Recent Improvements in Spatial Regression of Climate Data

Jouke H.S. de Baar, Irene Garcia-Marti
Royal Netherlands Meteorological Institute (KNMI)
Utrechtseweg 297, 3731 GA De Bilt
THE NETHERLANDS

jouke.de.baar@knmi.nl; irene.garcia.marti@knmi.nl

Keywords: multi-fidelity, Gaussian process, noise treatment, Rastrigin, climatology.

ABSTRACT

Climate and weather services rely, for a large part, on data obtained from in situ *measurement stations. For this purpose, the Netherlands operates an official network of high-fidelity stations. However, over recent years, data has become available from multiple lower-fidelity observation networks. We are currently investigating how this data, of different fidelity levels, can be combined in a single service. As in multi-fidelity surrogate-based vehicle optimisation, Gaussian process regression (or Kriging) is a key method in our field of application, and in both vehicle design optimisation and meteorological services we are challenged by similar issues in the efficient application of such multi-fidelity surrogates. In this work, we focus on quality control, noise treatment, inclusion of high-resolution covariates, and improvement of the reliability of the predicted local uncertainty, all in the context of spatial regression of multi-fidelity data.*

DISCLAIMER

KNMI is continuously researching possible ways to improve services. The methods and results in this work should be considered as fit for research purposes, not as part of official KNMI services. After sufficient research, KNMI aims to make new methods operational as soon as possible.

1.0 INTRODUCTION

Meteorological observations are fundamental to sustain a wide range of applications at national meteorological and hydrological services (NMHSs). Official national monitoring networks acquire weather and climate measurements, which are quality-controlled and subsequently introduced in a data assimilation process intended to refine the (regional to global) weather forecasts resulting from the numerical weather prediction process (NWP). Observations are also key for the statistical post-processing and verification of the NWP, which are stages intended to enhance the forecast, and a substantial number of observations is required for this purpose.

Typically, NMHSs rely on the high-fidelity observations acquired by high-end professional devices from the networks they operate. Official networks operated by NMHSs are often spatially distributed in a way that balances financial considerations with the optimal coverage for large-scale phenomena at monitoring sites that meet prescribed measurements conditions (e.g., regulations from World Meteorological Organization, WMO). Hence, NMHSs are in practice limited by practical and financial boundaries in the number of official measurements it can collect, which implies that large regions often remain unobserved (Garcia-Marti et al., 2022).

In the past two decades our global society has witnessed the advent of new technological and scientific advances that have enabled additional ways of collecting observations from the environment. The ubiquity of

wireless networks and the decreasing hardware prices, resulting in 'the internet of things', have made possible that today we can practically collect weather observations anywhere on Earth. Public government agencies are not alone anymore at the arduous task of monitoring the weather, because citizens, organizations, and companies are now able to contribute to this task, by establishing new observational networks.

The availability of these new networks appears to be offering two main advantages:

1) A fundamental trait of these novel networks is that they provide a higher spatial resolution than the official ones. This implies that these alternative weather observations could contribute to the Climate Services by adding spatial details, hence revealing new patterns and trends. We focus on this effect in Section 3.2.
2) The creation of high-resolution weather and climate maps has been shown to benefit from information contained in high-resolution covariates, like land use type, terrain altitude, terrain complexity, population density, etc. However, in particular in a small country like the Netherlands, the number of official stations is not only limited, but by focussing on the consistency of measurements (e.g., obtaining observations in areas of low population density, open terrain with grass, etc.), meteorological organisations could introduce biases for areas that differ substantially from these criteria.. This is likely to lead to extrapolation and over-fitting effects when modelling weather and climate as a function of these covariates. The availability of lower-fidelity stations, which often sample the distribution of covariates more homogeneously (e.g., crowd-sourced data is also available from urban environments, areas of different land use type, etc.), could help to improve such covariate-based approaches. We focus on this effect in Section 3.3.

However, due to various reasons, these lower-fidelity observations contain substantial measurement errors, which need to be addressed. The quantification and treatment of measurement errors (e.g., bias, noise) expressed in uncertainty metrics, is becoming a central theme in science. In this work, we present a multi-fidelity kriging method combining three datasets with variable quality levels for air temperature. Hence, we combine official observations from the Dutch Met Office (KNMI) (i.e., high-fidelity first-party data, 1PD), with measurements acquired along the Dutch road network by Rijkswaterstaat (Directorate-General for Public Works and Water Management) (i.e., medium-fidelity second-party data, 2PD), and crowdsourced data from the WOW-NL network (i.e., low-fidelity third-party data, 3PD) available via http://wow.knmi.nl.

1.1 Specific Contribution to this Workshop

We have chosen the Rastrigin test function to test our methods, as this approach was also used by other workshop participants (Mainini et al, 2022). Further, there are many parallels between the issues we face, and the issues that are encountered in multi-fidelity surrogate-based vehicle performance optimization. The main ingredients are the following. (1) Where the availability of high-fidelity data is limited, there is a potential of using lower-fidelity data as well. This lower-fidelity data typically arrives at lower cost, at larger quantities, and with more noise. (2) Gaussian process regression, or kriging, is a key method in this field. (3) Similar challenges are the effective treatment of noise, as well as improving the reliability of the uncertainty of our predictions. (4) In the future we will also be interested in `adaptive sampling': provided we have a certain budget, where should we place new stations, and of which fidelity level should they be? (5) Currently we are mostly interested in the global RMSE of our maps. However, our services are shifting to analysis, prediction, and warning of impact. In that sense, it will become more interesting to know what the location and severity of the worst impact will be, which is indeed close to surrogate-based optimisation approaches. (6) Either in understanding the climate or reducing vehicle noise and emissions, we share a common goal of making the world a better place for the generations to come.

2 METHODOLOGY

The general problem is to predict the output of a function f(x) conditional on observations $\mathbf{y} = f(\mathbf{x'})$, however, we will see that the observation process can introduce an unknown amount of bias and noise.

We start from the standard Gaussian process regression (GPR) predictors (Ficini et al, 2021):

$$E[f(x)] = m(x) + k(x,\mathbf{x'},\theta)\, K(\mathbf{x'},\mathbf{x'},\theta)^{-1} \{ \mathbf{y} - \mathbf{m}(\mathbf{x'}) \}, \tag{1}$$

$$var[f(x)] = k(x,x,\theta) - k(x,\mathbf{x'},\theta)^{T}\, K(\mathbf{x'},\mathbf{x'},\theta)^{-1}\, k(x,\mathbf{x'},\theta), \tag{2}$$

where we use a Gaussian kernel $k(x,x,\theta)$ which depends on length scale θ. In order to accommodate an unknown drift m that is a linear function of covariates M, we generalize (1) and (2) to (Christensen, 2001):

$$E[f(x)] = M(x)\, \boldsymbol{\beta} + k(x,\mathbf{x'},\theta)\, K(\mathbf{x'},\mathbf{x'},\theta)^{-1} \{ \mathbf{y} - \mathbf{M}(\mathbf{x'})\, \boldsymbol{\beta} \}, \tag{3}$$

$$var[f(x)] = k(x,x,\theta) - k(x,\mathbf{x'},\theta)^{T}\, K(\mathbf{x'},\mathbf{x'},\theta)^{-1}\, k(x,\mathbf{x'},\theta) + C_M(x,x,\theta,\boldsymbol{\beta}), \tag{4}$$

where C_M is an additional term that represents the uncertainty caused by the fitting of the drift (see Christensen, 2001 for details). The hyperparameters $\boldsymbol{\beta}$ act as linear model weights and remain to be estimated.

However, when dealing with multi-fidelity meteorological data, we observe significant bias and noise in the lower-fidelity measurements. Therefore, we model the observation process as $\mathbf{y} = f(\mathbf{x'}) + N(\mathbf{b},R(\mathbf{x'},\mathbf{x'},\boldsymbol{\varepsilon}))$, where we have added a normally distributed observation error with bias $\mathbf{b}$ and uncorrelated noise covariance matrix $R(\mathbf{x'},\mathbf{x'},\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^2\, \mathbf{I}$. For unknown bias, a proxy for the bias can be stacked as additional columns into $\mathbf{M}(\mathbf{x'})$, and as zero columns into M(x), such that we now have: [1]

$$E[f(x)] = M(x)\, \boldsymbol{\beta} + k(x,\mathbf{x'},\theta) \{ K(\mathbf{x'},\mathbf{x'},\theta) + R(\mathbf{x'},\mathbf{x'},\boldsymbol{\varepsilon}) \}^{-1} \{ \mathbf{y} - \mathbf{M}(\mathbf{x'})\, \boldsymbol{\beta} \}, \tag{5}$$

$$var[f(x)] = k(x,x,\theta) - k(x,\mathbf{x'},\theta)^{T} \{ K(\mathbf{x'},\mathbf{x'},\theta) + R(\mathbf{x'},\mathbf{x'},\boldsymbol{\varepsilon}) \}^{-1}\, k(x,\mathbf{x'},\theta) + C_M(x,x,\theta,\boldsymbol{\beta}). \tag{6}$$

The hyperparameters θ, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ can be estimated through maximum likelihood estimation (MLE) (Mardia and Marshall, 1984; Christensen, 2001). It should be noted that the MLE for $\boldsymbol{\beta}$ can be computed analytically, however, the MLE for θ and $\boldsymbol{\varepsilon}$ are found through an – expensive – optimization algorithm (either brute force or differential evolutionary, due to occurring local minima in the objective function).

2.1 Proposed Improvements

As we have seen in the introduction, an important challenge in spatial regression of weather and climate data is accurate downscaling (i.e., upsampling), which is in fact predicting a 'map' E[f(x)] and diag(var[f(x)]) on a high-resolution grid, conditional on a sparse network $\mathbf{x'}$ of observation stations. For example, in the Netherlands we currently aim at a grid resolution of 0.01° x 0.01° (roughly 1 km x 1 km), while the typical distance between official stations is in the order of 0.5° (roughly 50 km). The current hypothesis is that the use of additional lower-fidelity measurement stations in combination with high-resolution covariates can assist in this downscaling.

[1] Although the bias is technically part of the likelihood, in the analysis it is treated similarly to the covariates. The difference is that, while the covariates and the bias are both stacked into $\mathbf{M}(\mathbf{x'})$ in the analysis step, the bias is represented by columns of zeros in M(x) in the prediction step. In other words, we do want to subtract the bias and covariates from the observed data during analysis, but - where we do add the covariates back during prediction - we do not want to add the bias again during the prediction.

Therefore, we propose the following improvements to the methodology presented in Equations (1-6): In Section 3.2, we illustrate how we can use multiple proxies in order to learn, from the data, a model for bias and noise. It turns out that we can effectively deal with bias and noise that depends on the fidelity level. In Section 3.3, we illustrate how the use of high-resolution drift covariates as well as high-resolution noise proxies to further improve downscaling. In Section 3.4, we focus on improving the reliability of the posterior variance var[f(x)].

2.2 Relation to Multi-Fidelity Surrogate-Based Vehicle Optimisation

As can be distilled from Equations (1-6), our methodology is closely related to methods used in multi-fidelity surrogate-based vehicle performance optimisation (see for example: Forrester et al, 2007; Diez et al, 2014; Volpi, 2015; Baar, de, et al, 2015; Baar, de, et al, 2017; Bhattrai et al, 2018; Straten, van, et al, 2019; Wackers et al, 2022). The main difference is that in vehicle optimisation, the 'difference function' between fidelity levels is often considered to be a relatively smooth function of the design variables $\mathbf{x}$, while in the case of multi-fidelity spatial regression of meteorological data this assumption is not valid. Rather, for meteorological data, the difference function can be thought of as characterized by different levels of bias and noise, which in the advanced case can be modelled on (high-resolution and/or non-spatial) proxies. Despite this difference, issues like treatment of bias and noise, as well as improving the reliability of the posterior variance, arise in both fields of application.

3.0 IMPROVEMENTS TO MULTI-FIDELITY SPATIAL REGRESSION

3.1 Quality Control for Crowd-Sourced Meteorological Data

The 3PD air temperature observations coming from the WOW-NL network have undergone a quality control process. This process is a modification of (Napoly et al., 2018) that includes filters in two categories: mechanistic and statistic. The mechanistic filters check whether the station metadata is correct and assess whether a given station provides a sufficient daily and monthly coverage. The statistical checks first apply an elevation-based temperature correction, then check whether each observation is an outlier compared to a robust estimator, and finally calculates the Pearson correlation between each measurement and the hourly median. The quality control then attributes a quality label (i.e., lowest: M0, highest: M4) to each observation. In this study, we used only observations with the highest quality level.

3.2 MF Spatial Regression with Multiple Error Proxies

The treatment of noise in GPR has already been suggested, see for example (Forrester, 2007). However, often a single noise level is used. For some applications, the noise level can be different for different locations, and is based on a noise proxy (Baar, de, et al, 2014). Our hypothesis is that it can help us regress climate data if we expand this idea to use multiple proxies for the bias and noise in the data.

In this section, we focus on the improvements we can get if we model the bias and noise of our observations on proxies. The proposed models are

$$\mathbf{b} = c_{b1}\,\boldsymbol{\pi}_{b1} + c_{b2}\,\boldsymbol{\pi}_{b2} + \ldots + c_{bn}\,\boldsymbol{\pi}_{bn}, \tag{7}$$

$$\boldsymbol{\varepsilon}^2 = c^2_{\varepsilon 1}\,\boldsymbol{\pi}^2_{\varepsilon 1} + c^2_{\varepsilon 2}\,\boldsymbol{\pi}^2_{\varepsilon 2} + \ldots + c^2_{\varepsilon n}\,\boldsymbol{\pi}^2_{\varepsilon n}, \tag{8}$$

where the coefficients $\mathbf{c}$ remain to be estimated through MLE. Note that the bias $\mathbf{b}$ and noise $\boldsymbol{\varepsilon}$ can have distinct values for each measurement station, and, because the bias and noise are generally unknown, we assume that they can be modelled on a number of proxies $\boldsymbol{\pi}$. In the present section, the proxies contain zeros and ones, as an indication of the fidelity level. This results in a model where each fidelity level can have their own bias and noise level. However, in general it is also possible to build more complex models for bias and noise.

3.2.1 Mapping Simulated Wind Speed: Results and Discussion for Rastrigin Test Function

We test this approach for a modified rotated Rastrigin test function, which is one of the test functions that was selected for use in the present workshop (Mainini et al, 2022):

$$f_{rast}(x) = x_1^2 + x_2^2 + 1 - \cos(2\,\pi\,N_{osc}\,x_1) - \cos(2\,\pi\,N_{osc}\,x_2) + 6. \quad (9)$$

After a shift and rotation of the coordinates, the function has a global minimum that is centred in the middle of the Netherlands, as well as local minima and maxima. In this test function, we allow for local weather patterns, or 'regionalization', by variation of N_{osc}, the number of oscillations per degree. An illustration of the modified Rastrigin test function is shown in Figure 1, for increasing regionalization. As our $\mathbf{x'}$, we consider the actual locations of high-fidelity (1PD), medium-fidelity (2PD), and low-fidelity (3PD) stations, and to 2PD and 3PD we add an amount of bias and pseudo-random Gaussian noise.

We measure the performance of the spatial regression in two ways. Firstly, in order to quantify the accuracy of E[f(x)], we consider the prediction root mean squared error (RMSE) over the domain. Secondly, in order to quantify the reliability of var[f(x)], we consider rank histograms. In a rank histogram, we create ensemble members, and bin the prediction error by member (Hargreaves, 2010). This is illustrated for two different cases in Figure 2. Ideally, the rank histogram should be as flat as possible. Therefore, we compute the rank histogram standard deviation, and divide it by the worst possible value (i.e., all errors in the same bin). As such, we arrive at a performance measure between zero and one.

In our experiments we use a pseudo-random value for the Rastrigin rotation, the random part of the observation error, and – to avoid oscillation-to-grid interference effects – the 10^4 random locations at which we make the GPR prediction. We then repeat each experiment 20 times, each time resampling these random variables. In the figures we show the 25%, 50% and 75% percentile of the observed performance.

Figure 3(a) shows, for different methods, the RMSE for increasing regionality. The main point here is to illustrate that adding lower-fidelity stations can increase the attainable resolution, but only if we treat the noise in the data properly. Indeed, going from 1PD to 123PD improves performance, such that we can deal with higher regionality. However, it is important to model the noise in 23PD, since, when we model 'no noise', the RMSE quickly increases with increasing regionality (due to over-fitting, the RMSE explodes when directly using 23PD data – this effect might be less severe if we decrease the lower bound for the length scale hyperparameter estimation). The best performance is achieved when we model separate noise levels for 2PD and 3PD ('multi noise', as opposed to 'single noise' which assumes the same noise level for 23PD). It should be emphasized that we do not provide the actual noise levels to the GPR – instead, the noise levels are estimated through MLE. Figure 3(b) shows the corresponding rank histogram standard deviations.

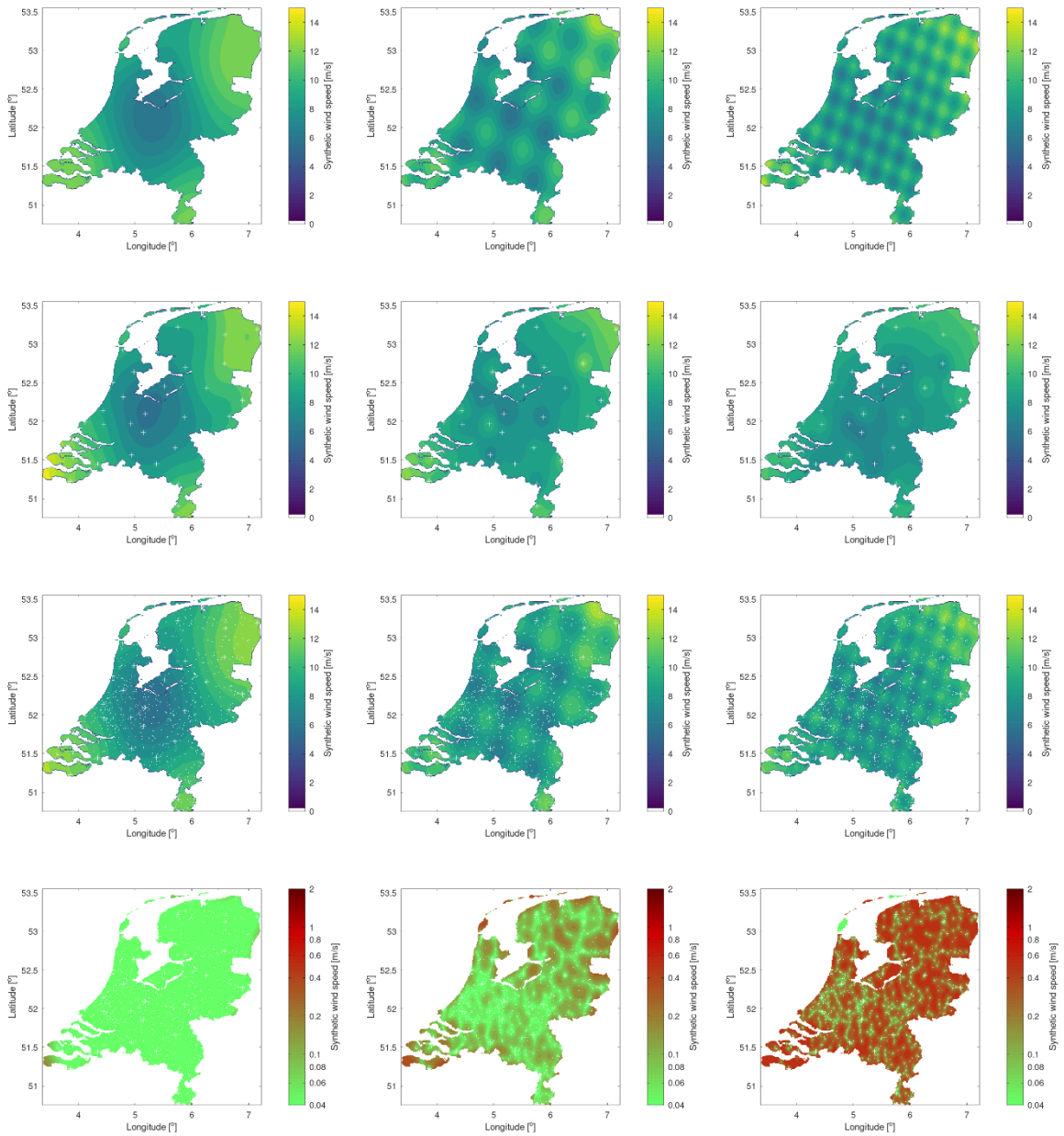
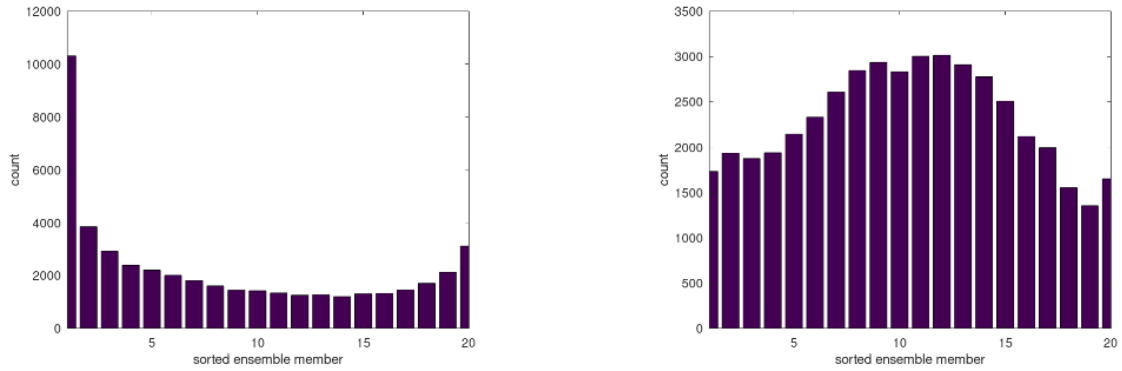
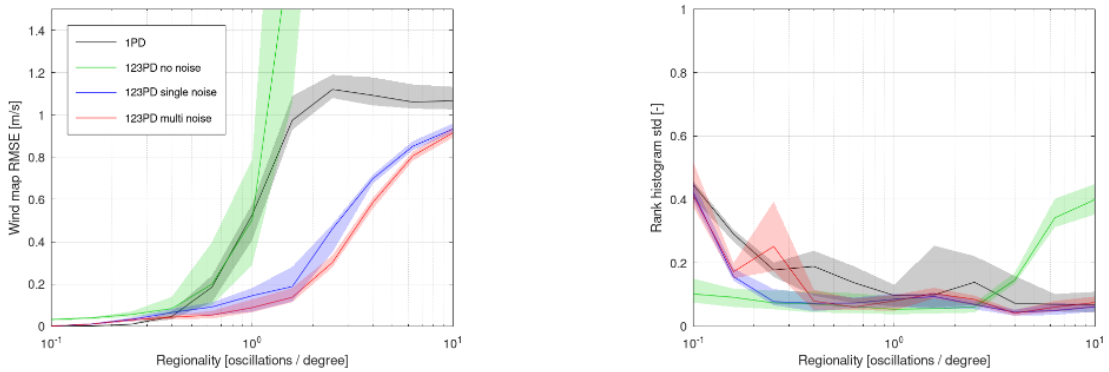


Figure 1: Illustration of the modified Rastrigin test function, used to provide synthetic wind speed data over the Netherlands. Left column for low regionality (N = 0.5 oscillations/degree), middle column for medium regionality (N = 1.5 osc/deg), and right column for high regionality (N = 3.0 osc/deg). Row one shows true function, row two shows regression mean based on official 1PD stations only, row three shows regression mean based on all 123PD stations, and row four shows regression uncertainty based on all 123PD stations.



(a) (b)

Figure 2: Illustration of rank histograms for high regionality for (a) 1PD and (b) 123PD.



(a) (b)

Figure 3: Quantified performance for increasing regionality, with (a) showing the map RMSE and (b) showing the map rank histogram standard deviation for different data and regression methods.

3.2.2 Mapping Real-World Air Temperature: Results and Discussion for Meteorological Data

In (Beekvelt, van, 2022), the same approach has been used for real world temperature data. Figure 4 illustrates that, when comparing 123PD results to 1PD results, we achieve more detail in our maps, as well as a lower predicted uncertainty. The 1PD, 2PD, and 3PD results are indicated as HF, MF, and LF, respectively.

3.3 MF Spatial Regression with High-Resolution Covariates and Error Proxies

The next step is to introduce high-resolution geospatial layers in the regression process. As an example, Figure 5(a-b) show maps of population density[2] (Central Bureau of Statistics, CBS) and percentage of land area covered by trees[3] (Dutch National Institute for Public Health and the Environment, RIVM). However, from Figures 6(c-e) it can be observed that the distribution of only 1PD stations over these covariates is quite limited, with a possibility of leading to over-fitting. The inclusion of 2PD and 3PD stations (locations as of 2020) has the potential of improving this distribution of stations over the range of covariates, and therefore possibly contributing to a reduction of over-fitting effects.

[2] CBS '100m vierkant statistiek' (100m square statistics): https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken

[3] Atlas Natuurlijk Kapitaal (Atlast of the Natural Capital): https://www.rivm.nl/atlas-natuurlijk-kapitaal

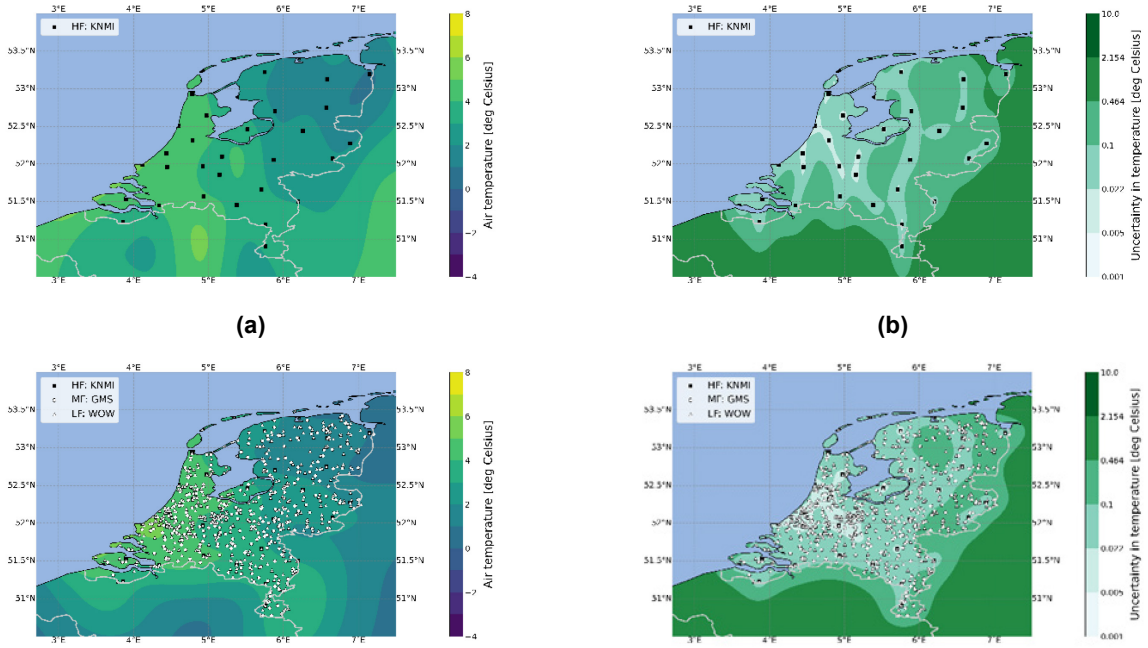


Figure 4: Daily mean temperature for 25 January 2019, with (a) regression mean based on 1PD, (b) regression uncertainty based on 1PD, (c) regression mean based on 123PD, and (d) regression uncertainty based on 123PD. Maps reproduced from (Beekvelt, van, 2022).

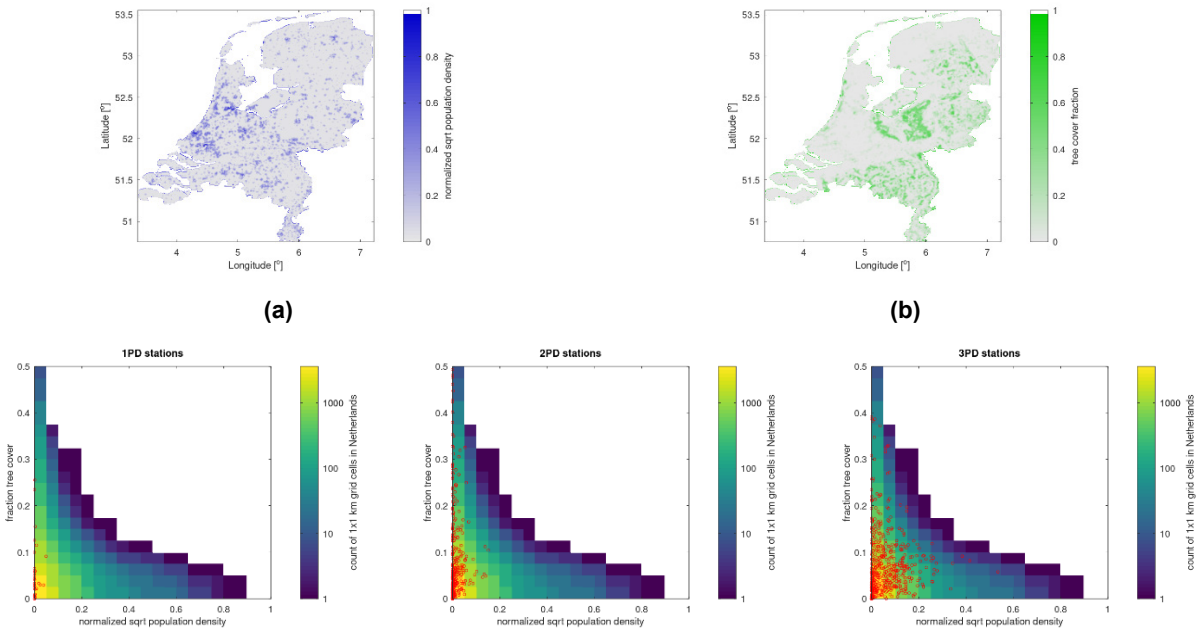


Figure 5: Example high-resolution covariates, showing (a) normalized square root of population density and (b) percentage of tree cover. In (c-e), we show, three times, the same two-dimensional histogram of the distribution of these covariates of the Netherlands. However, it can be seen that (c) 1PD stations cover only a small part of the range of these covariates, while (d) 2PD and (e) 3PD stations cover the covariates more extensively.

3.3.1 Mapping Simulated Wind Speed: Results and Discussion for Rastrigin Test Function

In this section, we consider the test function

$$f_{rastcov}(x) = f_{rast}(x) \quad - 2\ transformedPopDensity - 6\ treeCoverFraction, \tag{10}$$

where the coefficients are arbitrary. In this case, we only consider 1PD and 2PD station locations. Again, we add bias and noise to the 2PD. Importantly, now we add more 2PD noise in areas with higher population density.

Through M, we then feed the same covariates and noise proxies (but not the actual constants) to the GPR. Figure 6 compares results for medium regionality, while Figure 7 compares the quantified performance for increasing regionality. Now, after the introduction of relevant covariates, we see a large improvement when using 12PD over 1PD. Note that in 'single noise' we have used a constant noise proxy, while in 'multi noise' we base the noise model on a constant and on population density.

Our main observation here is that there seems to be a 'synthesis' effect: the combination of covariate information and 2PD improves the prediction. However, it becomes even more important to treat the noise adequately – as a function of proxies – as we do in the 'multi-noise' case.

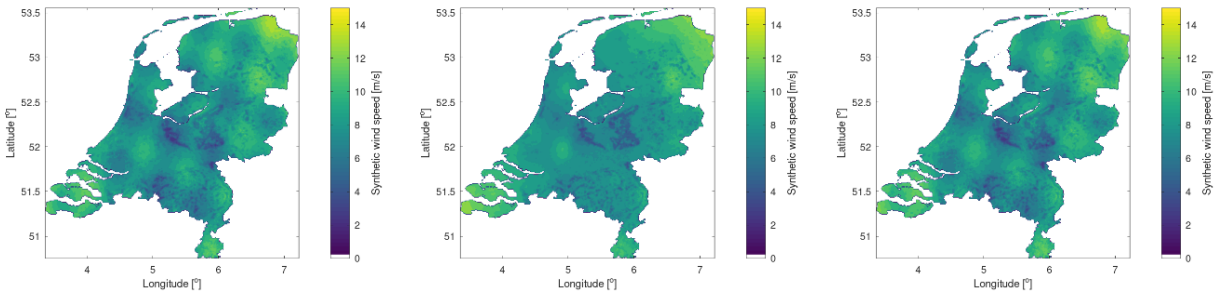
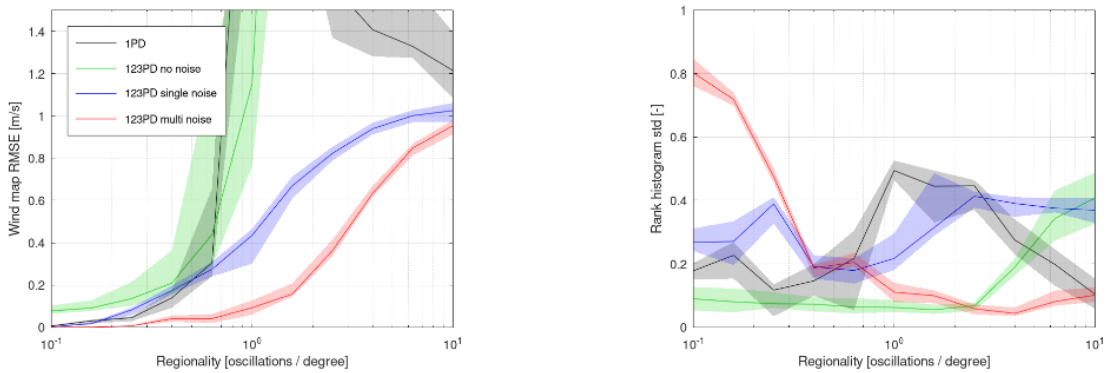


Figure 6: Spatial regression for medium regionality, using high-resolution covariates and noise proxy, showing (left) Rastrigin test function, (middle) regression mean based on 1PD, and (right) regression mean based on 12PD.



(a) **(b)**

Figure 7: Quantified performance for increasing regionality, with (a) showing the map RMSE and (b) showing the map rank histogram standard deviation for different data and regression methods.

In Figure 5(c-e) we have already seen that the 2PD and 2PD samples the covariates in a more equal way than the 1PD. As a result, we can observe an interaction effect, or 'synergy', between adding 23PD stations and covariates. As another test with the Rastrigin function, we have modified Equation (10) to include higher-order terms, resulting in five covariates. We can now study the effect of 'feeding' more 23PD stations and/or more covariates to the multi-fidelity analysis. The results of this test are illustrated in Figure 8. Without synergy, we would expect a result like Figure 8(a). In the case of anti-synergy, we would expect a result like Figure 8(b). However, in the case of synergy between the number of 23PD stations and the number of covariates, we would expect a result like Figure 8(c). From Figure 8(d), the actual result for the Rastrigin test-function, we observe synergy between the number of ingested 23PD stations and the number of ingested covariates.

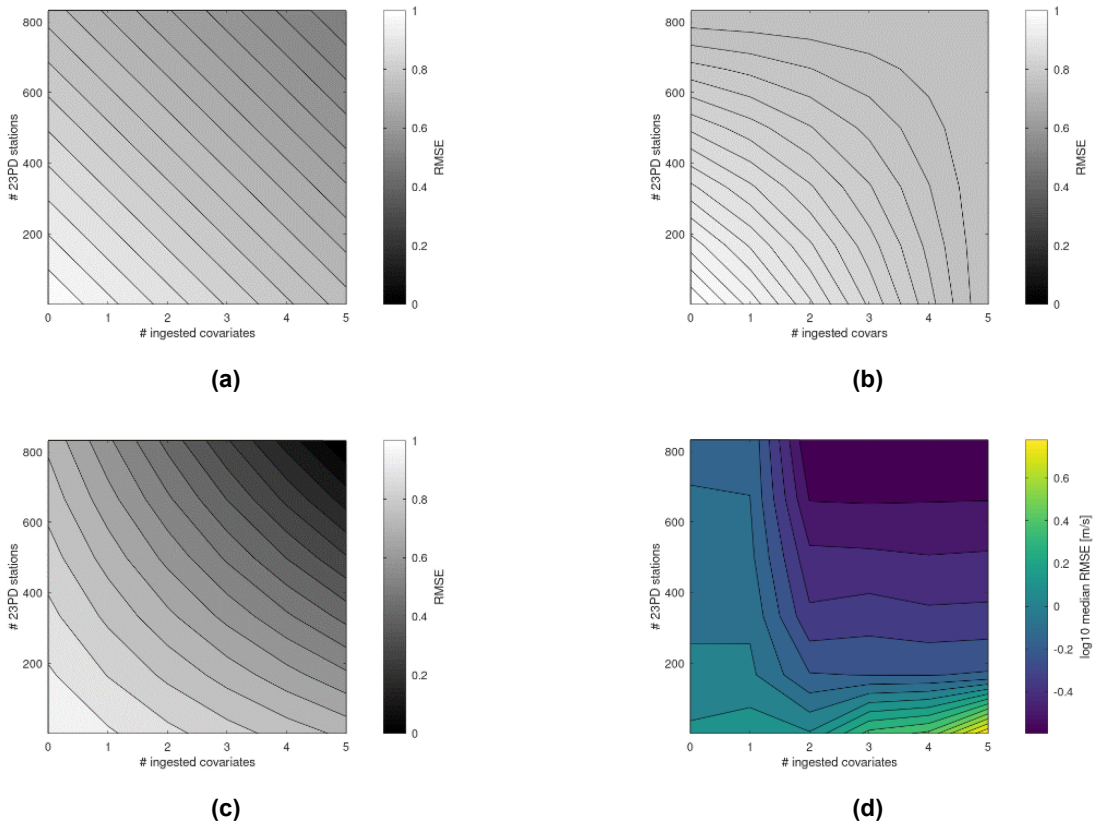


Figure 8: Hypothetical plots (a) without synergy, (b) with anti-synergy, and (c) with synergy between the number of ingested 23PD stations and the number of ingested covariates. The results for the Rastrigin test-function (d) show a significant synergy between these variables.

3.3.2 Mapping Real-World Temperature: Results and Discussion for Meteorological Data

As a preliminary study, we apply the same approach to real world data. In this case, we consider daily minimum and daily maximum temperatures, as observed by 1PD and 2PD networks on 25 January 2019. As covariates, we provide a second order basis of population density and tree cover to the GPR. Both population density and tree cover are thought to have an effect on local temperatures. However, this is only a preliminary study, and more (or more relevant) covariates might be considered in the future.

Figure 9 and Figure 10 show the results for daily minimum and maximum temperature, respectively. In both cases, going from (a) 1PD to (b) 2PD adds detail, and (d) introducing the covariates in 12PD adds even more local detail. However, introducing the covariates in 1PD (c) only seems to lead to over-fitting of the covariate model. This is an effect that can possibly also be seen in the sharp rise in Figure 7(a), and requires further investigation.

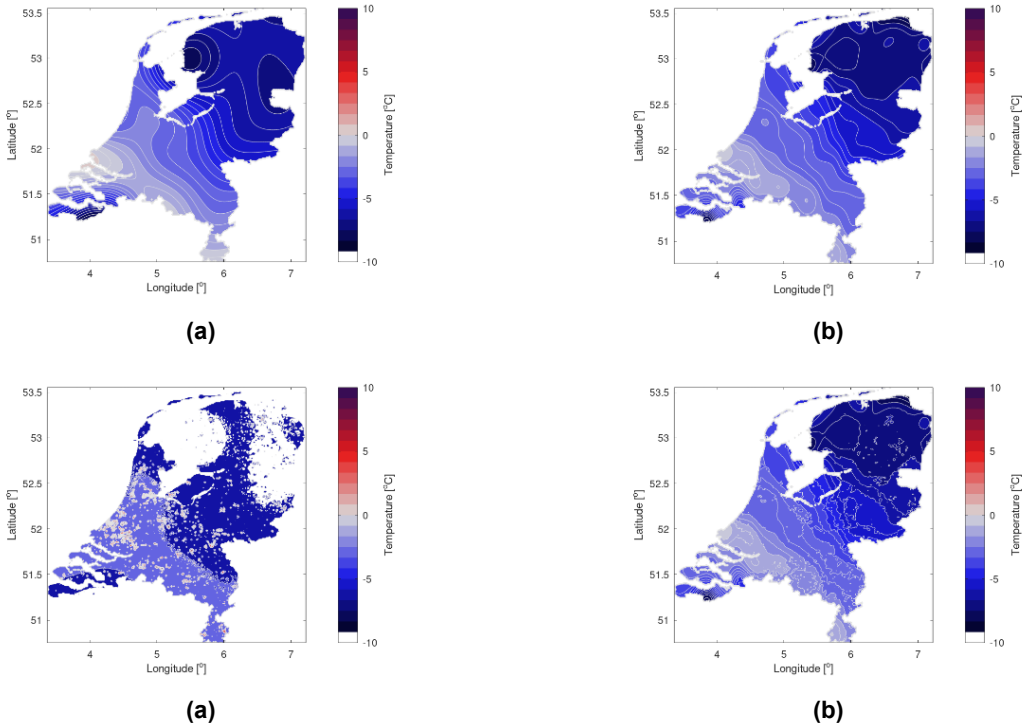


Figure 9: Daily maximum temperature on 25 January 2019, showing (a) 1PD without covariates, (b) 12PD without covariates, (c) 1PD with covariates, and (d) 12PD with covariates.

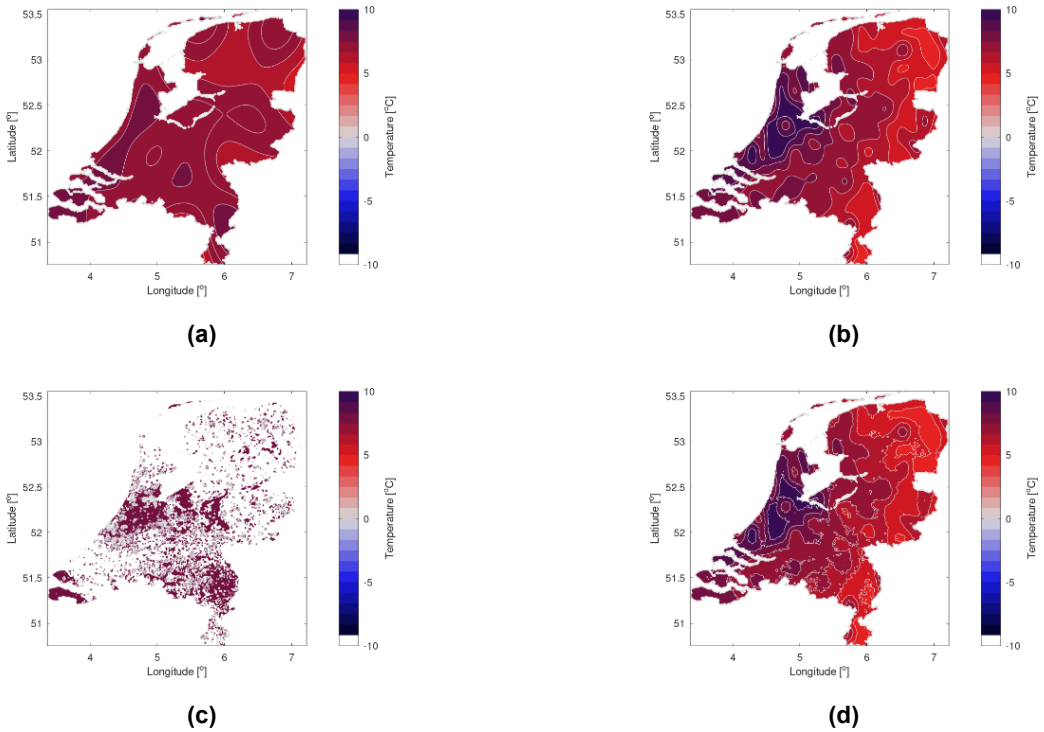


Figure 10: Daily maximum temperature on 25 January 2019, showing (a) 1PD without covariates, (b) 12PD without covariates, (c) 1PD with covariates, and (d) 12PD with covariates.

3.4 Improving the Reliability of the Posterior Variance Through Multi-Objective Hyperparameter Estimation

In the previous sections we have used MLE to optimize the hyperparameters, and have used prediction RMSE and rank histogram standard deviation to quantify the performance of E[f(x)] and var[f(x)], respectively. We have seen increased performance in terms of RMSE, thus E[f(x)], when introducing lower-fidelity data, especially in cases with high regionality (i.e., a large number of oscillations). However, we would also like to improve the performance of var[f(x)].

In an attempt to do so, we now switch to cross-validation based hyperparameter estimation (Witten et al, 2013) (in the multi-fidelity case, we only cross-validate by leave-one-out validation of the 1PD stations). We compare two approaches: In the first approach, we only target the cross-validation RMSE, and expect results that are close to the MLE optimisation. However, in the second approach, we use multi-objective minimization to target the RMSE as well as the rank histogram standard deviation.

In more detail, for the second, and new, objective, we divide the cross-validation-based GPR errors at the 1PD station locations by the individual cross-validation-based GPR uncertainty predictions. When building an experimental cumulative density function (CDF) of these relative errors, we would expect this CDF to be close to the Gaussian CDF. We quantify this by computing the RMS difference between the observed experimental CDF and the theoretical Gaussian CDF. This objective function can be seen to change for different settings of the hyperparameters.

As we now have two objective functions, the 'standard' RMSE and the RMS CDF difference, we can find a Pareto-optimal setting for our hyperparameters. Doing so, we hope to achieve low GPR prediction RMSE and low GPR prediction rank histogram standard deviation at the same time.

3.4.1 Mapping Simulated Wind Speed: Results and Discussion for Rastrigin Test Function

Here, we repeat the experiment of Section 3.2.1, but now for 1PD and 2PD stations only, the results are shown in Figure 11. With the multi-objective optimization, we observe similar performance for the prediction RMSE, and slightly more robust performance of the rank histogram standard deviation. However, the results are not as convincing as we expected, possibly due to the low number of 1PD stations used in the cross-validation. This requires additional investigation.

3.4.2 Mapping Real-World Temperature: Results and Discussion for Meteorological Data

For 1PD only, we have run a large-scale experiment for daily mean temperature in mixed terrain in and around Switzerland. We obtain the temperature data from ERA5-Land re-analysis simulations (Muñoz-Sabater et al, 2021). Figures 12(a-c) illustrate the terrain topology, a typical temperature map, and typical random station locations. In this case, altitude is provided to the GPR as a covariate for temperature.

Figure 12(d) shows an additional improvement that we made for this study, where in the kernel k(x,x) the distance is non-Euclidian: it receives a penalty if two locations are separated by complex terrain.[4] A typical resulting kernel is illustrated in this figure. It should be noted that this can lead to small negative eigenvalues in the covariance matrix K(**x'**,**x'**)+R, which, therefore, has to be corrected numerically.

Figure 12(e-f) show the quantified performance of the prediction RMSE and rank histogram standard deviation for January 2020. The RMSE is similar in both methods, while the higher scores of rank histogram standard

[4] In more detail: We compute two distances, or lags, between each pair of locations. The first lag is the standard Euclidian distance. For the second lag, we draw a straight line between the locations, and then compute the total travelled absolute vertical distance when moving from one location to the other. The total lag is the first lag plus a penalty factor times the second lag. The penalty factor is determined during hyperparameter estimation.

deviation that occur in single-objective hyperparameter estimation are significantly reduced when using the proposed multi-objective hyperparameter estimation in combination with the non-Euclidean kernels.

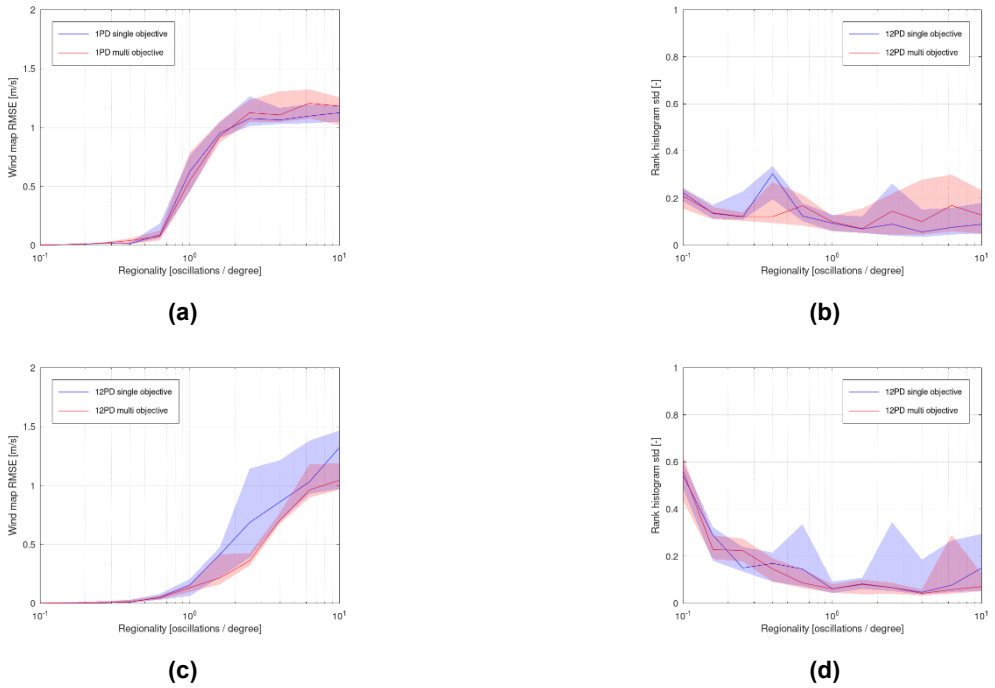


Figure 11: Quantified performance of multi-objective hyperparameter estimation for the Rastrigen test function, showing (a) map RMSE for 1PD, (b) rank histogram standard deviation for 1PD, (c) map RMSE for 12PD, and (d) rank histogram standard deviation for 12PD.

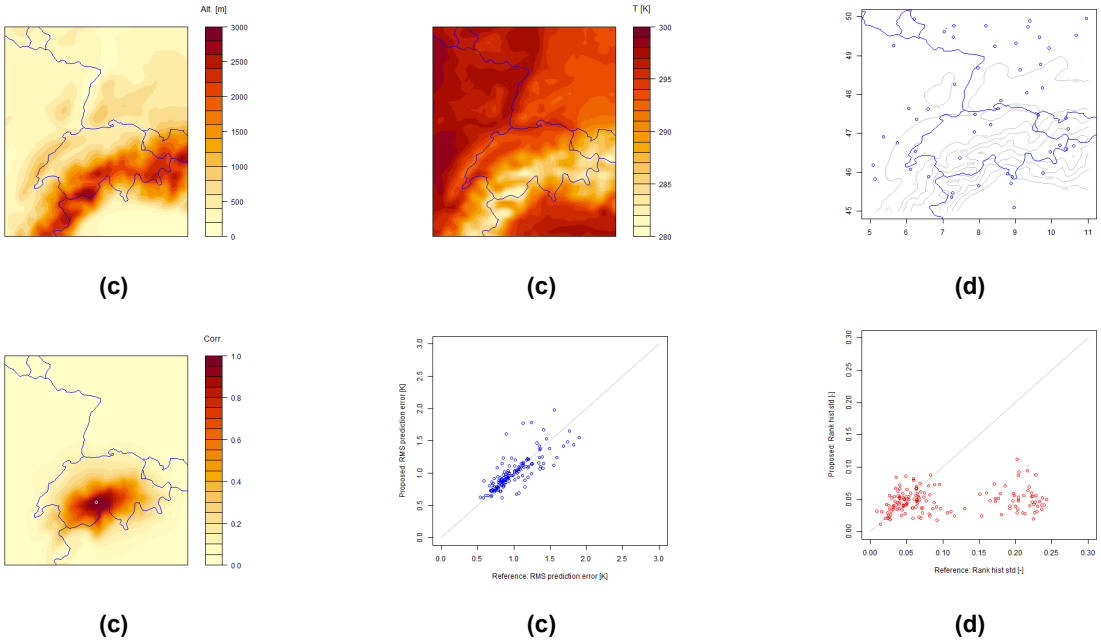


Figure 12: Map of temperature showing (a) terrain altitude, (b) example daily mean temperature from ERA5-Land reanalysis, (c) example station locations, (d) illustration of spatial correlation kernel using non-Euclidian distance, (e) map RMSE, and (f) map rank histogram standard deviation.

4.0 CONCLUSIONS

Over the recent years, within weather and climate observations and modelling, there is a growing interest in the availability and analysis of lower-fidelity and/or crowd-sourced data. How to merge such data of multi-fidelity sources into a single weather or climate service remains an active and ongoing research question. With respect to this development, on a technical level, we aim to answer the following questions:

1) Does the inclusion of lower-fidelity data help us to map weather and climate patterns of a higher spatial variability (i.e., regionality)? The results in Section 3.2 indicate that this is indeed the case. However, improvements due to the inclusion of lower-fidelity data can only be achieved when the bias and noise in the data are treated adequately.
2) Does the use of lower-fidelity data enable the inclusion of high-resolution covariate information to further improve prediction accuracy? The results in Section 3.3 indicate that there is a strong synergy between the inclusion of lower-fidelity data and high-resolution covariate information. Due to the small number of 1PD stations (i.e., concentrated in areas of low population, low vegetation cover, etc.), over-fitting of the covariate information becomes more likely. The inclusion of lower-fidelity stations strongly reduces this effect.
3) Can we improve the reliability of not only the GPR mean but also the GPR variance? Section 3.4 indicates that this might be possible, although the results are not completely convincing and indicate that further research is required.

We are convinced that these results can not only be interesting for the future treatment of 2PD and 3PD in meteorological applications, but also that these approaches and results are worth discussing in a broader research community in the generic context of multi-fidelity surrogate modelling.

As for climate and weather applications, we envision that the gradual inclusion of 3PD in the operational services of NMHSs will enable a plethora of new applications, especially in urban areas. It is particularly interesting to strive in the direction of urban applications, since severe weather conditions over communities make urban areas vulnerable. For example, heavy rainfall resulting in floods due to impervious surfaces (Berne et al., 2004), or increased urban temperatures due to the urban heat island effect (Theeuwes, 2017; Deilami et al., 2018), are hazardous conditions that could be monitored with 3PD at a finer spatial resolution than currently, thus enabling faster response times from decision makers. Therefore, even more *in situ* urban weather observations are desirable in the future. We hope that understanding the contribution of 2PD and 3PD will help cater to this need – or even motivate society to increase the amount of low-fidelity data collection in those areas most relevant for the analysis and prediction of weather- and climate impact.

ACKNOWLEDGEMENT

ERA5-Land data (Muñoz-Sabater et al, 2021) for Section 3.4.2 was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store.

5.0 REFERENCES

de Baar, J.H., Percin, M., Dwight, R.P., van Oudheusden, B.W. and Bijl, H., 2014. Kriging regression of PIV data using a local error estimate. Experiments in fluids, 55(1), pp.1-13.

de Baar, J., Roberts, S., Dwight, R. and Mallol, B., 2015. Uncertainty quantification for a sailing yacht hull, using multi-fidelity kriging. Computers & Fluids, 123, pp.185-201.

de Baar, J.H.S., Leylek, Z, Habib, A, Neely, A.J., and Ray, T, 2017. Multi-fidelity efficient global optimisation of the geometry of a transonic axial compressor, ISABE 2017, 3-8 September 2017, Manchester, UK.

Daniëlle van Beekvelt, 2022. Multi-fidelity spatial regression for air temperature predictions using first, second and third party data. MSc Thesis, Utrecht University, Department of Mathematics.

Berne, A., Delrieu, G., Creutin, J.D. and Obled, C., 2004. Temporal and spatial resolution of rainfall measurements required for urban hydrology. Journal of Hydrology, 299(3-4), pp.166-179.

Bhattrai, S., de Baar, J.H. and Neely, A.J., 2018. Efficient uncertainty quantification for a hypersonic trailing-edge flap, using gradient-enhanced kriging. Aerospace Science and Technology, 80, pp.261-268.

Christensen, Ronald, 2001. Advanced linear modeling : multivariate, time series, and spatial data; nonparametric regression and response surface maximization, Springer, New York.

Deilami, K., Kamruzzaman, M. and Liu, Y., 2018. Urban heat island effect: a systematic review of spatio-temporal factors, data, methods, and mitigation measures. International Journal of Applied Earth Observation and Geoinformation, 67 pp. 30-42).

Diez, M., He, W., Campana, E.F. and Stern, F., 2014. Uncertainty quantification of Delft catamaran resistance, sinkage and trim for variable Froude number and geometry using metamodels, quadrature and Karhunen-Loève expansion. Journal of Marine Science and Technology, 19(2), pp.143-169.

Ficini, S., Iemma, U., Pellegrini, R., Serani, A. and Diez, M., 2021. Assessing the performance of an adaptive multi-fidelity gaussian process with noisy training data: A statistical analysis. In AIAA AVIATION 2021 FORUM (p. 3098).

Forrester, A.I., Sóbester, A. and Keane, A.J., 2007. Multi-fidelity optimization via surrogate modelling. Proceedings of the royal society a: mathematical, physical and engineering sciences, 463(2088), pp.3251-3269.

Garcia-Marti, I.,Overeem, A., Noteboom, J.W., de Vos, L., de Haij,M., and Whan, K., 2022. From proof-of-concept to proof-of-value: Approaching third-party data to operational workflows of national meteorological services. International Journal of Climatology, pp.1-18.

Hargreaves, J.C., 2010. Skill and uncertainty in climate models. Wiley Interdisciplinary Reviews: Climate Change, 1(4), pp.556-564.

Mainini1, L., Serani, A., Rumpfkeil, M.P., Minisci, E., Quagliarella, D., Pehlivan, H., Yildiz, S., Ficini, S., Pellegrini, R., Di Fiore, F., Bryson, D., Nikbay, M., Diez, M., and Beran, P., 2022. Analytical Benchmark Problems for Multifidelity Optimization Methods, present workshop.

Mardia, K.V. and Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71(1), pp.135-146.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H. and Martens, B., 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), pp.4349-4383.

Napoly, A., Grassmann, T., Meier, F. and Fenner, D., 2018. Development and application of a statistically-based quality control for crowdsourced air temperature data. Frontiers in Earth Science, 6, pp. 1-16.

Van Straten, O.F., Celik, E., De Baar, J.H., Ascic, B. and De Jong, J.S., 2019. Improved hull design with potential-flow-based parametric computer experiments. In MARINE VIII: proceedings of the VIII International Conference on Computational Methods in Marine Engineering (pp. 321-333). CIMNE.

Theeuwes, N.E., Steeneveld, G.J., Ronda, R.J. and Holtslag, A.A., 2017. A diagnostic equation for the daily maximum urban heat island effect for cities in northwestern Europe. International Journal of Climatology, 37(1), pp.443-454.

Volpi, S., Diez, M., Gaul, N.J., Song, H., Iemma, U., Choi, K.K., Campana, E.F. and Stern, F., 2015. Development and validation of a dynamic metamodel based on stochastic radial basis functions and uncertainty quantification. Structural and Multidisciplinary Optimization, 51(2), pp.347-368.

Wackers, J, Pellegrini, R, Serani, A, Diez, M, Visonneau, M, 2022. Multi-fidelity Active Learning for Shape Optimization Problems Affected by Noise. ECCOMAS 2022.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning. New York: springer.

Weather Meteorological Organization (WMO), 2018. Guide to instruments and methods of observation. Volume III: observing systems. Available at: https://library.wmo.int/doc_num.php?explnum_id=9872