Report Geostatistics

Erlend Øydvin

December 22, 2021

1 Introduction

In this report we study how Ordinary Kriging can be used to interpolate precipitation measured as point values in the plane. First the data is presented and the problem is formulated. Then the theoretical framework, the empirical variogram and the Ordinary Kriging equations, are explained. Finally we apply the Kriging equations to the dataset and then discuss the results. The Kriging Equations have been implemented in Python and the code can be found in the following git-repository:
https://github.com/erlnwind/Geostatistics

1.1 Measuring Precipitation

Precipitation are often measured as spatial point measurements. This is traditionally done using for instance rain gauges. When we want to know how much it is raining over an area we then need to interpolate between the rain gauges. The challenge is to find a reasonable set of weights for the observations so that the interpolated values are optimal. As precipitation comes in many variations, some varying much in space and time, it is reasonable to weight the observations based on the variation that can bee seen in the data. For instance if the spatial variation in observed precipitation is high we should weight observations that lies close to the point we want to interpolate. If the spatial variation is low we can increase the weight of observations that lies further away. This way of looking at variations in the data also gives an indication on the uncertainty in the data. High variations in the observations indicate a high variance in the rainfall. It is possible to measure rainfall using a weather radar. A weather radar works by sending out an electromagnetic pulse and then measure the reflectivity. Reflectively is the amount of transmitted power returned to the radar receiver after hitting precipitation. The more precipitation, the more power is reflected back. Reflected power can thus be related to rain rate. In contrast to point observations, such as rain gauge, weather radar has the potential to cover an area. However, despite advances in radar technology, beam blockage and the vertical profile of reflectivity still remains a limiting factor [Chwala and Kunstmann, 2019]. Due to these limiting factors it can be hard to relate the precipitation measured by the radar to the precipitation that is measured on the ground. Conventional precipitation maps are thus created by calibrating measured radar reflectivity with rain gauges on the ground [Berne and Krajewski, 2013]. Point measurements on the ground are therefor still a very relevant contribution to rainfall maps.

1.2 Data

The data used in this report is point measurements from rain gauges from Norwegian Meteorological Institute. The data was downloaded using Norwegian Meteorological Institute's API Frost and the python package requests. In this report I have used data from 7 weather stations in the Oslo Region. The study area with weather stations can be seen in Figure 1. The data lists accumulated hourly precipitation for the 8th of August 2018 from 08:00 to 12:00 and the whole dataset is shown in Table 1. Here we can see that the measured precipitation in Ås (NMBU) and Rustadskogen is similar. This is reasonable as these observations lies close to each other. We can also see that its generally observed less precipitation in Kykkelsrud. This could be due to the fact that it lies a bit away from the fjord and thus in other meteorological conditions, however this dataset is not large enough to conclude anything.

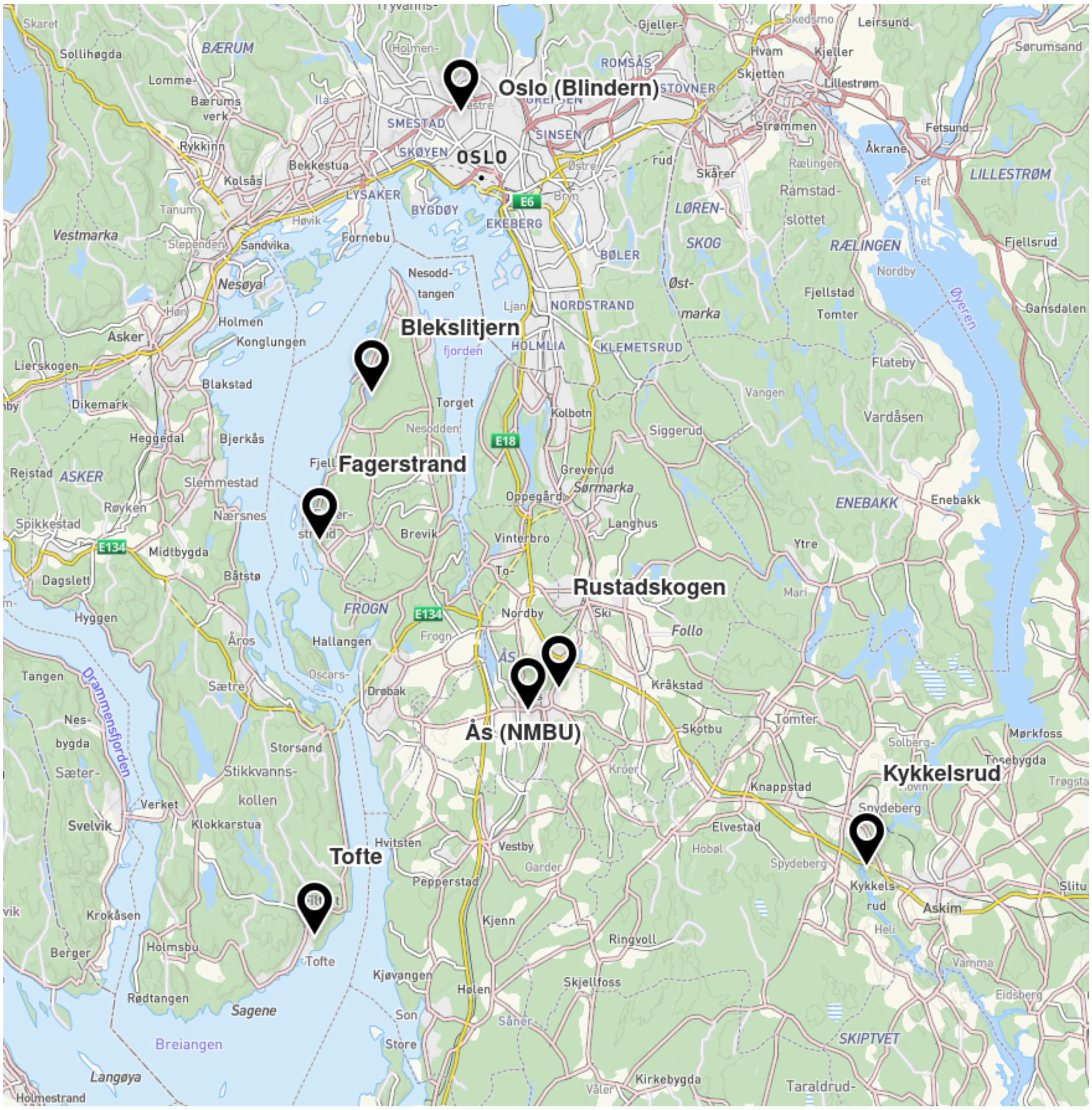

Figure 1: Map over the Oslo fjord area and weather stations used in this report.

Time/Stations	08:00	09:00	10:00	11:00	12:00
Kykkelsrud	0.6	1.1	0.2	0.1	0.9
Fagerstrand	0.3	0.6	2.1	2.	0.3
Blekslitjern	0.2	0.6	2.1	1.9	0.4
Ås (NMBU)	1.6	1.4	0.9	2.2	2.1
Rustadskogen	2.1	1.5	0.9	2.	1.
Oslo (Blindern)	0.2	0.8	1.5	3.	0.4
Tofte	1.6	0.9	1.7	4.3	1.1

Table 1: Sum hourly rainfall for weather stations shown in Figure 1 on the 8th of August 2018 from 08:00 to 12:00.

2 Methods

2.1 The Variogram

Central in the Ordinary Kriging method is to study the spatial variance in the observations. This spatial variance is modeled using the variogram. Points that lie close are more likely to be similar than points that lie far from each other and the distance where this similarity decreases depends on the spatial variance of the field. For instance, if the data shows that the rainfall varies much between stations, even for observations that lies close to each other, then the variogram shows that variance increases rapidly. Likewise, if the data shows that observations that are close to each other are more similar then the variogram shows that the variance increases more slowly. This spatial similarity is characterized by the semivariance function (also called semivariogram or variogram) and is given by

$$\gamma(\vec{u}) = \frac{1}{2} Var\left(Z(\vec{u}+\vec{h}) - z(\vec{u})\right) = \frac{1}{2} E\left(\left[Z(\vec{u}+\vec{h}) - z(\vec{u})\right]^2\right) \tag{1}$$

Here $\vec{u}$ is a point in space and time and $\vec{h}$ is a vector separating the two points. Equation 1 builds on what is called the intrinsic hypothesis and that the rainfall field is an isotropic process. The intrinsic hypothesis assumes that the mean of all $Z(\vec{u})$ is constant and that the difference between two points $Z(\vec{u}+\vec{h}) - Z(\vec{u})$ has zero mean and constant variance [Webster, 2001]. In other words this means that the spatial correlation and the semivariogram is the same for all positions in the study area $\vec{u}$. An isotropic process assumes that the spatial correlation is independent of direction, which is not always true for precipitation as wind decreases the variance in its direction.

Figure 2 shows a variogram. Here the x-axis represents h, the distance between two points, and the y-axis the corresponding variance γ between these two points. We can observe that γ increases as h increases until it converges towards a value called the sill. The sill represents the study area variance. The value of $\gamma(0)$ is called the nugget. The nugget have two different interpretations as it can represent uncertainty in finding precipitation right next to the weather station and it can also represent uncertainty in the observation itself. The semivariance function for our study field can be found by comparing the spatial distance between the stations with the squared difference (semivariance) between them as follows

$$\gamma(h) = \frac{1}{2}(y_i - y_j)^2 \tag{2}$$

Here h is the distance between the measurements i and j, y_i and y_j is the value of measurement i and j respectively. Plotting $\gamma(h)$ as a function of h then generates the red dots in Figure 2. The red dots generally have small values close to $h = 0$ and spreads more out as h increases. The realizations are usually divided into averaging bins as follows

$$\gamma * (h) = \frac{1}{2N(h)} \sum_{i,j \in R(h)}^{N(h)} (Z(u_i) - Z(u_j))^2 \tag{3}$$

where $N(h)$ is the number of observations in each bin, and $R(h)$ is given by

$$R(h) = \Delta h_k \leq |u_i - u_j| \leq \Delta h_{k+1} \tag{4}$$

where $i, j = 1, 2..$ N(h), Δh_k is a separation interval and k = 0, 1, ..., h_{max} is the number of separation interval classes. h_{max} is maximum separation distance. After the semivariance for the observations and the different distances has been found the next step is to find a empirical model for $\gamma(h)$ that fits these observations. There exists several models that can be used to fit the observations, including the spherical and the linear model [Deutsch and Journel, 1997]. In this project I have implemented the general exponential function given by

$$\gamma_z(h) = C_0 + C_1\left(1 - e^{\alpha\left(\frac{h}{h_R}\right)^{\beta}}\right) \tag{5}$$

where C_0 represents the nugget value, $C_0 + C_1$ represents the sill, h is the spatial distance between measurements, h_R is the range and α is the practical range (usually set to -3) [Rasmussen and Williams, 2006].

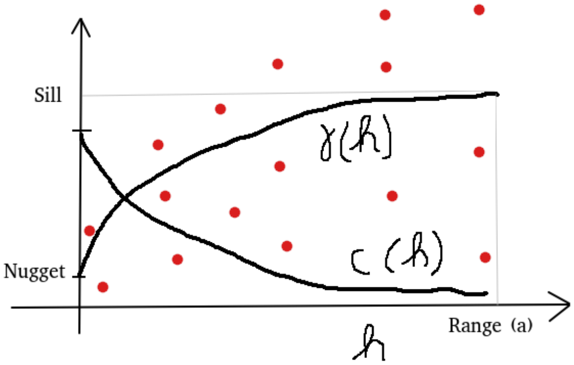

Figure 2: Semivariance function $\gamma(h)$ together with its corresponding covariance function $C(h)$. The red dots illustrates the semivariance found by Equation 5.

β defines the shape of the function and if set to 1 the model is called exponential and if set to 2 the model is called Gaussian. When fitting Equation 5 to oberserved data C_0 can be chosen or set as a free variabel whereas C_1 is usually found through some curve fitting algorithm. The covariance $C(h)$ and semivariance $\gamma(h)$ are related by

$$Cov_z(h) = C_0 + C_1 - \gamma_z(h) \tag{6}$$

$$= C_1 e^{\alpha\left(\frac{h}{h_R}^{\beta}\right)}. \tag{7}$$

2.2 Deriving the Ordinary Kriging Equations

The Ordinary Kriging method seeks to find the optimal value for a point in the plane where we do not have any measurements based on the measurements that lies around it. This amounts to finding the optimal weights λ for the observations $Z(u_i)$ and taking the sum so that

$$\hat{Z}(u_0) = \sum \lambda_i Z(u_i), i = 1, 2..., n. \tag{8}$$

Here $\hat{Z}(u_0)$ is the estimated value in point u_0, λ_i is the weight for observation i and $Z(u_i)$ is observation i. The goal is thus to find the optimal set of weights λ_i. Next we require this model $\hat{Z}(u_0)$ to be identical to the observations in the points where we have done observations so that

$$E\left(\hat{Z}(u_0) - Z(u_i)\right) = 0, i = 1, 2..., n \tag{9}$$

By inserting Equation 8 into Equation 10 we get

$$E\left(\sum \lambda_i Z(u_i) - Z(u_i)\right) = 0, i = 1, 2..., n \tag{10}$$

which can be written as

$$\sum \lambda_i E\left(Z(u_i)\right) = E\left(Z(u_i)\right), i = 1, 2..., n \tag{11}$$

which implies that

$$\sum \lambda_i = 1, i = 1, 2..., n \tag{12}$$

which states the sum of the weights should be equal to 1. Next we seek a model that minimizes the variance between the predictions from the variogram and our to be found model under the constraints from Equation 12. We thus define the following cost-function

$$C(\lambda) = \frac{1}{2} Var\left(Z(u_n) - \hat{Z}(u_n)\right) + \mu_L\left(\sum \lambda_i - 1\right) \tag{13}$$

where $Z(u_n)$ is the value estimated by the variogram, $\hat{Z}(u_n)$ is the prediction from the model, μ_L is the lagrange multiplier and $\sum \lambda_i - 1$ is the contraint from Equation 12. If we assume that the data in the study area is a product of an intrinsic and isotropic process Equation 1 holds and we can rewrite Equation 13 into

$$C(\lambda) = \frac{1}{2}E\left(\left(Z(u_n) - \hat{Z}(u_n)\right)^2\right) + \mu_L\left(\sum \lambda_i - 1\right) \tag{14}$$

By inserting Equation 8 into Equation 13 we get

$$C(\lambda) = \frac{1}{2}E\left(\left(Z(u_n) - \sum \lambda_i Z(u_i)\right)^2\right) + \mu_L\left(\sum \lambda_i - 1\right) \tag{15}$$

We can minimize the cost-function $C(\lambda)$ by taking the derivative with respect each λ_i as follows

$$\frac{\partial C}{\partial \lambda_i} = \frac{1}{2}E\left(\frac{\partial}{\partial \lambda_i}\left(Z(u_n) - \sum \lambda_i Z(u_i)\right)^2\right) + \frac{\partial}{\partial \lambda_i}\mu_L\left(\sum \lambda_i - 1\right) = 0 \tag{16}$$

The derivative of λ with respect to λ_i is just 1 for all i. Then by the chain rule Equation 16 can be written as

$$\frac{\partial C}{\partial \lambda_i} = E\left(\left(Z(u_n) - \sum \lambda_i Z(u_i)\right)\frac{\partial}{\partial \lambda_i}\left(Z(u_n) - \sum \lambda_i Z(u_i)\right)\right) + \mu_L = 0 \tag{17}$$

By taking the sum and the derivative only the ith λ_i contributes to this cost function and we get

$$\frac{\partial C}{\partial \lambda_i} = E\left(\left(Z(u_n) - \lambda_1 Z(u_1) - \lambda_2 Z(u_2) - ... - \lambda_n Z(u_n)\right)\left(-Z(u_i)\right)\right) + \mu_L = 0 \tag{18}$$

Which can be written as

$$\frac{\partial C}{\partial \lambda_i} = E\left(\left(Z(u_n) - \lambda_1 Z(u_1) - \lambda_2 Z(u_2) - ... - \lambda_m Z(u_m)\right)\left(-Z(u_i)\right)\right) + \mu_L = 0 \tag{19}$$

Note also that

$$\frac{\partial C}{\partial \mu_L} = \lambda_1 + \lambda_2 + ... + \lambda_m = 1 \tag{20}$$

Equation 19 and Equation 20 form a system of equations that can be written as

$$\begin{bmatrix} E(Z(u_1)Z(u_1)) & E(Z(u_1)Z(u_2)) & \dots & E(Z(u_1)Z(u_m)) & 1 \\ E(Z(u_2)Z(u_1)) & E(Z(u_2)Z(u_2)) & \dots & E(Z(u_2)Z(u_m)) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ E(Z(u_m)Z(u_1)) & E(Z(u_m)Z(u_2)) & \dots & E(Z(u_m)Z(u_m)) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \\ \mu_L \end{bmatrix} = \begin{bmatrix} E(Z(u_n)Z(u_1)) \\ E(Z(u_n)Z(u_2)) \\ \vdots \\ E(Z(u_n)Z(u_m)) \\ 1 \end{bmatrix} \tag{21}$$

In Appendix A it is show that

$$\gamma(u_1, u_2) = {\mu_z}^2 - E\left[Z(u_1)Z(u_2)\right] \tag{22}$$

Where $\gamma(u_1, u_2)$ is the semivariance function, ${\mu_z}^2$ is the variance and $E\left[Z(u_1)Z(u_2)\right]$ is the expected occurrence of $Z(u_1)$ and $Z(u_2)$. By inserting Equation 22 into Equation 21 the minus sign and the ${\mu_z}^2$ cancels out on both sides and we get

$$\begin{bmatrix} \gamma(u_1, u_1) & \gamma(u_1, u_2) & \dots & \gamma(u_1, u_m) & 1 \\ \gamma(u_2, u_1) & \gamma(u_2, u_2) & \dots & \gamma(u_2, u_m) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ \gamma(u_m, u_1) & \gamma(u_m, u_2) & \dots & \gamma(u_m, u_m) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \\ \mu_L \end{bmatrix} = \begin{bmatrix} \gamma(u_1, u_n) \\ \gamma(u_2, u_n) \\ \vdots \\ \gamma(u_m, u_n) \\ 1 \end{bmatrix} \tag{23}$$

Another way to express this system is in terms of co-variance. By inserting Equation 6 into Equation 23 we get that

$$\begin{bmatrix} cov(u_1, u_1) & cov(u_1, u_2) & \dots & cov(u_1, u_m) & 1 \\ cov(u_2, u_1) & cov(u_2, u_2) & \dots & cov(u_2, u_m) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ cov(u_m, u_1) & cov(u_m, u_2) & \dots & cov(u_m, u_m) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \\ \mu_L \end{bmatrix} = \begin{bmatrix} cov(u_1, u_n) \\ cov(u_2, u_n) \\ \vdots \\ cov(u_m, u_n) \\ 1 \end{bmatrix} \tag{24}$$

Equation 24 can be written in matrix notation as

$$\boldsymbol{cov}\ \boldsymbol{X} = \boldsymbol{cov_{eo}} \tag{25}$$

It can be shown that this system is positive definite. We then find the weights X by left-multiplying by the inverse of the covariance matrix and get that

$$\boldsymbol{X} = \boldsymbol{cov}^{-1}\boldsymbol{cov_{eo}} \tag{26}$$

It can be shown that by inserting Equation 26 into Equation 13 that the error, or the Kriging Variance, can be expressed as

$${\sigma_k}^2 = Cov(0) - \boldsymbol{X}^T\boldsymbol{cov_{eo}} \tag{27}$$

This is shown more in detail in the lecture notes [Kitterød, 2021].

2.3 Implementing code and testing

Tho Kriging method was implemented as a python module. It was tested and shown to generate the same results as other python implementations of ordinary Kriging, in this case it was tested against the python package pykrige.

3 Results

The experimental variogram was found using 3 averaging bins. α was set to -3 as is a common approximation. The range h_r was set to 0,5 as this seems to be a good approximation for the variogram. C_0 was set to 0 as this seems like a reasonable guess for most of the observations. Both the Exponential ($\beta = 1$) and the Gaussian model ($\beta = 2$) given by Equation 3 was tested. Then the variogram was found using the module curve-fit from the python package scipy.optimize. Next the estimated variograms was used together with the Ordinary Kriging Equations given by Equation 24 to calculate interpolated values for our study area.

Interpolated precipitation together with the Kriging Variance for the Gaussian model is shown in Figure 3. Here we can see that precipitation generally is low from 08:00 to 10:00, then increases around 11:00 and decreases for most locations again 12:00. The Gaussian model seems to be able to predict higher precipitation outside observations. The Kriging variance shows that the uncertainty is generally low close and in between the observations, creating a trail between the observations, and it increases when the difference between observations is large. Most of the Gaussian variograms seems to fit the data very good except the variogram for 12:00 where observations that lies far from each other are similar. Interpolated precipitation with the Kriging Variance for the Exponential model is shown in Figure 4. Here we can also see that precipitation is low from 08:00 to 10:00, increases at 11:00 and decreases again at 12:00. Generally the exponential model seems to generate a more smooth surface than the Gaussian, with the maximum values centered at the observations. The Kriging Variance is for the exponential model also more centered around the stations and it seems to increase when there are variations among stations. Visually the exponential variograms seems to fit the data, but the Gaussian seems to fit better at least for the events at 08:00, 09:00 and 11:00.

4 Discussion and conclusion

A Gaussian an a Exponential semivariogram was fit to the observations and Ordnary Kriging was performed to interpolate precipitation for the study area. Generally the methods seems able to produce results that agree with the input data, little rain from 08:00 to 10:00, then more rain at 11:00 and then less rain at 12:00.

The methods also shows that the Kriging variance is low close to the observations and then generally higher further away and higher for events with more spread in data. The Gaussian method also seems to predict lower variance between the observations, implying less uncertainty in the predictions. The Gaussian model seems able to generate interpolated maximum and minimum that are not exactly

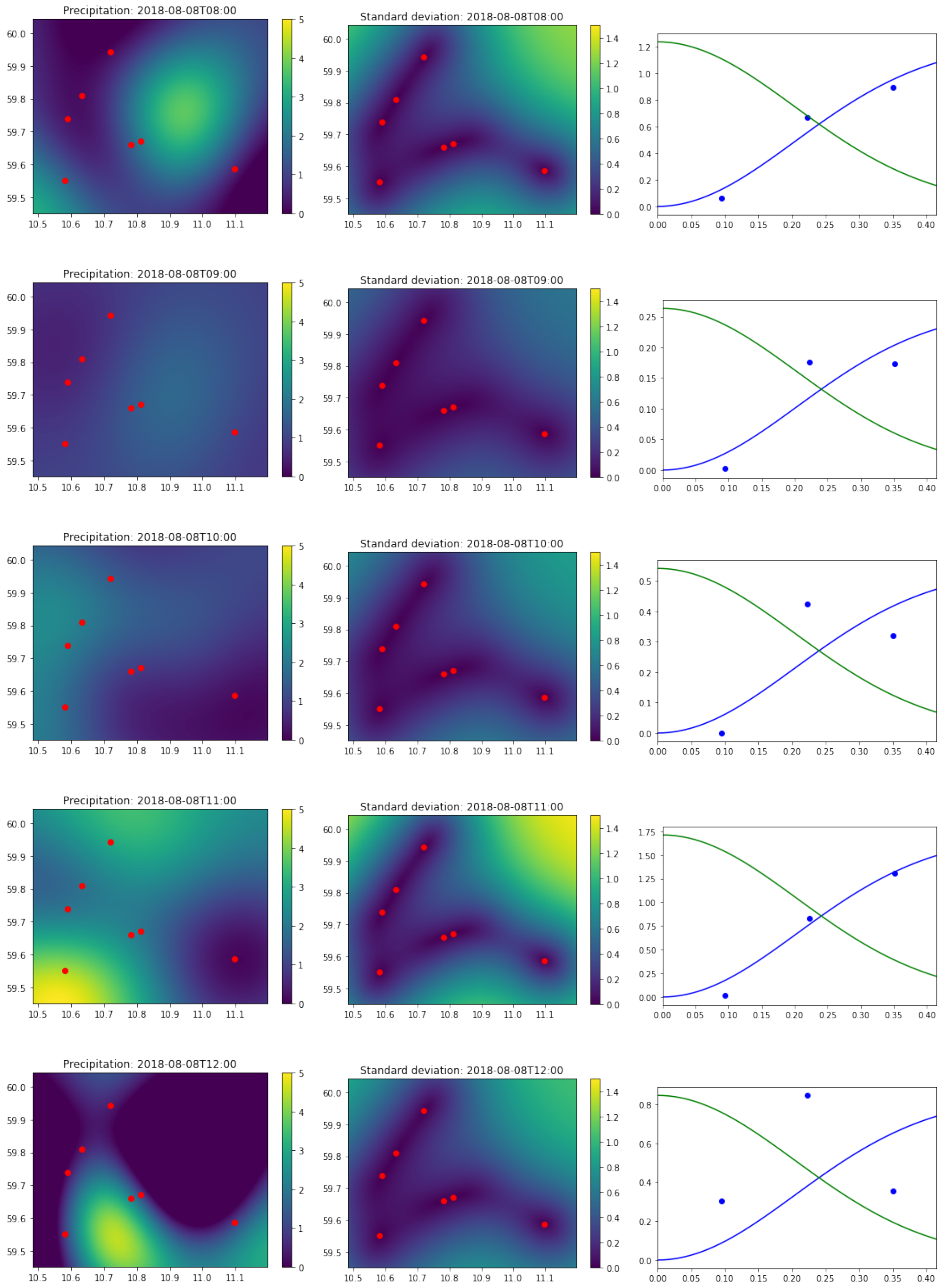

Figure 3: Gaussian variogram: Left column: Kriging interpolation of precipitation. Middle column: Kriging variance for interpolated values. Right column: Semivariogram. The red dots show locations with observations, the blue dots shows the bins average semivariance given by Equation 3.

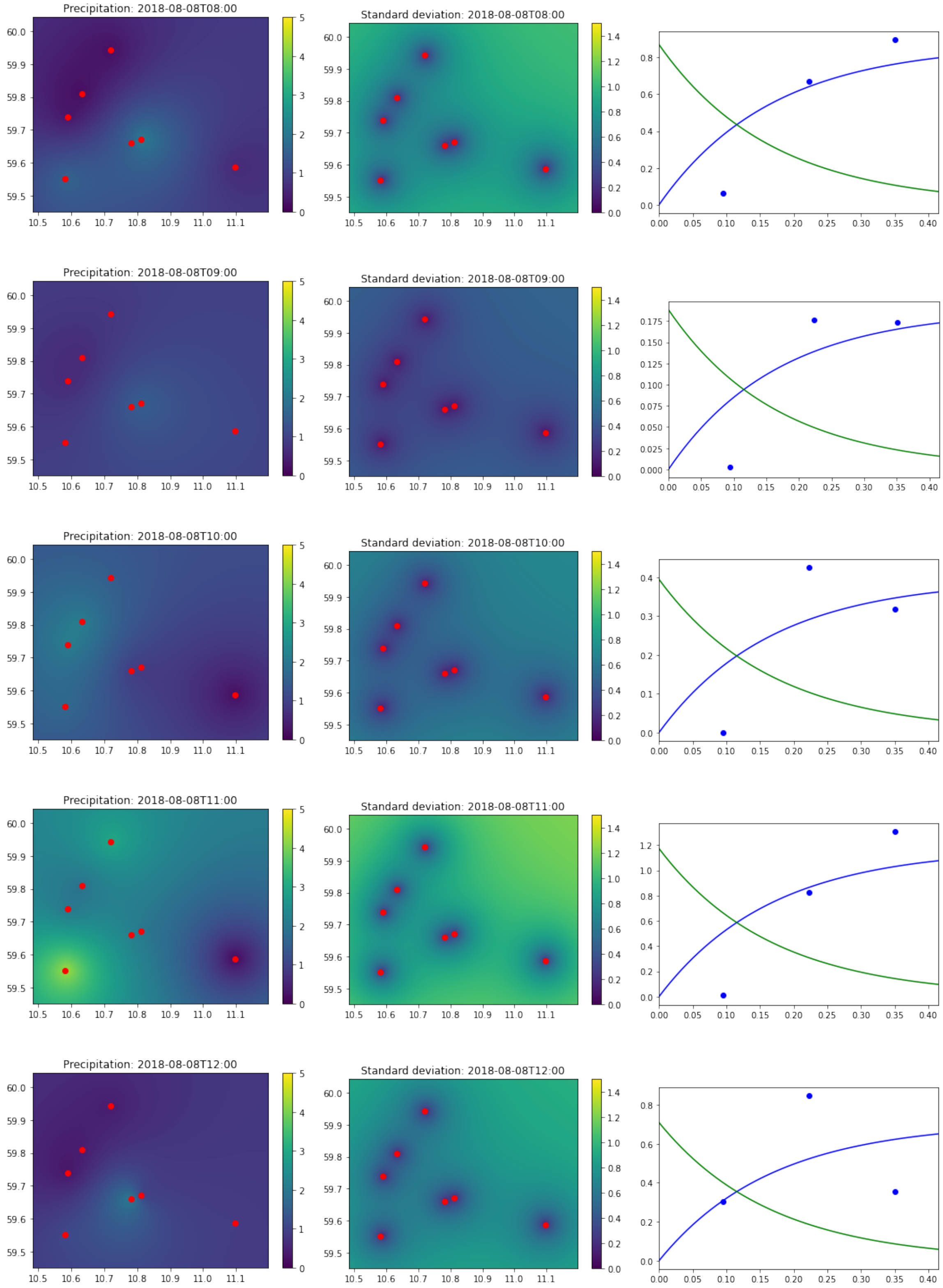

Figure 4: Exponential variogram: Left column: Kriging interpolation of precipitation. Middle column: Kriging variance for interpolated values. Right column: Semivariogram. The red dots show locations with observations, the blue dots shows the bins average semiavariance given by Equation 3.

at the observations, whereas the exponential model seems to have extreme values exactly at the observations. What the correct predicted precipitation and uncertainty should be is not known, but it is reasonable to expect that the maximal value of the field definitely can be larger that the largest observations and that uncertainty is lower between the stations. A way to validate and find the optimal semivariogram model could be to calculate the mean squared error using k-fold cross-validation, bootstrap or other resampling methods. These methods solves the Kriging equations on a subset of the observations and then uses the left out observations to compute the prediction error. Other more Kriging-oriented cross-validation techniques also exists [Dubrule, 1983]. Visually the Gaussian variogram seems to better fit the variance in the observations.

References

References

[Berne and Krajewski, 2013] Berne, A. and Krajewski, W. (2013). Radar for hydrology: Unfulfilled promise or unrecognized potential? *Advances in Water Resources*, 51:357–366.

[Chwala and Kunstmann, 2019] Chwala, C. and Kunstmann, H. (2019). Commercial microwave link networks for rainfall observation: Assessment of the current status and future challenges. *WIREs Water*, 6(2).

[Deutsch and Journel, 1997] Deutsch, C. V. and Journel, A. G. (1997). Geostatistical software library and user's guide.

[Dubrule, 1983] Dubrule, O. (1983). Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6):687–699.

[Kitterød, 2021] Kitterød, N. O. (2021). *Introduction to Geostatistics - Lecture notes.*

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning.

[Webster, 2001] Webster, R. (2001). *Geostatistics for environmental scientists.* John Wiley & Sons, Chichester, England New York.

Appendix A:

$$\gamma(u_1, u_2) = \frac{1}{2}\left(Z(u_1) - Z(u_2)\right)^2 \tag{28}$$

$$\gamma(u_1, u_2) = \frac{1}{2}Var\left(Z(u_1) - Z(u_2)\right) \tag{29}$$

$$\gamma(u_1, u_2) = \frac{1}{2}\frac{1}{N}\sum\left(Z(u_1) - Z(u_2)\right)^2 \tag{30}$$

$$\gamma(u_1, u_2) = \frac{1}{2}E\left(\left(Z(u_1) - Z(u_2)\right)^2\right) \tag{31}$$

$$\gamma(u_1, u_2) = \frac{1}{2}E\left(Z(u_1)Z(u_1) - Z(u_1)Z(u_2) - Z(u_2)Z(u_1) + Z(u_2)Z(u_2)\right) \tag{32}$$

$$\gamma(u_1, u_2) = \frac{1}{2}E\left(E\left(Z(u_1)^2\right) - 2E\left(Z(u_1)Z(u_2)\right) - E\left(Z(u_2)^2\right)\right) \tag{33}$$

$$\gamma(u_1, u_2) = -E\left(Z(u_1)Z(u_2)\right) + \frac{1}{2}E\left(Z(u_1)^2\right) + \frac{1}{2}E\left(Z(u_2)^2\right) \tag{34}$$

$$\gamma(u_1, u_2) = -E\left(Z(u_1)Z(u_2)\right) + \frac{1}{2}\left({\mu_1}^2 + {\mu_2}^2\right) \tag{35}$$

$$\gamma(u_1, u_2) = -E\left(Z(u_1)Z(u_2)\right) + {\mu_Z}^2 \tag{36}$$

$$\tag{37}$$