reflecting a lack of prior knowledge of the signs and dependence structures of the components of  $\mathbf{d}_i$ .

Predictions can be made at any point in continuous space and time, with associated uncertainties given.



The model can be fitted efficiently using Markov Chain Monte Carlo, in C++ via R. An R package and shiny app are in development.

## Results

The model was fitted to the simulated data and predictions made. Example predictions are presented in Figure 4, for 1024 points over a short section of river of length 200m and width 50m, for 14<sup>th</sup> January 2015. Here, the response dataset is the in situ data for 10 point locations and the explanatory variable is the satellite data over a 10m by 10m grid.

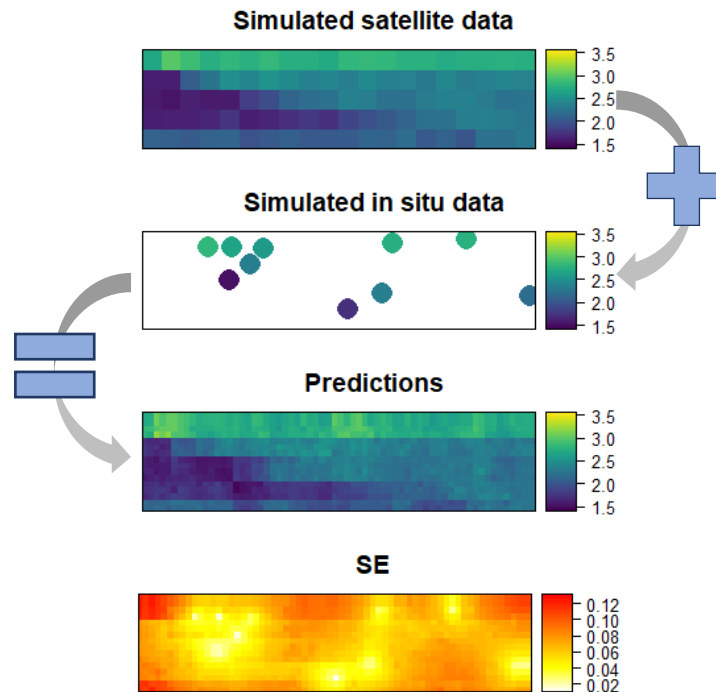


Figure 4: Simulated satellite data, in situ data and predictions from the nonparametric statistical downscaling model for a 200m by 50m stretch of river on 14<sup>th</sup> January 2015. Standard errors of predictions are also shown.

The predictions are able to reproduce the spatial patterns in the data well, with the spatial patterns of the satellite data retained after calibration with the in situ data. The standard errors of predictions are lowest close to the in situ data locations, as expected.

Figure 5 shows predictions for a single point location over time, with that point location having been left out of the data and predicted using the model fitted to data for the 9 other point locations. The model is able to reproduce the temporal patterns in the data well, but 95% credible intervals are wide. These wide intervals may be due to the large spatial region to which the model is fitted.



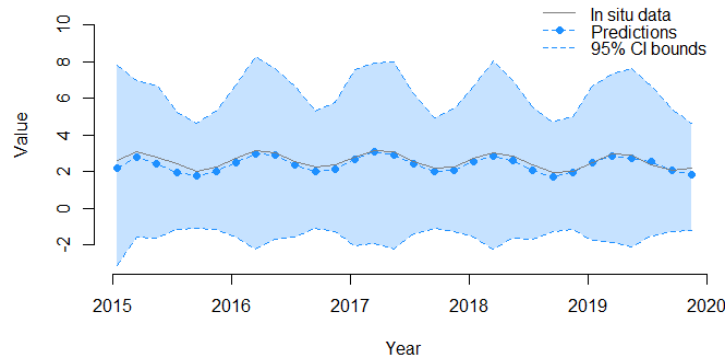


Figure 5: In situ data and predictions at the data location from the nonparametric statistical downscaling model fitted to the other 9 point locations, with 95% credible interval bounds shown.

## Future work

Extensions to model multiple sources of data are in development. Two main options are:

- Include the satellite and drone data as covariates in the same model, i.e. have an additional line  $\mathbf{z}_i | \mathbf{e}_i, \sigma_z^2 \sim N(\boldsymbol{\omega}_i \mathbf{e}_i, \sigma_z^2 \mathbf{I})$ , with  $c_{ij} | \alpha_{ij}, \beta_{ij}, \delta_{ij}, d_{ij}, e_{ij}, \sigma_c^2 \sim N(\alpha_{ij} + \beta_{ij} d_{ij} + \delta_{ij} e_{ij}, \sigma_c^2)$ .
- 2-stage model with satellite and drone data fused to create a new surface SaDr, with this product then fused with the in situ data.

The first option is simple to implement and close in speed to the model described previously, but it may not be appropriate to treat the satellite and drone data in the same way, due to differences in their spatial and temporal resolutions. The second option is potentially more flexible, without the need for both satellite and drone data to be available at the prediction sites for at least some timepoints, allowing for optimal use of the available data, but it results in a more complex model that is slower to run. Future work will investigate the optimal model structure to deal with the three main data sources in this application.

Further developments will focus on adapting the method to address river distance, which may differ substantially from Euclidean distance, and to address the directionality created by river flow direction, potentially following some of the work of Ver Hoef et al (2006). Incorporating river flow itself as an additional parameter may also be investigated.

## Acknowledgements

The Ramganga Water Quality Fusion project is funded by the UK EPSRC Global Challenges Research Fund (EP/T003669/1).

## Team members:

University of Glasgow: Dr. Surajit Ray, Prof. Marian Scott, Prof. Claire Miller, Dr. Craig Wilkie  
University of Stirling: Prof. Andrew Tyler, Dr. Peter Hunter, Dr. Veloisa Mascarenhas  
UK Centre for Ecology and Hydrology: Dr. Dan Read, Dr. Mike Bowes  
The James Hutton Institute: Dr. Rachel Helliwell, Dr. Miriam Glendell  
Indian Institute of Technology, Kanpur: Prof. Rajiv Sinha, Dr. Manudeo Singh, Dr. Bharat Choudhary, Umar Farooq

Further information on the project is available at <https://surajitstat.github.io/ramganga/>.



## References:

- Berrocal, V.J., Gelfand, A.E. & Holland, D.M. (2010). 'A spatio-temporal downscaler for output from numerical models'. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**(2):176-197.
- Brynjarsdóttir, J. & Berliner, L.M. (2014). 'Dimension-reduced modeling of spatio-temporal processes'. *Journal of the American Statistical Association*, **109**(508):1647-1659.
- Fuentes, M. & Raftery, A.E. (2005). 'Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models'. *Biometrics*, **61**(1):36-45.
- Ludwig, G., Chu, T., Zhu, J., Wang, H., Koehler, K. et al. (2017). 'Static and roving sensor data fusion for spatio-temporal hazard mapping with application to occupational exposure assessment'. *The Annals of Applied Statistics*, **11**(1):139-160.
- Nguyen, H., Cressie, N. & Braverman, A. (2012). 'Spatial statistical data fusion for remote sensing applications'. *Journal of the American Statistical Association*, **107**(499):1004-1018.
- Sahu, S.K., Gelfand, A.E. & Holland, D.M. (2010). 'Fusing point and areal level space-time data with application to wet deposition'. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**(1):77-103.
- Ver Hoef, J.M., Peterson, E. & Theobald, D. (2006). 'Spatial statistical models that use flow and stream distance'. *Environmental and Ecological Statistics*, **13**:449-464.
- Wilkie, C.J., Miller, C.A., Scott, E.M., O'Donnell, R.A., Hunter, P.D., Spyarakos, E. & Tyler, A.N. (2019). 'Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support'. *Environmetrics*, **30**(3):e2549.
- Yarali, E. & Rivaz, F. (2020). 'Incorporating covariate information in the covariance structure of misaligned spatial data'. *Environmetrics*, **31**(6):e2623.

