MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master’s Thesis

Vertical localisation for convective-scale data assimilation using a 1000-member ensemble

verfasst von / submitted by

David Hinger, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 614 (Master Meteorologie)
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Metorologie
Betreuer / Supervisor:	Univ.-Prof. Mag. Dr. Martin Weissmann
Mitbetreut von / Co-Supervisor:	Dr. Tobias Necker, M.Sc.

Contents

Abstract

The use of small ensemble sizes, due to computational restrictions in operational data assimilation, causes several inaccuracies, mainly because of spurious correlations. To mitigate sampling errors, different localisation methods have been developed over the last twenty years. This thesis seeks to find optimal localisation methods for vertical correlations of convective-scale forecast errors. This is done by using a convective-scale 1000-member ensemble as an assumed true depiction of the forecast error correlations and comparing it with localisations applied to 40-member ensembles randomly drawn from the 1000-member ensemble. The method of using the Gaspari Cohn function (Gaspari and Cohn, 1999) is optimised by using a variable dependent approach and also combining it with the statistical sampling error correction by Anderson (2012). Furthermore, a way to calculate the optimal weighting function is developed, to learn where the flaws of the currently used functions are and to use it directly for localisation. Lastly, the possibility of using machine learning for localisation is tested by using a random forest.

All developed methods bring improved results compared to the localisation of the Deutscher Wetterdienst (DWD). The best improvement is achieved by the optimal weighting function. It also shows that the shape of an optimal localisation function differs between self- and crosscorrelations. The random forest achieves clear improvement but shows many aspects which have to be considered building a stable and well-working machine learning tool. Using a variable dependent Gaspari Cohn function brings half as much improvement as the other two methods, but it shows the difference in the localisation of different parameters.

Zusammenfassung

Die Verwendung kleiner Ensemblegrößen, aufgrund begrenzter Rechenkapazitäten, in der operationellen Datenassimilation führt zu einer Reihe von Ungenauigkeiten, die hauptsächlich auf zufällige Korrelationen zurückzuführen sind. Um dieses Problem zu lösen, wurden in den letzten zwanzig Jahren verschiedene Lokalisierungsmethoden entwickelt. In dieser Arbeit wird versucht, optimierte Lokalisierungsmethoden für vertikale Korrelationen zu finden. Dazu wird ein convective-scale 1000-Member Ensemble als näherungsweise wahre Darstellung der Vorhersagefehlerkorrelationen angenommen. Dieses wird mit Lokalisierungen verglichen die auf 40-Member Ensembles, gezogen aus dem 1000er-Ensemble angewendet werden. Die Gaspari-Cohn-Funktion (Gaspari and Cohn, 1999) wird durch die Verwendung eines variablenabhängigen Ansatzes und durch die Kombination mit der Sampling Error Correction von Anderson (2012) optimiert. Darüber hinaus wird eine Methode zur Berechnung der optimalen Gewichtungsfunktion entwickelt, um zu erkennnen, wo die Schwachstellen der derzeit verwendeten Funktionen liegen, und um diese direkt zur Lokalisierung zu nutzen. Schließlich wird die Möglichkeit des Einsatzes von Machine Learning für Lokalisierung mit Hilfe eines Random Forest getestet.

Alle entwickelten Methoden bringen verbesserte Ergebnisse im Vergleich zur Lokalisierung des Deutschen Wetterdienstes (DWD). Die größte Verbesserung wird durch die optimale Gewichtungsfunktion erreicht. Es zeigt sich auch, dass sich die Lokalisierungsfunktion für Eigen- und Kreuzkorrelationen unterschiedliche Formen annimmt. Der Random Forest erziehlt ebenfalls eine deutliche Verbesserung, zeigt aber auch die vielen verschiedenen Aspekte, die berücksichtigt werden müssen, um ein stabiles und gut funktionierendes Machine Learning Tool zu entwickeln. Die Verwendung einer variablenabhängigen Gaspari-Cohn-Funktion bringt eine halb so große Verbesserung wie die beiden anderen Methoden, zeigt aber den Unterschied bei der Lokalisierung verschiedener Parameter.

1. Introduction

Data assimilation takes a crucial part in numerical weather predictions as it delivers the initial state of the atmosphere. All weather predictions rely on such an estimation of the state as a starting point for computation. This means the better the initial state is described, the better the weather forecast gets. The challenge in data assimilation is to get the best information out of many different types of observations and transform the unevenly distributed measurements into a homogeneous model grid. In order to deal with these challenges, the observations are combined with previous forecasts. To get the best combination, different data assimilation methods have been developed, with the most popular ones being the 3D-Var (Lorenc, 1986) and 4D-Var (Le Dimet and Talagrand, 1986) respectively and the Kalman Filter (Kalman and Bucy, 1961). This thesis focuses on the ensemble Kalman filter, which chooses the combination of observations and forecasts by weighting both of them relying on their estimated error correlations and covariances. In the beginning, the weights were calculated based on climatology.

A breakthrough in data assimilation was achieved by the use of ensembles instead of a climatology and the development of the Ensemble Kalman Filter (Evensen, 1994). This increased the quality of the estimated initial state and in the following weather forecasts. The Ensemble Kalman Filter calculates the error correlations out of the ensemble samples. These error covariances are calculated in the background error covariance matrix, short B-Matrix. In a further development, a hybrid version was established using both climatology and ensembles.

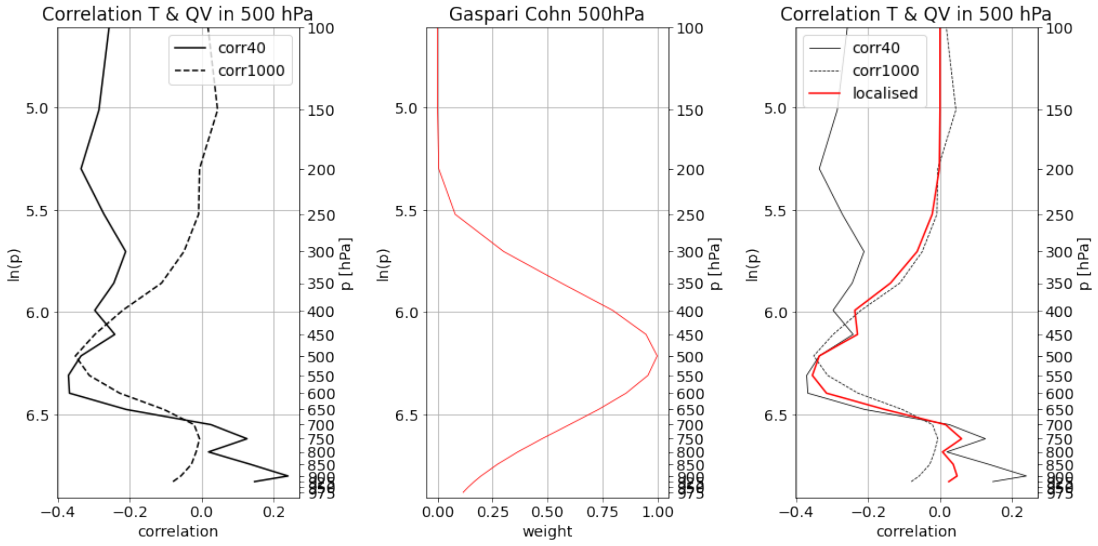
However, by using ensembles to calculate the error covariances, new problems arise. The reason is the use of low ensemble size because of limited computational power, as the common operational ensembles have sizes between 20 to 250 members, with 250 being the exception. Small ensemble sizes, as being not representative samples in statistical terms, cause spurious correlations which have no real physical relevance.

To tackle this problem, localisation methods are applied for data assimilation (Houtekamer and Mitchell, 1998),(Houtekamer and Mitchell, 2001). These methods mostly work by cutting off and damping spatial correlations after a specific distance. They rely mostly on predefined functions like the one developed by Gaspari and Cohn (1999). An example of a vertical localisation of a false correlation from a 40-member ensemble can be seen in Figure 1. It shows the comparison of a correlation of temperature (T) with specific humidity (QV) in 500 hPA at the same grid point between a 40- and 1000-member ensemble (Fig.1a), the applied localisation function, in this example a Gaspari Cohn function, (fig.1b) and the correlation localised with this function (Fig.1c).

Figure 1 shows that the correlation of the small ensemble (corr40) differs from the big ensemble (corr1000), which can be assumed to be a good depiction of the truth. The goal of localisation is to get the 40-member correlation into the same shape. This is done well by the localisation functions as especially the false correlations above 400 hPa and beneath 700 hPa are strongly damped or cut off.

However, these distance-based methods do not work perfectly. Problems are cutting of real physical correlations on long distances or that not all observations are on a fixed location but are integrated measurements from satellites and that the optimal damping differs from grid point to grid point.

Besides distance-based methods using different functions, other approaches to localisation have been developed, mainly using statistical methods like the sampling error correction (SEC) by Anderson (2012) or the global group filter (GGF) by Lei and Anderson (2014).



(a) *Correlation of 40- & 1000-member ensemble*

(b) *Localisation function*

(c) *Localised 40 member correlation*

Figure 1: Example of the localisation of a correlation of a 40-member ensemble compared to the correlation of a 1000-member ensemble; shown is the correlation of temperature (T) to specific humidity(QV) in 500 hPa

This thesis focuses on the optimisation of distance-based methods currently developed and on exploring how a perfect weighting function would have to look like. It is also tested if machine learning could be a possible approach to localisation in the future.

1.1 Approach

In order to be able to optimise and develop new localisation methods, an accurate depiction of the forecast error correlations is needed for evaluation. For this purpose, a convective-scale 1000-member ensemble is used in this thesis. Although 1000 members are a large number, it does not deliver a perfect depiction of the error correlations and therefore is only a very good approximation, which for the case of this thesis is assumed as the truth. As operational numerical weather predictions use ensembles with a much smaller number of members, 40-member ensembles drawn from the 1000-member ensemble are used to apply and test the different methods. This number of members is also run by the Deutscher Wetterdienst (DWD) in its ICON prediction model (Reinert et al., 2021).

The results of the localisation are then compared to the 1000-member ensemble, and the quality is compared to the uncorrected data and the localisation of the DWD. This thesis focuses only on the correction of vertical correlations. The main differences between vertical and horizontal correlations are the higher number of horizontal grid points and, therefore, the greater distance they span. A differentiation between vertical and horizontal localisation is also done by the DWD (Potthast, 2019). Its vertical localisation method serves as a benchmark for the methods developed in this work.

1.2 Research Questions

The overarching goal of evaluating different localisation methods is to improve the forecast error correlations. This leads to a better initial state of a forecast model and therefore increases the quality of the weather predictions. With this goal in mind, the research questions of this work are split into two parts and read as follows:

1. Localisation with currently used methods
   a) What is the optimal localisation scale for the Gaspari Cohn function?
   b) How does the optimal function, for correcting the error correlations, look like?
2. Could machine learning be a good alternative to conventional methods?
   a) How well works a random forest compared to currently used and optimised distance-based methods?
   b) Which configurations are needed to optimise a random forest?

2. Theoretical background

2.1 Ensemble data assimilation

In the whole process of weather forecasting, data assimilation takes the crucial part of defining the initial state of the atmosphere. The more precise the current state of the atmosphere is depicted, the more reliable the numerical weather prediction (NWP) gets. A great number of observations of different types (e.g. radiosondes, satellites, surface instruments, ...) are exploited to get the needed information. Most types of observations are not placed in a dense grid all over the world, rather there are some places where the observation density is high and others where nearly no information from instruments is available. To get information on the sparse regions of the world and transform them to the model grid, the observations are combined with the latest forecasts. Furthermore, do forecasts contain information about older observations and therefore increase the amount of exploited information. It is also important to know how the information from one observation location distributes to the next location. Therefore, it is important to know about the forecast error correlation between two different variables, different types of measurements, and different physical locations. With ensemble data assimilation it is possible to evaluate the forecast error correlation dynamically with every analysis step. Two types of data assimilation methods are common. There are variational methods like 3D-Var (Lorenc, 1986) and 4D-Var (Le Dimet and Talagrand, 1986) which are based on iteratively reducing a cost function. On the other hand, there are ensemble approaches as the Ensemble Kalman Filter (EnKF) (Evensen, 1994) and ensemble square-root filters (EnSRF) (Whitaker and Hamill, 2002) like the Local Ensemble Transform Kalman Filter (LETKF)(Hunt et al., 2007). They solve explicit equations to get the analysis state of the atmospheric system. (Petrie and Dance, 2010)(Ensemble-Data-Assimilation, n.d.)

The following sections describe the ensemble Kalman filter, the used localisation methods and the functionality of the random forest.

2.1.1 Ensemble Kalman Filter

To get the best estimate of the initial state of a numerical weather prediction sequential methods combine observations **y** with a model forecast x_b (background state). Both are weighted on the information of their errors. The initial state x_a (analysis) is computed from the background state with an added correction. This correction is obtained from the difference between observation and background state multiplied by the optimal weight matrix **K** which is given through a Kalman gain matrix: (Necker, 2019)

$$x_a = x_b + K(y - Hx_b) \tag{1}$$

$$K = BH^T(HBH^T + R)^{-1} \tag{2}$$

H is a linearised operator that transforms from model space to observation space, **B** is the background error covariance matrix and **R** is the observation error covariance matrix.

A further development to the EnKF is the LETKF by Hunt et al. (2007). This Kalman Filter uses only local observations inside a certain radius around each gridpoint in its analysis. Furthermore, the

B- Matrix is not calculated explicitly. This leads to the reduction of computational resources, which makes it usable in operational NWP.

In the ensemble Kalman Filter (Evensen, 1994) the background error covariance matrix is calculated using an ensemble forecast. It consists of the sample covariance calculated from the difference between each ensemble member and the ensemble mean.

$$B = P^b = \frac{1}{N-1}\sum_{n=1}^{N}\left(x_n^b - \overline{x^b}\right)\left(x_n^b - \overline{x^b}\right)^T \tag{3}$$

where **N** is the ensemble size, x_n^b is the state of the n-th ensemble member and $\overline{x^b}$ is the ensemble mean.

B - Matrix

The ensemble background error covariance matrix then takes the form of an n x n matrix:

$$B = \begin{pmatrix} cov(x_1) & cov(x_2,x_1) & \cdots & cov(x_n,x_1) \\ cov(x_1,x_2) & cov(x_2) & & \\ \vdots & & \ddots & \\ cov(x_1,x_n) & \cdots & & cov(x_n) \end{pmatrix} \tag{4}$$

The general definition of the covariance between two vectors **x** and **y** is:

$$cov(x,y) = \frac{1}{n-1}\sum_{i=n}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{5}$$

where x_i,y_i are the i-th value of the vectors and $\overline{x}$, $\overline{y}$ are the vector means.
The Pearson correlation coefficient $\rho_{x,y}$ between two vectors **x** and **y** is defined as:

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{6}$$

Concluding that the terms under the fraction line depict the standard deviations σ_x, σ_y of the vectors and the term above the derivation line is equal to the covariance in Equation 5, the correlation coefficient can be written as

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} \tag{7}$$

Therefore the background error covariance matrix consists of the information about the correlation and the standard deviation

$$B = \begin{pmatrix} \sigma_1^2 & \rho\sigma_2\sigma_1 & \cdots & \rho\sigma_n\sigma_1 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ \rho\sigma_1\sigma_n & \cdots & & \sigma_n^2 \end{pmatrix} \tag{8}$$

The B- Matrix is essential as it ensures a physically consistent and balanced forecast model. The correlation inside the matrix also brings information about how observation information spreads geographically and between different variables. Under the assumption of the same errors, this can be understood as that the updated initial state at one point is the same update at the next point multiplied by the correlation coefficient.

Due to the use of small ensembles in operational NWP systems, problems regarding the correlations can occur. The main problem is that small ensemble sizes are not statistically representative which leads to undersampling. This causes problems such as sampling errors, spurious correlations at long range, inbreeding, and filter divergence.(Petrie and Dance, 2010)

To tackle this problem, a variety of methods have been developed. These methods are mostly called localisation, though this only depicts a certain type of method. They try to correct the correlations or covariances in the Kalman Gain to get a true picture of the forecast error correlations.

2.2 Localisation

In the last twenty years, a variety of different methods has been developed to correct the covariances or correlations in the B- matrix of ensemble data assimilation and tackle the problems caused by small ensemble sizes (spurious correlations, inbreeding, filter divergence, etc.). These methods can be split roughly into two groups: distance-based and statistical methods. A third approach explored in recent years is using machine learning (e.g. Moosavi et al. (2019)). It also has to be differentiated how the methods are applied to the data assimilation process. They can be applied to model space or observation space. The model space localisations are applied directly to the background error covariance (B-Matrix) and typically apply distance-based methods. On the other hand, observation space localisation tapers the correlations between the model space and observation space. This can be done with distance-based methods, R-matrix inflation, or the use of statistical methods. In this thesis, the focus is on model space localisation. Both types of methods are shortly described in the following.

2.2.1 Model space localisation

Model space localisation mostly uses distance-based localisation. These methods, also known as covariance localisation, apply a tapering function that weights correlation points near the outgoing correlation point higher than points further away. They rely on a "cutoff length" which defines a distance at which all correlations further away are set to zero. This method mainly tackles spurious correlations as they are mostly eliminated due to the cutoff length.
A localisation matrix using a weighting function is applied to the background error covariance matrix via a Schur Product (Schur, 1911)

$$(A \circ B)_{ij} = A_{ij} B_{ij} \tag{9}$$

which is an element-wise multiplication between two matrices of the same form. Applied to the Kalman Gain (Eq. 2) with the localisation matrix denoted as **C** it takes the form of

$$K = (C \circ B) H^T [H(C \circ B) H^T + R]^{-1} \tag{10}$$

Examples of functions utilised in distance-based methods are the Gaspari Cohn function (Gaspari and Cohn, 1999) used by e.g. Anderson et al. (2005), Houtekamer et al. (2005), Campbell et al. (2010) or satellite weighting functions (Miyoshi and Sato, 2007). Other approaches are a multivariate (Roh et al., 2015) or a scale dependent localisation (Buehner and Shlyaeva, 2015).

In this thesis, the distance-based localisation method used to optimise localisation applies the weighting function postulated by Gaspari and Cohn (1999). It also serves as a benchmark for other methods. For more information see section 2.2.3.

2.2.2 Observation space localisation

In observation space localisation the distance-based localisation matrices C_1, C_2 are applied after the multiplication of **B**-matrix and forward operator **H**. Which denotes as:

$$K = C_1 \circ (BH^T)[C_2 \circ (HBH^T) + R]^{-1} \tag{11}$$

(Lei and Whitaker, 2015)

Alternatively to distance-based methods, statistical methods can be applied in observation and model space localisation. They try to correct the correlations by using different types of filters like the hierarchical filter developed by Anderson (2007) or the Global Group Ensemble Filter (Lei and Anderson, 2014). The method used in this work to find an optimised localisation method is the Sampling Error Correction by Anderson (2012). It is combined with the distance-based method using the Gaspari Cohn function (see sec. 3.3.1).

2.2.3 Gaspari Cohn function

The most common function for distance-based localisation was defined by Gaspari and Cohn (1999).

$$C_0(z,c) = \begin{cases} -\frac{1}{4}(|z|/c)^5 + \frac{1}{2}(z/c)^4 + \frac{5}{8}(|z|/c)^3 - \frac{5}{3}(z/c)^2 + 1, & 0 \le |z| \le c, \\ \frac{1}{12}(|z|/c)^5 - \frac{1}{2}(z/c)^4 + \frac{5}{8}(|z|/c)^3 + \frac{5}{3}(z/c)^2 - 5(|z|/c) + 4 - \frac{2}{3}c/|z|, & c \le |z| \le 2c, \\ 0, & 2c \le |z| \end{cases} \tag{12}$$

The function is applied differently in horizontal and vertical localisation. In the following, only the vertical approach is described. The Gaspari Cohn function has a symmetric shape when the height is given as the logarithm of pressure (ln(p)). It has a Gaussian-like form. The distance between the correlation level and the other levels is given as **z**. At the correlation level, the Gaspari Cohn function is set to one.

The function decreases from one to zero so that spurious and physically not relevant correlations further away are not considered. A length scale **c** is defined so that at two times **c** the correlations are set to zero. The length scale is defined as

$$c = \sqrt{\frac{10}{3}} l \tag{13}$$

where **l** is a free to choose localisation scale (LS), normally between 0 - 1. This is done to get an optimised version of the original Gaspari Cohn function (Lorenc, 2003). This thesis seeks the optimal localisation scale, that corrects the sample correlations in the best way. In Figure 2 the Gaspari Cohn function (GC) is shown at a correlation level of 500 hPa for four different localisation scales. If the localisation scale is bigger more correlations are considered in the localised data.

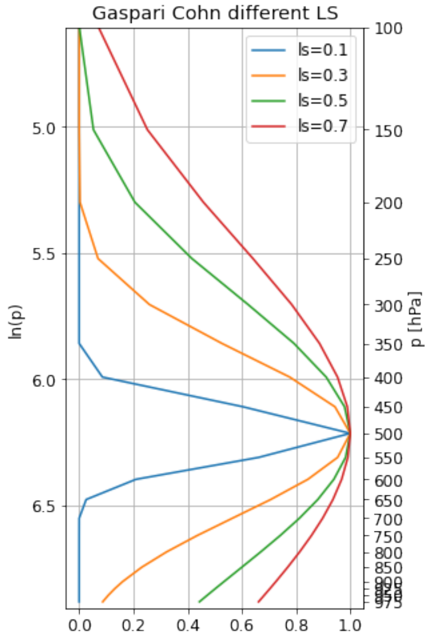


Figure 2: Example of a Gaspari Cohn function at the 500 hPa correlation level, for 4 different localisation scales

Gaspari Cohn DWD

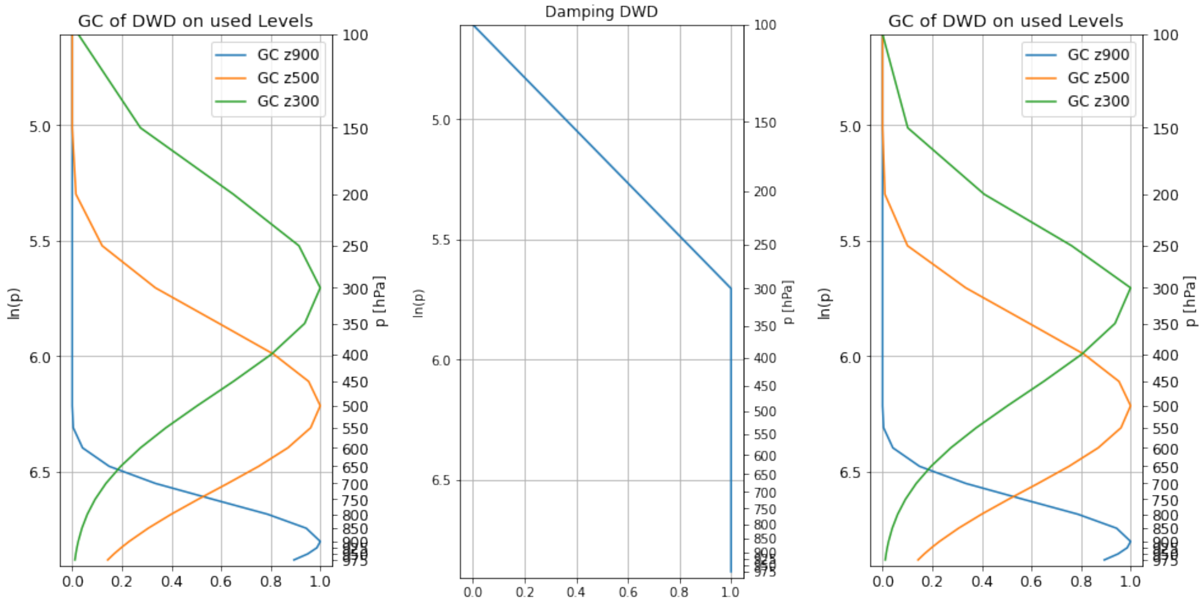
The Gaspari Cohn function as applied by the DWD serves as a benchmark to evaluate the results of the optimised Gaspari Cohn function, the optimal weighting function, and the random forest experiments. Therefore, the localisation scale settings of the DWD (Potthast, 2019) are applied to the data of this work (see cha. 3).

A fixed localisation scale for every correlation level is used, increasing linearly from 0.1 at the lowest level to 0.5 at the top level of the model domain. Originally the DWD starts with a localisation scale of 0.075 (Potthast, 2019), but because of the fewer levels in this thesis, the start is set to 0.1. An example of the broadening with the height of the Gaspari Cohn function as utilised by DWD is shown in Figure 3a. Furthermore, a damping term is introduced for upper tropospheric levels above 300 hPa. It decreases linearly in ln(p) to zero at the top level (see Fig. 3b). The final Gaspari Cohn function as used by the DWD can be seen in Figure 3c.(Potthast, 2019)

2.2.4 Sampling error correction (SEC)

The sampling error correction by Anderson (2012) exploits a look-up table to find the correction $\gamma_{m,p}$ for a given sample correlation ρ. It only depends on the size **m** of the ensemble and a prior distribution **p**. The corrected correlation ρ_{sec} is then calculated as

$$\rho_{sec} = \gamma_{m,p}\,\rho \tag{14}$$



(a) DWD GC without damping

(b) DWD damping

(c) DWD GC with damping

Figure 3: (a) Example of the DWD Gaspari Cohn function of three different levels, (b) in ln(p) linear damping for levels above 300 hPa, (c) GC plus damping, in this form used by the DWD

It is assumed that the sampling error only comes from the correlation coefficient and that all correlation coefficients are drawn from the prior distribution U[-1, 1], with this information $\gamma_{m,p}$ is calculated with an offline Monte Carlo technique. The look-up table won from this calculation is used to correct the overestimation of correlations caused by spurious correlations. It is a simple and practicable method, because the only input needed is the ensemble size and the calculated ensemble correlation. (Anderson, 2012)(Necker, 2019)

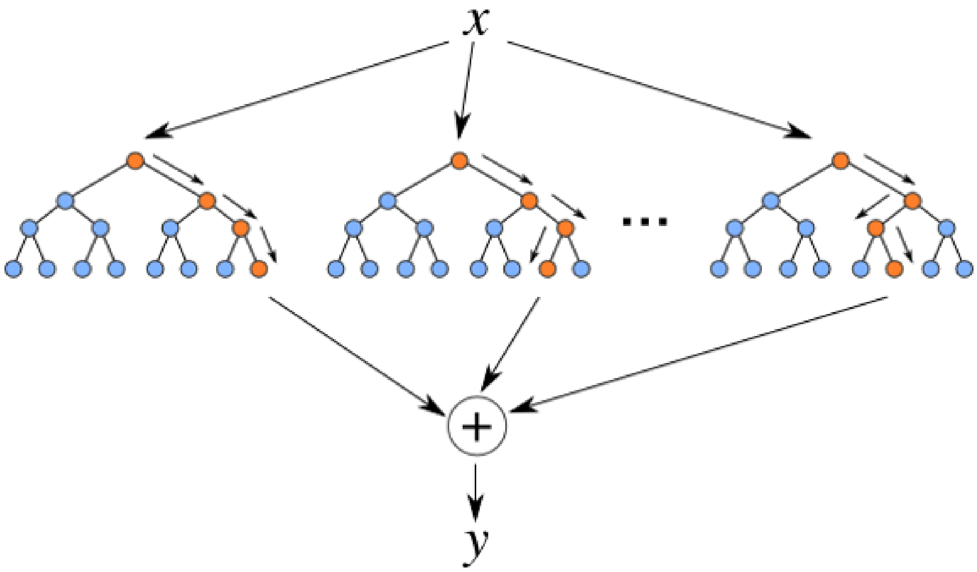
2.3 Random forest

In order to test an alternative approach that does not rely on fixed functions, machine learning is explored in this thesis. This wide range field gained more and more popularity in the new millennium, mainly due to faster computers. In line with this work, it is not possible to cover all possible approaches so a relatively simple method, but a very popular one, the random forest (Breiman, 2001), is chosen to test localisation with machine learning. This supervised learning technique is selected because it is easy to implement. Furthermore, it allows one to look at its structure, which helps in finding the optimal configuration for successful localisation.

How does it work?

A random forest (Breiman, 2001) consists of multiple decision trees. Each of these decision trees is trained with a different randomly chosen subsample of the training data. The subsamples are drawn with replacement, this is called bagging (Breiman, 1996). Randomisation of the data, by using multiple different trees, is done to reduce the variance of the estimator, because single decisions trees tend to have high variance which leads to overfitting. The result of training the random forest are multiple different structured decision trees, which produce different predictions. The final prediction is

then calculated by taking the mean of the predictions from the decision trees. (Breiman, 2001)(Ho, 1995)(Pedregosa et al., 2011)
A simple sketch of this procedure can be seen in Figure 4. Here x depicts the test data and y the predicted outcome. Shown in orange are the different prediction paths which lead to the different guesses. They are then averaged to get to the final prediction y.



Source: `https://ai-andi.com/random-forest/`

Figure 4: Concept of a random forest

Decision tree

A decision tree uses training data to build a flowchart-like construct, which can be utilised to predict an outcome (target) based on values from different parameters (predictors). The target and the predictors are connected, for every target value, there is a value of every predictor. For example, the target of this thesis is the correlation of the 1000-member ensemble, so at every gridpoint in the domain, there is one target value connected to one value of each predictor like temperature, humidity and pressure.
A decision tree uses the predictors to split the training data into groups with the lowest possible variance in the target values. At every splitting point, called leaf (circles in Figure 4), the data is split into two groups, based on if the values of a specific predictor are above or below a deciding value. Then these groups are split again until there is only one target value left in a group or there is no variance between the target values in a group. As the goal of the splitting is to find invariant groups of the target, the deciding value is chosen so that the resulting groups have the lowest possible variance. The criteria to get this is the mean squared error (MSE) of the target values. Therefore, at every leaf, a decision tree searches for the predictor and deciding value where the lowest MSE in the split groups is achieved. A simple decision tree with the 1000-member correlation as the target is shown in Figure 5. This tree is cut after the third splitting. At every leaf are shown in order: the deciding predictor and value, the MSE of group, the number of target values in the group and the mean of the target values in the group.

After finding all branches and leaves, the prediction with a decision tree starts with the first leaf. The connected predictors of the test data follow through the leaves at each deciding whether they are

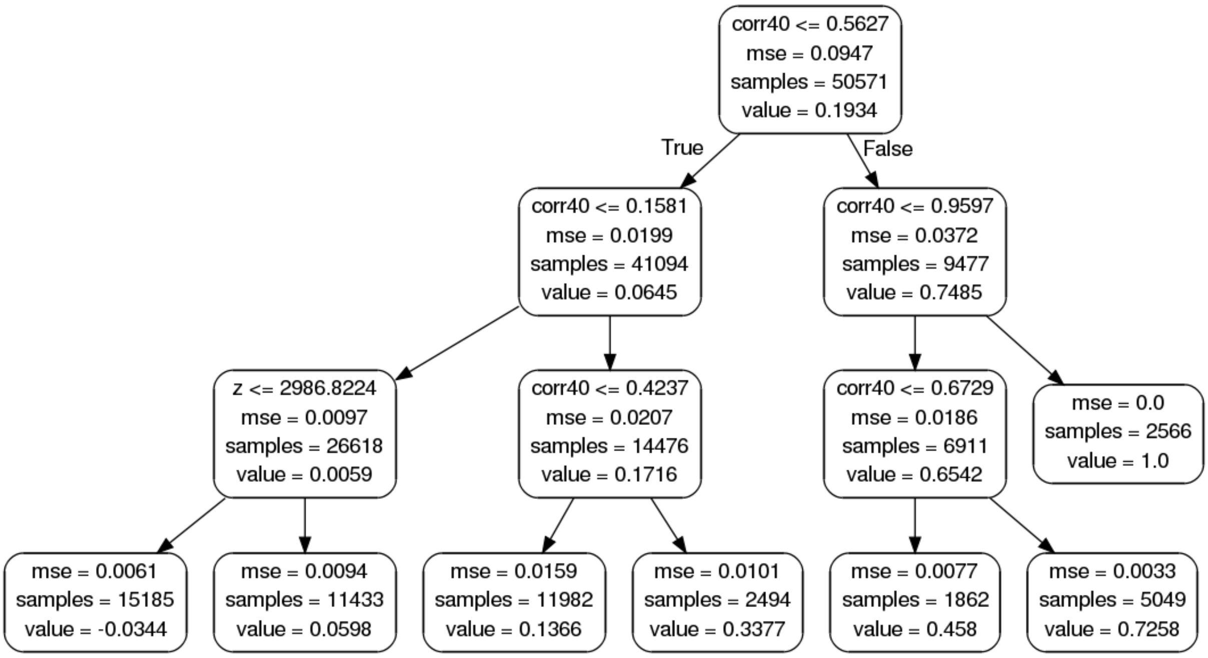


Figure 5: Example of a decision tree trained to predict the 1000-member ensemble correlation; at every leaf are shown in order: deciding predictor and value, MSE of the group, number of target values in the group, mean of the target values in the group

above or below the deciding value. When the end of the path (orange dots in Figure 4) is reached the last leaf gives the prediction which is the remaining target value at this leaf.

3. Data and methods

3.1 1000-member ensemble

The data used to calculate the correlations are five convective-scale 1000-member ensemble forecasts, which were simulated for the time period of 29.5.2016 to 2.6.2016. Every forecast has a lead time of 14 hours, but only the three-hour forecast is utilised in this thesis. It consists of 30 vertical levels, 20 of which are used in this study. Calculations for this data were done at the K-Computer of the RIKEN Center for Computational Science.

The data assimilation is done on a domain with 15 km grid spacing, for which the boundary conditions of the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP) are taken. The different ensemble members are calculated by combining the perturbations of a global 20-member GFS forecast combined with 1000 random perturbations that are scaled with a climatology from the Climate Forecast System Reanalysis (CFSR) data set (Saha et al., 2010). (Necker et al., 2020)

3.1.1 Correlations

Vertical correlations were calculated for 25 subsamples, containing 40 members, of the 1000-member ensemble and for the whole ensemble. They were computed for the three-hour forecast on five days at all available model levels by Tobias Necker. This means at every gridpoint in every level exists a correlation to all other grid points above and beneath it. An example of a correlation in 850 hPa to 6 other levels is shown in Figure 6. In this example, there are seven correlation values to the point in 850 hPa. In the data of this thesis, 20 values correlated to all other levels in the vertical column exist at every gridpoint.

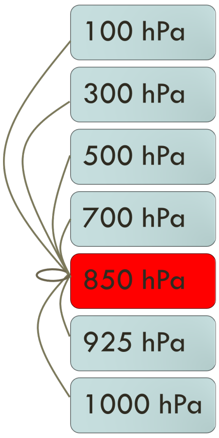


Figure 6: Example of a vertical Correlation in 850hPa

3.1.2 Domain

The domain of the data is located over central Europe, primarily Germany and the Alps in the southern parts of the domain. It has a size of 252x350 gridpoints, with a spacing of 3km (see Fig. 7)(Necker et al., 2020). To avoid problems in calculation at the domain borders, it is cut to a size of 200x200 gridpoints (see Fig. 7).

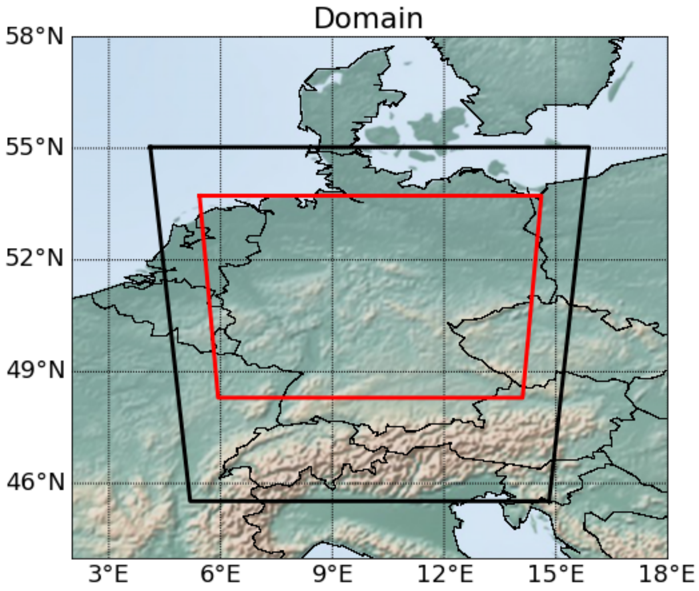


Figure 7: Domain of used 1000-member ensemble, ***black****: full domain, red: used Domain*

3.1.3 Parameters

There are two groups of available parameters. The first are all parameters, which are calculated for all levels (see tab. 1). These are used to calculate the vertical correlations and in further consequence to test the localisation methods. They consist of the temperature (t), three wind parameters (u, v, w), four parameters that deliver information about the water in the atmosphere (qv, rh, qhydro, dbz) and the height of the level (z).

Table 1: Table of available parameters calculated in all levels

	Full name	**Unit**
t	temperature	K
u	u-wind	m/s
v	v-wind	m/s
w	w-wind	m/s
qv	water vapor mixing ratio	kg/kg
rh	Relative humidity	%
qhydro	Mixing ratio of all hydrometeors	kg/kg
dbz	Radar reflectivity	dBZ
z	Height of Level	m

The second group consists of ground parameters (see tab. 2). They are only applied in the localisation with a random forest. Particularly, the precipitation transformed from a precipitation rate in mm/s to an hourly precipitation in mm/hr. Likewise used is the sea level pressure (slp).

Table 2: Table of available ground parameters

	Full name	**Unit**
slp	Seal level pressure	Pa
max_dbz	Maximum radar reflectivity	dBZ
t2	2m temperature	K
u10	10m u-wind	m/s
v10	10m v-wind	m/s
q2	2m water vapor mixing ratio	kg/kg
prec	Surface precipitation rate	mm/s

3.2 Perfect weighting

In distance-based localisation, a fixed predefined function is used which is fitted to get the best outcome. These functions are mostly in a Gaussian shape with a peak of one at the correlation level, like the one from Gaspari and Cohn (1999). They have the advantage of being easy to compute and are normally of a smoothed shape. Furthermore, the Gaspari Cohn function guarantees a positive definite localisation matrix. That is important because the correlation matrix (B-matrix) has to be positive definite, as a negative definite matrix would lead to negative variances, which are not defined mathematically.

As this thesis is about optimising localisation it is important to learn about the flaws of the currently used functions and how an ideal localisation function would look like. For this reason, the actual weighting that would be necessary to get from a correlation to the localised correlation is calculated as shown in Equation 15. Here ρ is the correlation which gets localised, ρ_{loc} is the localised correlation and w depicts the applied weighting.

$$\rho \cdot w = \rho_{loc} \Rightarrow w = \frac{\rho_{loc}}{\rho} \tag{15}$$

With this method, it is possible to calculate the perfect weight and in further consequence, when done for a complete vertical profile, the perfect weight/localisation function. The only requirement is to know the physically true correlation at each point.

The true correlation surely can never be known, so an approximation that gets as close as possible to the truth has to be considered. For this purpose, correlations of a 1000-member ensemble (see section 3.1) are assumed as the true depiction of forecast error correlations and the perfect weight (**w**) is calculated with

$$w = \frac{\rho_{1000}}{\rho_{40}} \tag{16}$$

where ρ_{1000} is the correlation of the 1000-member ensemble and ρ_{40} a correlation of a 40-member ensemble.

This calculation is done for all available subsamples and correlations. The mean of all vertical profiles then gives a mean perfect vertical weighting function, further on called optimal weighting function.

Filtering outliers

A problem that occurs, when calculating the optimal weighting function is very high or low weights with magnitudes above 10^2. These weights only work for the grid point they are calculated at, but when the mean of all profiles is computed, this high weight distorts the mean weight to a not useful one. This occurs due to a high difference between the correlation of the 40-member ensemble and the 1000-member ensemble, mainly when the 40-member correlation is near zero. For example, at one point $\rho_{40} = 10^{-6}$ so virtually zero and $\rho_{1000} = 10^{-3}$ so also near to zero. The resulting perfect weight (Eq. 16) would be $w = 1000$. If this weight would be taken into account at other points it would cause a miserable localisation. For that reason, the calculated perfect weights have to be filtered so that the result is not distorted in any direction and a good localisation can be achieved.

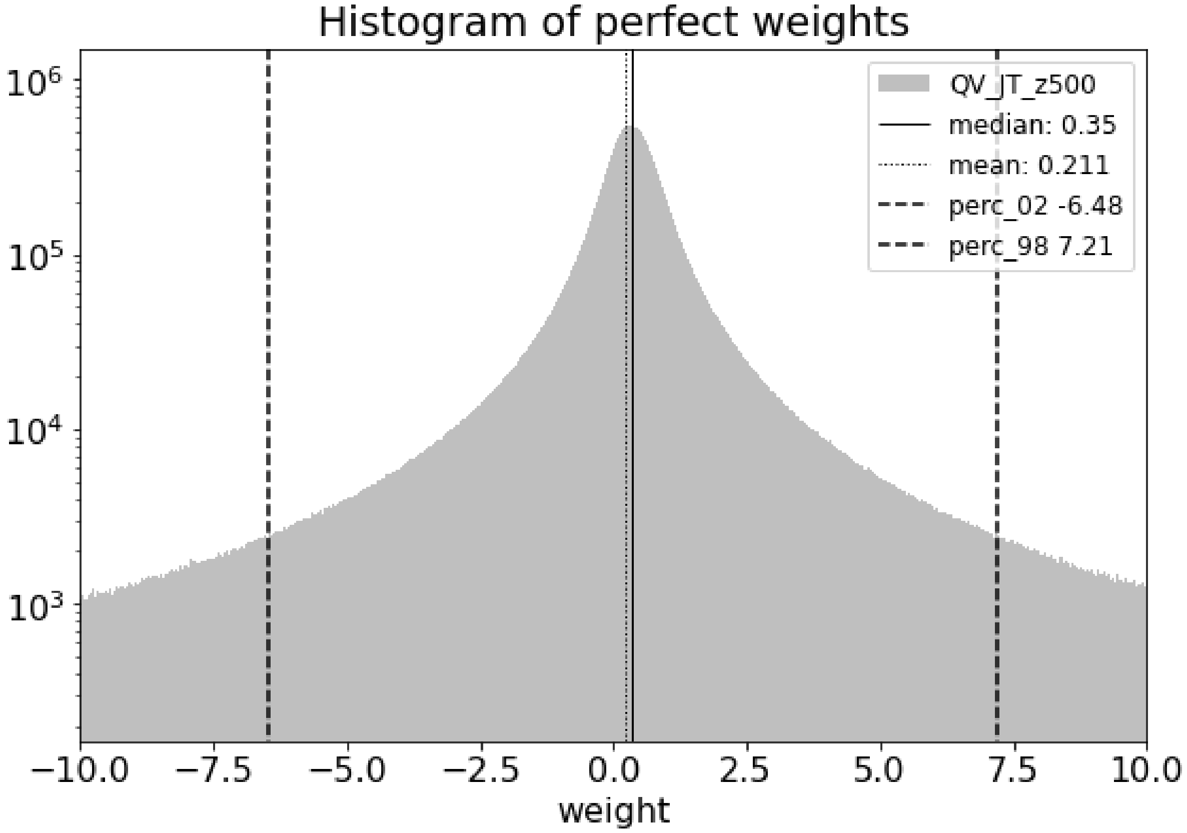


Figure 8: Histogram of weights of all grid points for the Correlation of QV with T in 500 hPa; the dashed lines show the 2% and 98% percentile which are used as the threshold for the outlier filtering, the solid line depicts the median, the dotted line the mean

In this thesis, the threshold values are set to be the 2% and 98% percentile. This means the lowest and highest 2% of the weights are filtered out and ignored in the calculation of the optimal weighting function. This threshold was chosen because it brought the best result without cutting too much of the data. Averaged over all types of correlations, the lower threshold is -6.11 ± 0.84 and the upper threshold is 6.8 ± 0.87. An example for the distribution of the calculated weights is shown in Figure 8. Used is the correlation of specific humidity with temperature in 500 hPa. It shows a symmetrical distribution with 96% of the weights lying in the area between the dashed lines which depict the 2% and 98% percentile.

3.3 Optimal Gaspari Cohn

As one goal of this work is to optimise current localisation methods, an empirical approach is used to get an optimised Gaspari Cohn function. Therefore a perfect localisation scale that achieves the best localisation for the specific forecasts time, subsample, correlation level, and pair of correlated parameters (see tab.1) is sought. Then, by taking the mean of all perfect localisation scales at each correlation level and for each pair of parameters, an optimal variable dependent Gaspari Cohn function can be calculated. A non-parameter dependent function is achieved by taking the mean of all variable dependent optimal localisation scales.

The perfect localisation scale is found by calculating a Gaspari Cohn function for each localisation scale between 0,05 and 2 with $\Delta = 0,05$. These functions are then used to localise correlations of 40-member subsamples drawn from the 1000-member ensemble (see Section 3.1). After the localisation, the root mean squared error (RMSE) between the localised correlation and the 1000-member correlation is calculated for every vertical profile in the domain (see Section 3.1.2) and averaged over the whole domain. The perfect localisation scale is then the one achieving the lowest RMSE.

3.3.1 Gaspari Cohn + SEC

In this work, the SEC (see sec. 2.2.4) is applied as one possibility to find an optimised localisation method. To achieve this the SEC is applied to the correlation of a 40-member ensemble drawn from the 1000-member ensemble (see Sec. 3.1) and then an optimal Gaspari Cohn function (see sec.2.2.3) is found for the corrected correlations. This function can be utilised to localise a sampling error corrected ensemble and so is a combination of a statistical and a distance-based method.

3.4 Random forest

3.4.1 Implementation

For the implementation in python for this thesis, a random forest regressor of the scikit learn tool (Pedregosa et al., 2011) is used. A single random forest for the whole domain is trained. This is done for a specific correlation of one parameter in a profile with another at a specific vertical level (see Fig. 6).

The training data consists of the values of the chosen predictors at every gridpoint if available. The minimum selection of predictors, further on called basis random forest, are the correlation of the 40-member ensemble, as this is to be optimised, and the parameter, which is correlated, from every level in the vertical profile, to a parameter at a specific level. As an example, if the correlation of the specific humidity (QV) and the temperature (T) at 500 hPa has to be localised, the predictors of the basis random forest would be the 40-member correlation and QV.

If the predictor is a surface parameter only the gridpoints at the lowest level have these values in the training data. Also, no values are given at gridpoints where the model levels are beneath the orography.

The target value is always the correlation of the 1000-member ensemble. Therefore, when training a random forest it will try to get as close to these correlation values using the given predictors. The random forest then delivers a corrected correlation value at every grid point in the domain in its prediction.

3.4.2 Optimisation

There are many different options to optimise a random forest. For example, the number of trees can be chosen freely, but it has to be considered that with each tree added the computation time increases. Knowing that at some point, adding more trees will not improve the random forest anymore, because the number of vastly different drawn subsamples is limited due to the size of data. A number of trees has to be chosen for which the random forest improves significantly and the computation time does not exceed the available resources. In this work, the number of trees is set to 10.

One major point in the attempt to optimise a random forest is how the data is handled. A bigger data size does not guarantee better results. It is more important to single out the data that brings improvement, the redundant data, and in the worst-case data that trains the random forest in a false direction and makes it useless for different test data. Therefore, neglected in the random forest is all kind of meta-information, because it would train the random forest to predict correlations on the information of geographical position and time. For example, the random forest would learn that at a specific longitude, latitude, and time, the same correlation is always the correct one. This is a problem as the weather situation is never strictly the same at one point and so the random forest would be ineffective when executed on any data other than the training data. So in order to train the random forest only on physical relations, no metadata and only physical parameters (see tab.1&2) are used as predictors. However, the random forest can be improved by choosing the right or best parameters. It is not as straightforward as with the metadata to choose the right predictors, because the physical connections between different parameters and especially their influence in training a random forest on correlations are not always very clear. It is possible that some information, when used in training, will lead to worse results, because there may be a random statistical connection between them and the calculated correlations, but no physical relevance. As it is difficult to theoretically find the right predictors, a so-called parameter test was developed that allows optimising the random forest in relation to the basis random forest.

Parameter test

The goal of the parameter test is to determine which parameters improve and which worsen the random forest. The basis random forest serves as the reference value. This means the basis random forest is trained with four of the five available days of the data (see cha.3) and then tested on the remaining day. The RMSE between the correlations of the 1000-member ensemble and the received corrected correlations of the random forest is calculated for every vertical profile in the domain and averaged over the whole domain. In addition, the relative difference between the RMSE of the basis random forest and the RMSE between the 1000- and 40-member correlation is calculated.

Now the same calculation is done with a new random forest using a third predictor from the available physical parameters in addition to the predictors of the basis random forest. Then the difference between the RMSE and the relative improvement of the basis random forest respectively the new random forest is calculated.

This computation is done for all parameters in Table 1 as well as the precipitation (prec) and the sea level pressure (slp) (tab.2). The results are then compared to each other and all predictors which achieve better relative improvement than the basis random forest are added to the basis predictors and applied in an optimised random forest.

Filtering data

Besides changing the size of data, the number of trees or adding different predictors another method to improve the random forest is to filter the training data. The target is to eliminate values that have a high divergence from the values assumed as the truth to which the random forest is trained. In this case, it is about the correlations of the 40-member ensemble, which are way higher or lower than the ones of the 1000-member ensemble.
Therefore before the computation of the random forest takes place the perfect weights as calculated in Equation 16 are calculated for every grid point in the training data. Then all grid points where the calculated weight is not between the 2% and 98% percentile are eliminated from the training data similar to the calculation of the optimal weighting function (see sec.3.2). The now filtered data is used to train the random forest.

4. Results

This chapter presents the results obtained by evaluating different localisation methods. The results can be split into three parts: the optimisation of the localisation with Gaspari Cohn (see sec.3.3), the searching for an ideal weighting function (sec.3.2) and localisation using a random forest (sec.3.4). To test these methods, they are applied to the correlations of 40-member subsamples drawn from the 1000-member ensemble. The results are compared with the RMSE between the correlations of the 1000-member ensemble and the 40-member ensemble or the currently used method of the DWD. The goal of every localisation is to reduce this RMSE and therefore get correlations similar to the 1000-member ensemble as this is assumed to depict the true correlations of the atmospheric state in the best approximation. In the shortcuts of the correlations, the parameter named first is the parameter correlated from every level in the vertical profile with the one at the correlation level. For example, if the specific humidity (QV) of every level in the profile is correlated with the temperature (T) at the correlation level, the shortcut would be QVT.

4.1 Optimal Gaspari Cohn

As the Gaspari Cohn function is the most common in distance-based localisation, the attempt is to find the best configuration of the function. To do this, the optimal localisation scale (LS) is sought as described in section 3.3. This is done for two cases. First, for the case of a non-variable dependent function, where the localisation scale differs between different correlation heights but not between different correlations of variables.

In Figure 9 the relative improvement of the RMSE of the localised correlation in comparison to the RMSE of the 40-member correlations is shown for different localisation scales of a non-variable dependent Gaspari Cohn function. It is calculated as the mean of localisations with different localisation scales of five days á 25 subsamples and all available parameters. The curves for three different correlation levels can be seen. Averaged are the self- and crosscorrelations of the following parameters:

1. Temperature (T)
2. Specific Humidity (QV)
3. Mixing ration of hydrometeors (QHYDRO)
4. u - wind (U)
5. v - wind (V)
6. vertical wind (W)

The optimal localisation scales are given by the positions of the peaks of the curves, marked by dotted lines. As can be seen, the optimal localisation scales are in an area between 0.35-0.7. This is slightly higher than the ones applied by the DWD especially in the lowest atmospheric levels where the DWD uses localisation scales around 0.1. In fact, Figure 9 shows that if the localisation scale is chosen too low, the localisation can result in a worse result than the unlocalised ensemble. This is because with very low localisation scales nearly all correlations at neighbouring levels are set to zero, therefore no real correlation is concluded. This is not the reality in most cases. On the other hand, if the localisation scale is chosen high, the quality of the localisation also declines, but not as drastically as it does with low localisation scales. With higher localisation scales the relative improvement asymptotes zero, because the higher the localisation scale gets, the fewer levels are cut off or damped as the Gaspari Cohn function gets wider, resulting in zero when every level is equally considered, meaning no localisation is applied.

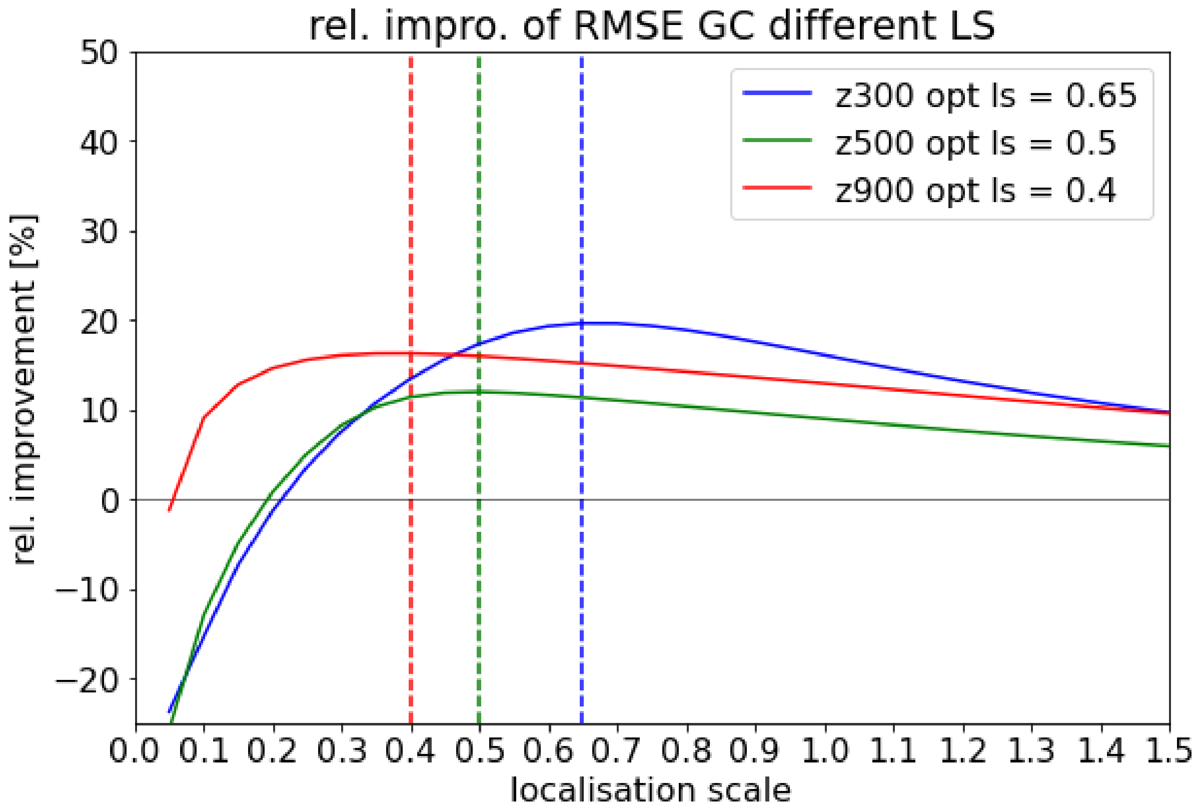


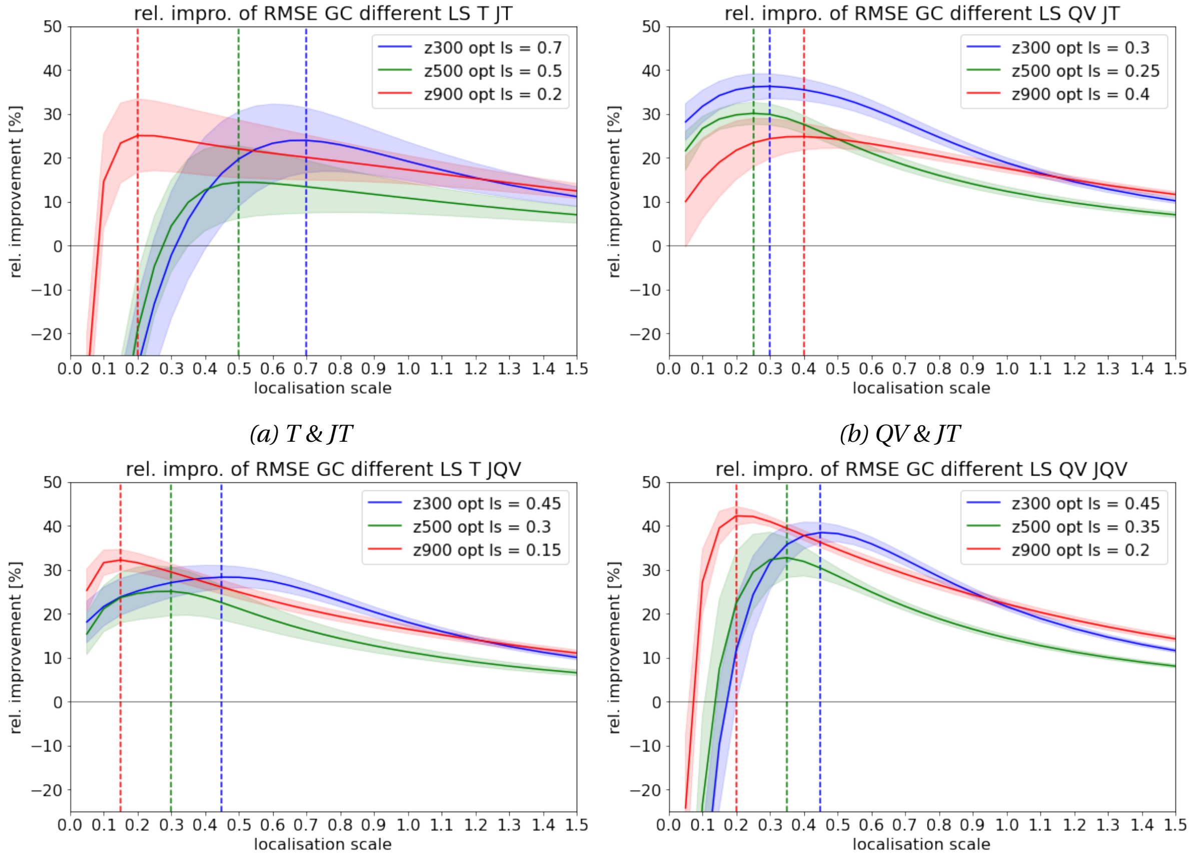
Figure 9: Comparison between localisation scale (x-axis) and relative improvement to 40-member ensemble (y-axis), in three correlation levels: 900 hPa, 500 hPa, 300 hPa; mean of all Correlations on 5 days and 25 subsamples; averaged are correlations of T, QV, QHYDRO, U, V, W ; dashed lines mark the maximum improvement

Variable dependent

In opposition to a single function for all correlations, a variable dependent function is a more precise attempt in localisation. It takes the different ranges of correlations between certain variables into consideration, as some correlations impact levels farther away and others have very small impact range and therefore need smaller localisation scales. In order to find the optimal variable dependent localisation scales, the same procedure as for the non-variable dependent localisation scale is applied (see section 3.3), with the only difference being that it is done for all different types of correlations separately.

Figure 10 shows the connection between localisation scale and relative improvement (same as in Fig.9) for the self- and crosscorrelations of temperature (T) and specific humidity (QV). The variable denoted with a J before it is the one on a specific level to which the other parameter from all other levels is correlated (see Fig. 6).

In comparison to the non-variable dependent localisation scale (Fig.9) the relative improvement at the maxima is about 10% - 20% higher for the variable dependent localisation scales. Especially well works the variable dependent function for the selfcorrelation of QV (Fig.10d), with peaks up to 45%. A major difference to the non-variable dependent function is the range in which the optimal localisation scales are. It ranges mostly from 0.15 to 0.45 as for the non-dependent the range was 0.35 - 0.6. An exception is the selfcorrelation of T (Fig. 10a) in the middle to high levels which have bigger localisation scales, but it has to be considered that especially for a height of 500 hPa the curve is very



(a) T & JT

(b) QV & JT

(c) T & JQV

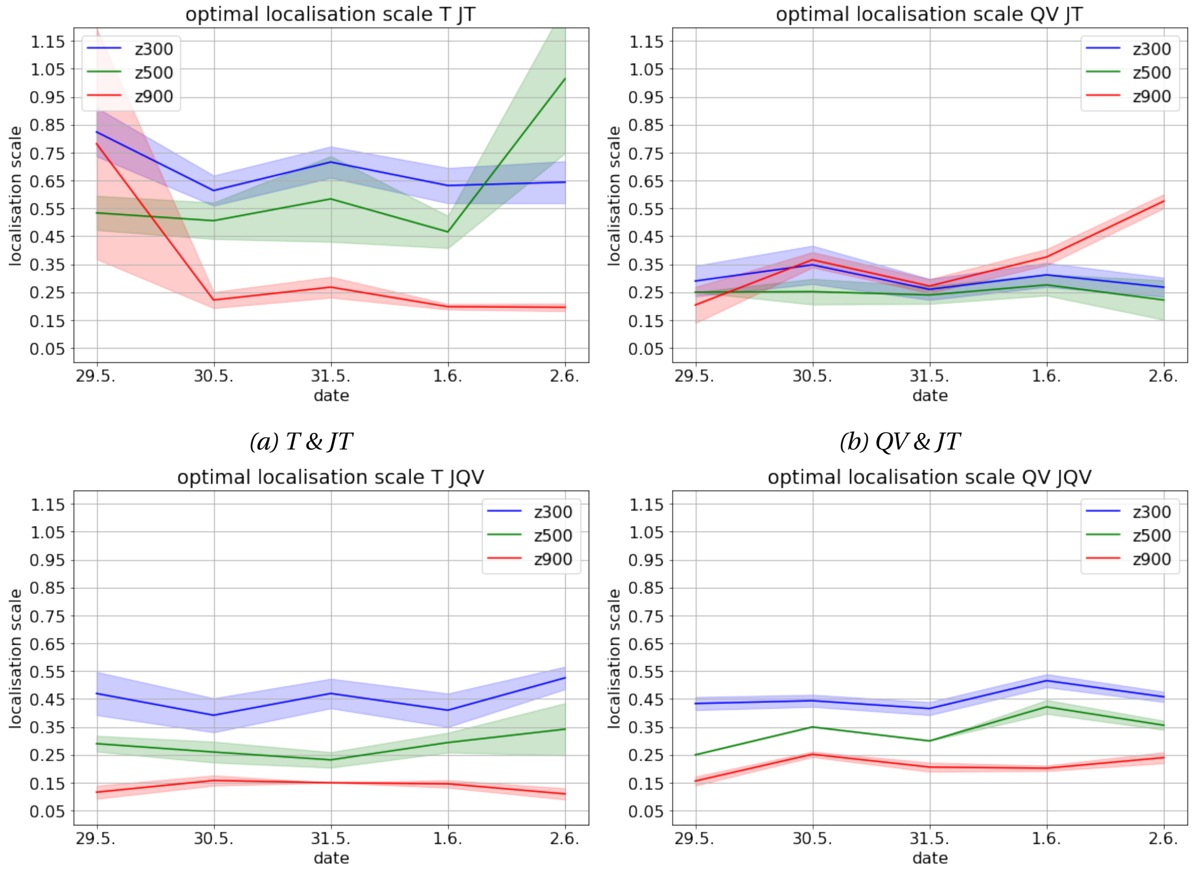
(d) QV & JQV

Figure 10: Comparison between localisation scale (x-axis) and relative improvement to 40-member ensemble (y-axis), for four correlations between temperature (T) and specific humidity (QV) (selfcorr. top-left, bottom-right; crosscorr.: top-right, bottom-left), in three correlation levels: 900 hPa, 500 hPa, 300 hPa; mean of 5 days and 25 subsamples, shaded areas show the standard deviation of the averaged data, dashed lines mark the maximum improvement

flat so the difference, for example, between a localisation scale of 0.4 and 0.8 is very low, which allows to chose the localisation scale less exactly. In contrast, the curves of the other three correlations show a much clearer peak with a greater in- and decrease before and after it. Therefore the localisation scale used has to be chosen much more carefully. Another point where the temperature selfcorrelation differs from the correlations with specific humidity is in the higher standard deviation of the relative improvement depicted by the shaded areas. This is possible due to the higher deviation of the temperature between different locations and days or weather situations. Figure 10 shows the difference between localisation scales for different correlations and how important it is to choose the right localisation scale as it can result in a highly improved outcome of localisation, but when chosen wrongly it could worsen the localisation drastically.

Variability in time

As the weather and so the atmospheric state varies in time, not only a variable dependent but also a weather situation specific localisation scale should be considered, as different conditions may impact the correlations differently. In order to test how the optimal localisation scale changes with different situations, the optimal localisation scale has been calculated for all five available days of the data. The results can be seen for the correlations of temperature and specific humidity in Figure 11. The optimal localisation scale is shown for each of the five days for three vertical levels with the standard deviation of all 25 subsamples depicted by the shaded areas. It has to be pointed out that a low number of consecutive days is used, therefore a similar weather situation prevails on all days. Nevertheless, it gives a good view of the variability in time.



(a) *T & JT*

(b) *QV & JT*

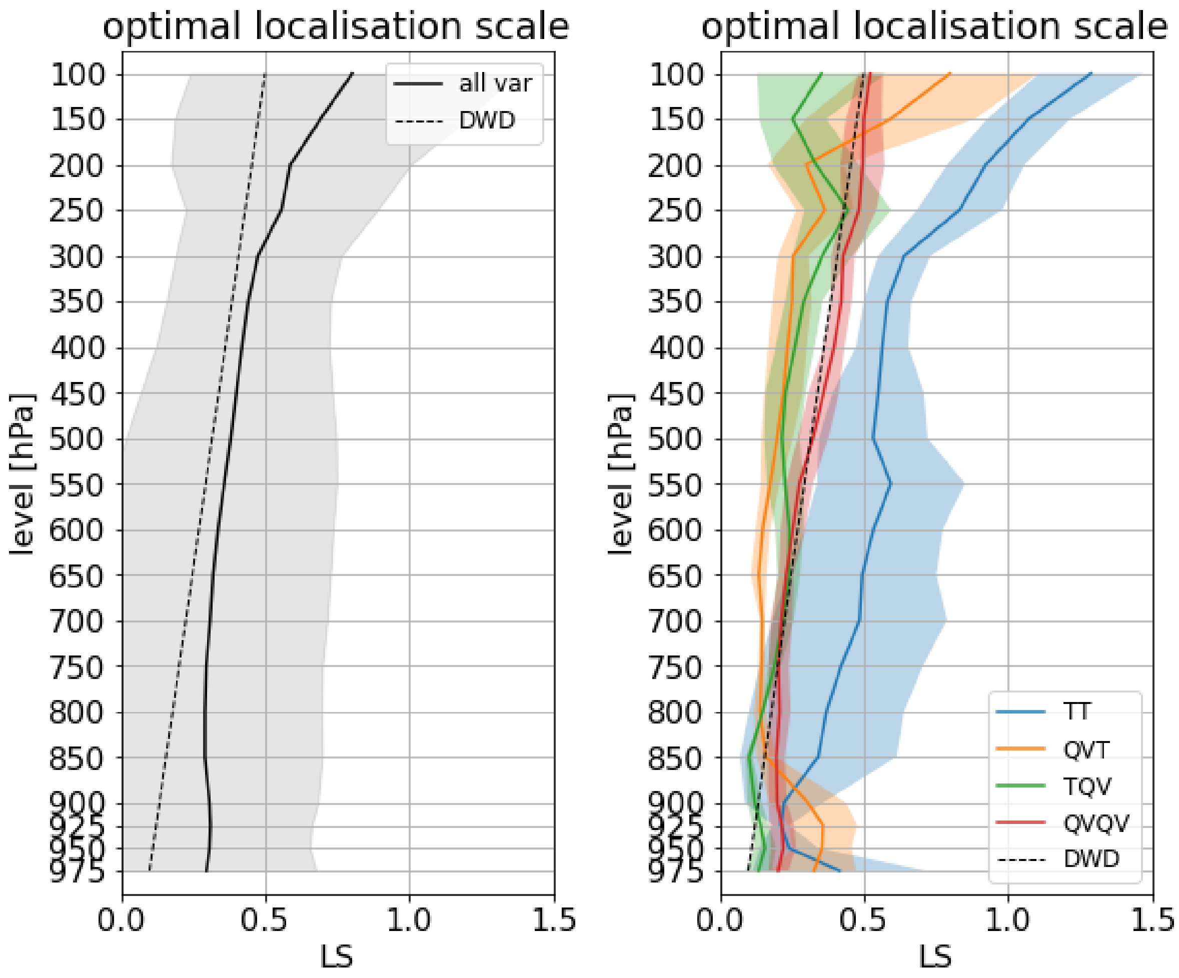
(c) *T & JQV*

(d) *QV & JQV*

Figure 11: Optimal localisation scale for the 15 UTC forecast on five days; shown are correlations between temperature and specific humidity (selfcorr.: top-left, bottom-right; crosscorr.: top-right, bottom-left), in three correlation levels: 900 hPa, 500 hPa, 300 hPa; the shaded areas show the standard deviation of all 25 subsamples

The variability between different days is low as it mostly differs with a deviation of about ±0,1. Great exceptions can be seen for the selfcorrelation of temperature (Fig.11a). However as pointed out previously in Figure 10a the difference in the relative improvement between certain localisation scales for the selfcorrelation of T is low, also the standard deviation for the outliers is much higher than for the other days which is explained by a very flat curve in Figure 10a. Therefore, choosing a localisation scale only depending on the correlation and not the weather situation can be done without worsening the result of the localisation. Another minor point that can be seen in Figure 11, is the increase of the localisation scale with height. Therefore, for low near-surface levels the localisation scale should be rather small and for higher tropospheric levels the Gaspari Cohn function should be broad. This meets with the way the DWD chooses the localisation scale in its data assimilation, by linearly increasing it with height (see sec.2.2.3).

Because localisation is not only done for three model levels rather for all model levels, the optimal localisation scale is calculated for all twenty levels of the 1000-member ensemble. The results are shown in Figure 12a for the non-variable dependent function and in Figure 12b for the variable dependent function of the correlations of temperature and specific humidity. The non-variable dependent localisation scales are averaged over the self- and crosscorrelations of T, QV, QHYDRO, U, V, W (see enumeration in 4.1). As a comparison, the localisation scales of the DWD like function are shown as dotted line.



(a) non-variable dependent

(b) variable dependent

Figure 12: optimal localisation scale at 20 correlation levels for a non-variable dependent function averaged over parameters: T, QV, QHYDRO, U, V, W, and for four different correlations (T & T, QV & T, T & QV, QV & QV), mean of 5 days and 25 subsamples, shaded areas show the standard deviation

The non-variable dependent has slightly higher localisation scales for the middle atmospheric levels (750hPa - 350hPa) than the DWD and also takes the same linear form. Bigger differences can be seen for high and low near-surface levels. In these areas, the optimal localisation scales are larger than the DWD, which means a wider Gaspari Cohn function is applied. Interesting is that in contrast to all other areas the optimal localisation scales in the bottom levels are almost constant with a value of about 0.3. This difference to the localisation scales of the DWD in the top and bottom levels indicates that in this area an improvement of the localisation is likely achievable. Another important instance is the high standard deviation of the non-variable dependent localisation scales, shown as grey shaded area. This indicates the high variability between different sorts of correlations depending on the correlated parameters and also the variability through changing weather conditions of some parameters. It is an indication that by using variable dependent localisation scales, the quality and especially the consistency can be improved, as the variability of the localisation scales for variable dependent localisation is smaller, as seen in the coloured shaded areas in figure 13b.

Looking at the variable dependent localisation scale (Fig.13b) the selfcorrelation of the temperature (TT) stands out with generally higher localisation scales than the three other correlations. The only exception is in the area between 975hPa - 850hPa where the correlation of specific humidity to temperature (QVT) has higher localisation scales and the temperature selfcorrelation is close to the other correlations. This bulge in the curves may be related to changes in the correlations of temperature due to the cooling effects of precipitation. This would also explain the higher standard deviation in this area as this would be caused by the variability of situations with rainfall, either geographical or in terms of time. In contrast to the selfcorrelation of temperature, the other three correlations are very similar to the DWD. The crosscorrelations (TQV, QVT) have except for the highest and lowest levels, slightly lower localisation scales than the DWD and therefore have a narrower Gaspari Cohn function. The selfcorrelation of humidity shows nearly the same shape except for also slightly higher localisation scales in the near-surface levels. This tendency to higher localisation scales than the DWD in the bottom levels, which is shown for both forms of optimal localisation scales, is one important solution to optimising the currently used method of Gaspari Cohn localisation.

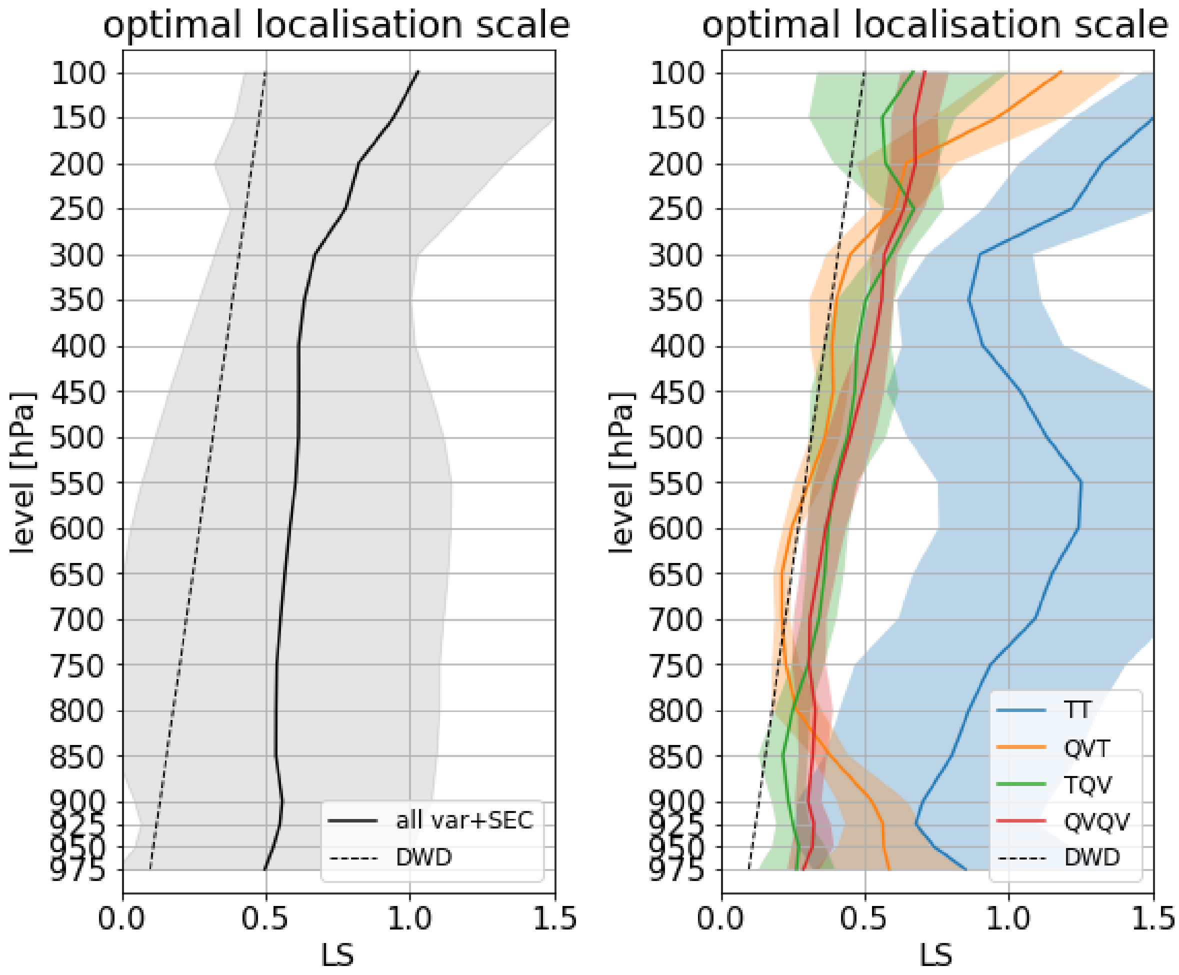
The paper of Destouches et al. (2021), where optimal localisation scales for different cases of weather situations are sought, shows very similar results for the selfcorrelations of temperature and specific humidity, with localisation scales around 0.3 for humidity and slightly higher ones for the temperature (Destouches et al., 2021, Fig.11). The temperature seems to have higher variability. This can also be seen in Figure 12b in this thesis which confirms that the localisation scales of temperature are more dependent on the weather situation than the localisation scales of humidity.

4.1.1 Gaspari Cohn + SEC

As previously shown, the optimal Gaspri Cohn function delivers more insight into how to optimise localisation. But as there are different possible approaches to the correction of the error covariance matrix, a combination of a distance-based and a statistical approach seems to be a good way to improve localisation even further. To explore this idea the optimal Gaspari Cohn is combined with the statistical method of the Sampling Error Correction (SEC)(described in section 2.2.4).
The combination is done by firstly applying the SEC to the data and then finding the optimal localisation scale of the Gaspari Cohn for the sampling error corrected data. This data is then localised with the Gaspari Cohn function defined by the new localisation scale.

The newfound optimal localisation scales for all vertical levels can be seen in Figure 13. The results are shown for a non-variable dependent approach (Fig.13a) and a variable dependent function for the correlations of temperature (T) and specific humidity (QV) (Fig. 13b).

The SEC shifts the optimal localisation scale to higher values compared to when it is not used. This may be due to the effects of the sampling error correction because the spurious correlations are reduced before the Gaspari Cohn function is applied and therefore the risk of getting false information further away from the correlation level is reduced. This allows using a broader function. The shapes of the curves are very similar to the ones without the SEC because the SEC only increases the localisation scale by a certain factor but does not change the information of how different correlations or areas have to be localised. A minor negative fact is that by using the SEC not only the optimal locali-



(a) non-variable dependent

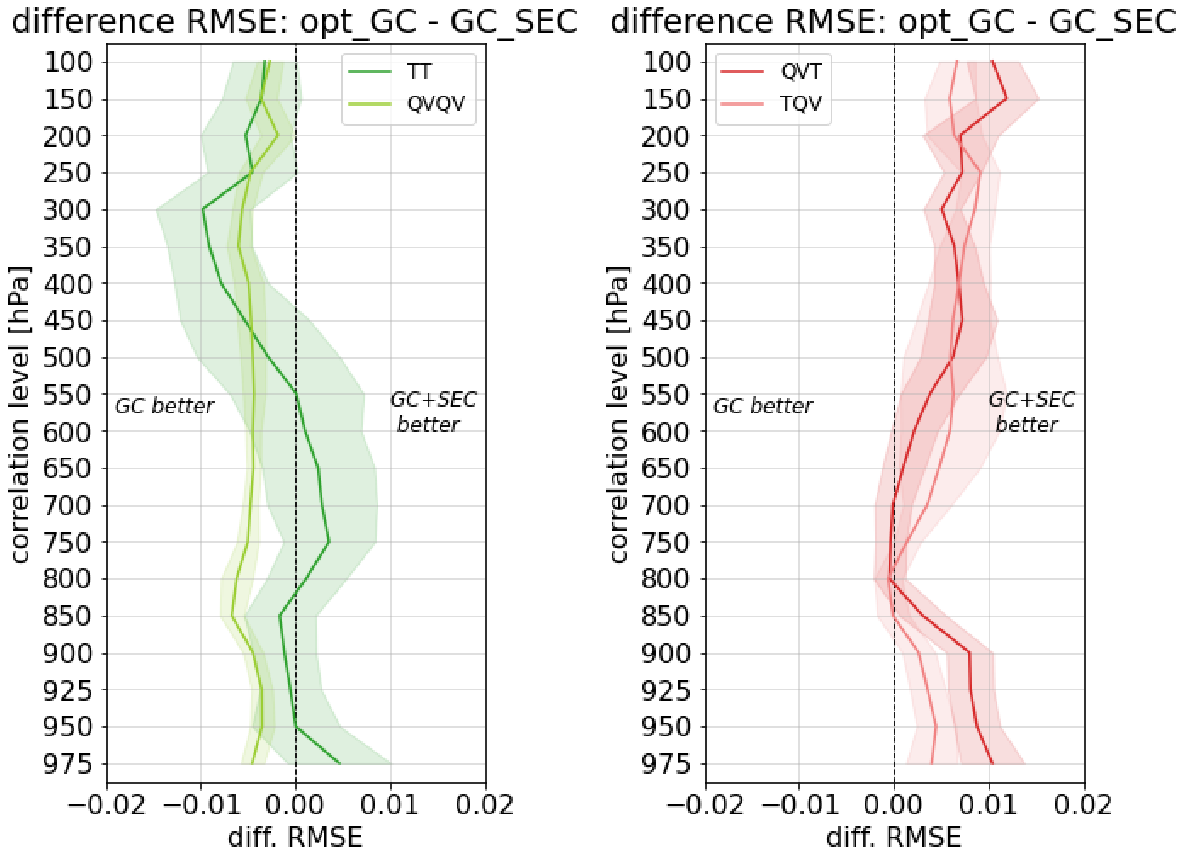
(b) variable dependent

Figure 13: optimal localisation scale for GC + SEC at 20 correlation levels a non-variable dependent function averaged over parameters: T, QV, QHYDRO, U, V, W and for four different correlations (T & T, QV & T, T & QV, QV & QV), mean of 5 days and 25 subsamples, shaded areas show the standard deviation

sation scale but also the variability depicted by the shaded areas increases, which makes it difficult to prove if the optimal localisation scale performs well for every type of weather situation. Although it is only a minor problem, as seen in Figures 9 & 10, the higher the localisation scale gets the smaller the difference of the improvement between neighbouring localisation scales gets.

As the optimal localisation scales are now known, it is interesting to know if the combination of the Gaspari Cohn function and sampling error correction achieves better results than only the optimal Gaspari Cohn function. Figure 14 shows the difference between the RMSE of the optimal Gaspari Cohn function and the Gaspari Cohn function plus SEC. Shown are the self- and crosscorrelations of temperature and specific humidity. Positive results mean that the combination of Gaspari Cohn function and SEC works better, and vice versa negative results mean that the optimal Gaspari Cohn function without SEC achieves better results.

The combination of both methods works better for crosscorrelations (Fig.14b) of all levels except for the area between 850hPa - 700hPa where there is no difference. In contrast to that is the result of the selfcorrelations (Fig.14a). Here the use of the sampling error correction decreases the quality of



(a) selfcorrelations

(b) crosscorrelations

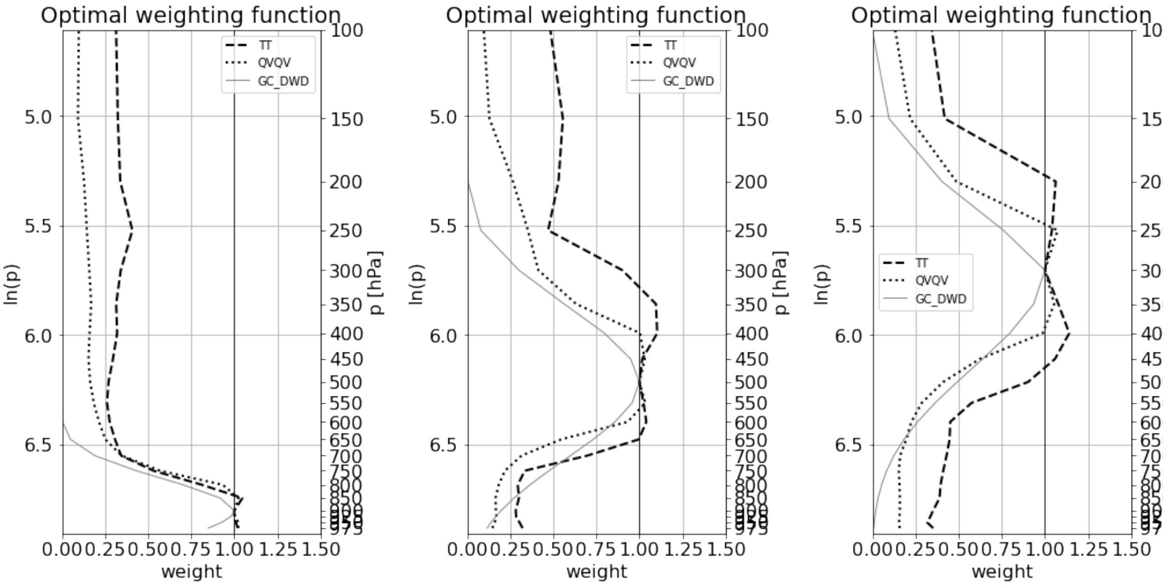
Figure 14: Difference between RMSE of perfect Gaspari Cohn function and perfect Gaspari Cohn function for sampling error corrected correlations, shown are correlations between temperature and specific humidity: selfcorrelations (TT, QVQV); crosscorrelations (TQV, QVT) as mean of 5 days and 25 subsamples, shaded areas are the standard deviation

the localisations, with an exception in the correlation of temperature with itself in the mid-levels. As these results show clear tendencies that the combination of Gaspari Cohn function and SEC works better only for crosscorrelations, it has to be pointed out that the differences shown in Figure 14 are very low with values between -0,01 and 0,01. This means there is a very low improvement, but it nevertheless shows the theoretical difference between the localisations of different types of correlations. It has to be considered that the assumptions in the calculation of the SEC tables are chosen generally. The results could possibly be enhanced when the SEC tables are computed with more specific information regarding the distribution of the correlations.

4.2 Optimal weighting function

After optimising the Gaspari Cohn localisation by using variable dependent localisation scales and combining it with the Sampling Error Correction, a look at the localisation function itself is done. The goal is to see if a Gaussian-shaped function, like the Gaspari Cohn, is the best approach or if the function can be improved furthermore. To explore this the perfect weighting function to correct the correlations of a 40-member ensemble to the correlations of a 1000-member ensemble is calculated. This is done by dividing both of them as described in section 3.2. The averaged optimal weighting function is used to explore its shape in comparison to the Gaspari Cohn function, to learn about possible flaws of a Gaussian-shaped function, and directly as a function for distance-based localisation.

Figure 15 shows the optimal weighting function for the selfcorrelations of temperature (TT) and specific humidity (QV). Shown are three levels (900hPa, 500hPa, 300hPa) depicting the low, middle and high troposphere. The weighting functions are compared to the respective localisation function of DWD.



(a) z=900hPa

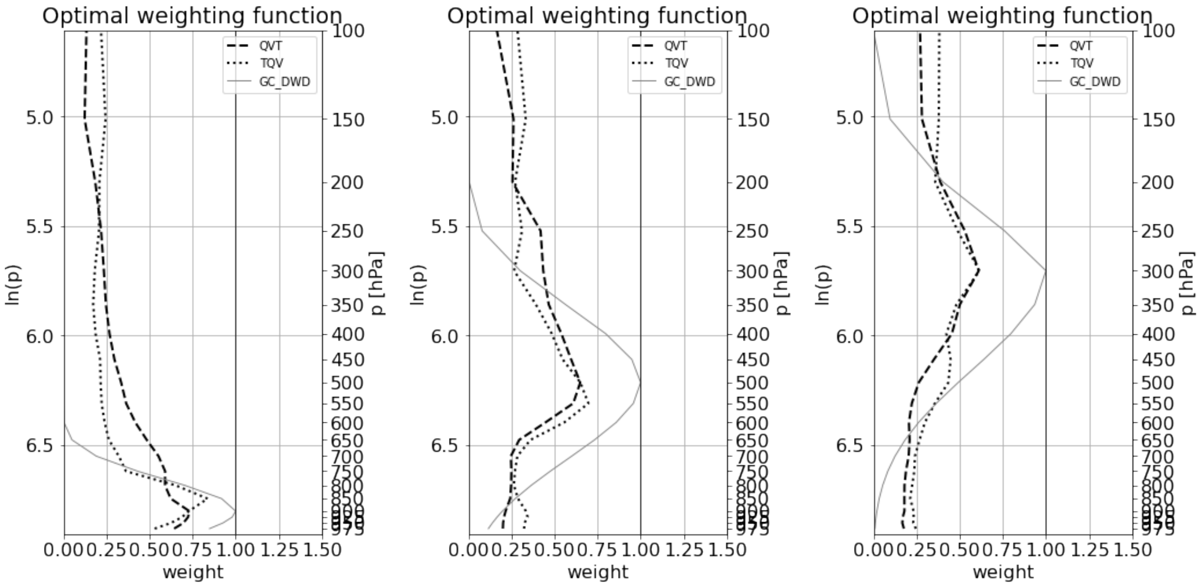
(b) z=500hPa

(c) z = 300hPa

Figure 15: Mean optimal weighting function of selfcorrelations of temperature (dashed) and specific humidity (dotted), in three correlation levels: (a) 900 hPa, (b) 500 hPa, (c), 300 hPa; compared to the GC of the DWD (grey line)

It can be seen that the weighting function in difference to the Gaspari Cohn function does not have a single peak at the correlation level, but rather a plateau-like peak, in some cases double-peak shape, around the 2-3 neighbouring model levels. This discovery applies to all tested selfcorrelations, not only the two shown here (see Appendix). Therefore, to improve localisation regarding selfcorrelations, a function with a plateau-like peak would increase the quality of the correlations. Despite the different types of peaks, the overall shapes of the DWD function and the optimal weighting are similar, especially in the lower levels. As also seen in the calculation of the optimal localisation scales in Figure 10 it seems that selfcorrelation of temperature needs a broader localisation function than currently used, whereas the breadth for the correlation of humidity to itself is similar to the one already applied. Another minor point is that different to the Gaspari Cohn function the optimal weighting function does not "cut off" (set to zero) the correlation at any point. It has to be considered that even large ensemble data has some sort of statistical noise and this may be the reason why there is no cutting off. Because no matter how far a point is away from the correlation level, there will always be a signal, although this may have no physical connection. Also does the "cut-off" reduce the computational resources in data assimilation (e.g. LETKF), because not all grid points have to be considered in the calculations.

The optimal weighting for the crosscorrelations (QVT, TQV) between temperature (T) and specific humidity (QV) is shown in Figure 16. As done previously, they are compared to the Gaspari Cohn function of the DWD.



(a) z=900hPa *(b) z=500hPa* *(c) z = 300hPa*

Figure 16: Mean optimal weighting function of crosscorrelations between temperature and specific humidity: QVT (dashed), TQV (dotted); in three correlation levels: (a) 900 hPa, (b) 500 hPa, (c), 300 hPa, (QVT, TQV); compared to the GC of the DWD (grey line)

The main difference between the optimal weighting of crosscorrelations and selfcorrelations, or the Gaspari Cohn function respectively, is the peak not being at one but rather between 0.5 and 0.75. This applies, same as the plateau peak of the selfcorrelations, to all tested crosscorrelations. Therefore, as the shape of the weightings is also Gaussian-like, it concludes that the Gaspari Cohn function already depicts a close to reality function, but it could be improved by shifting the peak. In addition, a variable dependent peak could be a good opportunity to optimise the Gaspari Cohn localisation. Despite the problem regarding the peak, Figure 16 confirms that the Gaspari Cohn function of the DWD is slightly broader than an optimal function would be, as already seen in the calculations of the optimal localisation scales for crosscorrelations (see Figure 10). The reason the optimal weighting is not cutting off any correlations is probably the noise of the used data as explained previously. Following Houtekamer and Mitchell (1998), the expected noise for the case that the true correlation is zero can be calculated with the ensemble size **N**

$$noise = \frac{1}{\sqrt{N}} \tag{17}$$

With this equation, the noise for the case of true correlation being zero, for the 40-member ensemble is 0,158 and 0,032 for the 1000-member ensemble. Following equation 16, this leads to an error of the optimal weighting, for the case of the perfect weight being zero, of 0,2.
This meets with the optimal weightings for correlations far away from the correlation level, where the weightings are mostly lower than 0,25. Therefore, a "cut-off" in these areas, such as the Gaspari Cohn has, can be justified with the noise of the data.

Summarising, the optimal weighting function brings important information on how a localisation function should look like and how the currently applied functions could be adapted to improve the localisation quality. These changes mainly regard the way the peak of the function is chosen. A differentiation between cross- and selfcorrelations is appropriate as they show big differences with a plateau-like peak for selfcorrelations and a damped peak for crosscorrelations. These changes could be applied to the Gaspari Cohn function. Another option would be to use the optimal weighting function directly for localisation. Therefore, it can be chosen between a variable dependent function or, to simplify calculations, the usage of just two functions, one for selfcorrelations and the other for crosscorrelations. More not discussed, optimal weighting functions can be seen in the Appendix.

4.2.1 Quality of localisation

After exploring the optimised distance-based localisation methods (optimal Gaspari Cohn, optimal weighting function). The quality of localisations using these methods is examined and compared to not localised correlations and to localisations done with the method utilised by the DWD.

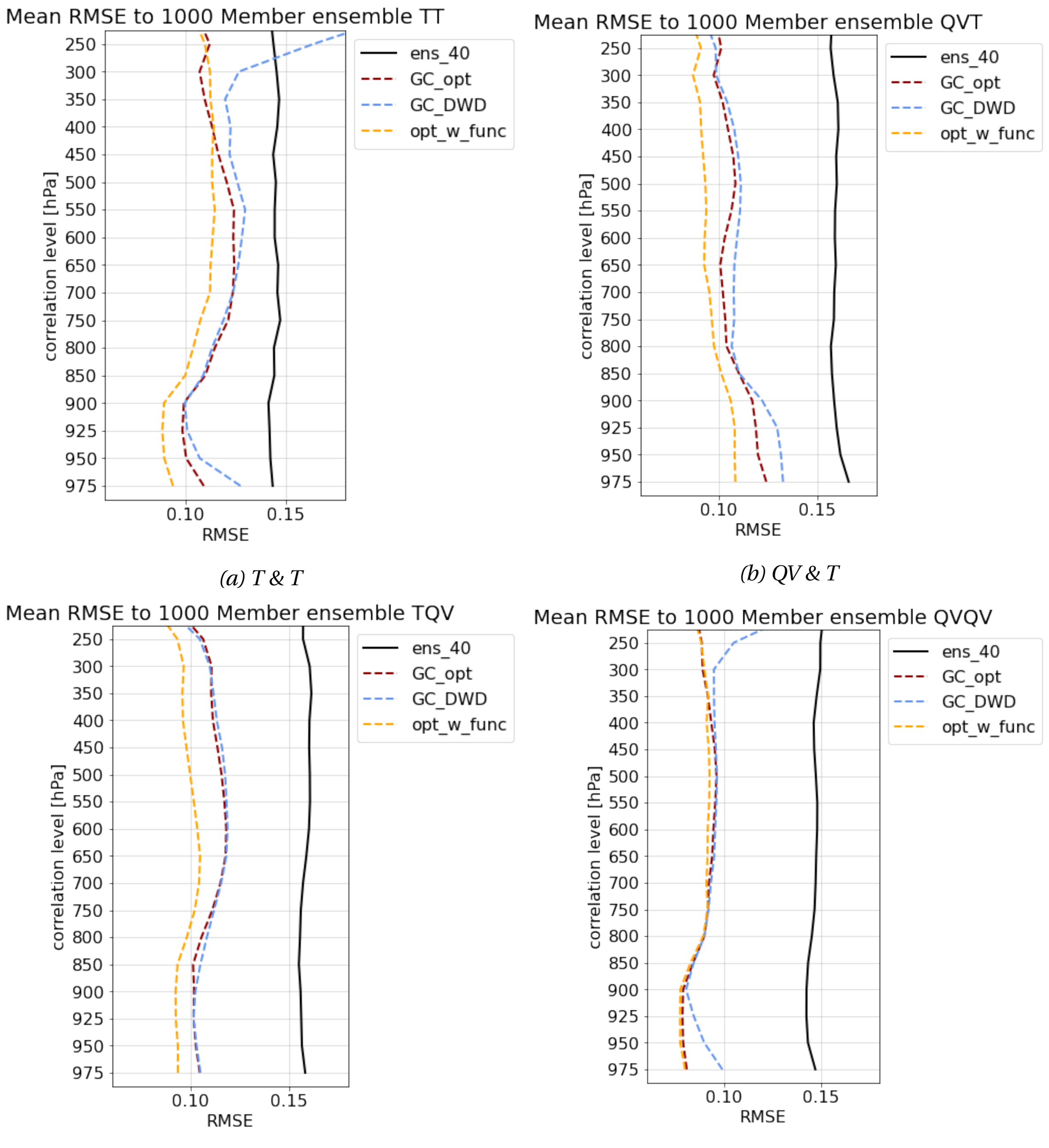
Therefore, localisations are calculated using the previously described methods. They are computed for every available correlation at every level on all 25 possible 40 member subsamples drawn from the 1000-member ensemble at 5 days. The mean RMSE between the 40-member ensemble and the localised correlations of all vertical profiles in the domain is calculated and averaged. The resulting mean RMSE are compared to the RMSE of the correlations of the 40-member ensemble to the 1000-member ensemble.

Figure 17 shows the comparison of applying the optimal weighting function directly (opt_w_func), using a variable dependent optimised Gaspari Cohn function (GC_opt) and the Gaspari Cohn function of the DWD (GC_DWD). The RMSE of the 40-member ensemble (ens_40) show the minimum benchmark, so if the RMSE is higher than this, the used method would even worsen the outcome of the localisation. The combination of the optimised GC and the Sampling Error Correction is not shown because of the small difference to optimised GC as seen in Figure 14. Shown are the correlations of temperature (T) and specific humidity (QV).

Firstly, all three displayed methods lower the RMSE significantly and therefore improve the correlations. As expected, the optimal weighting function performs better than both other methods. This is because the optimal weighting function depicts the closest to perfect localisation function. The optimised Gaspari Cohn function and the DWD Gaspari Cohn function have very similar results, except for the lower levels where in three of four shown cases the optimised Gaspari Cohn function brings better results. This reflects the conclusion of Figure 12 that the DWD chooses its localisation scales near the surface too small. The wider optimised Gaspari Cohn function works better, but can not compete with the optimal weighting function. The only exception is the selfcorrelation of specific humidity (Fig.17d) where all methods have a similar RMSE through the vertical profile, except for the DWD in the lower and top levels. Here, the best results are generally achieved with an RMSE under 0.10 over the whole profile.

While the 40-member ensemble has nearly constant RMSE, the localisations work better in the lower parts than in the middle atmosphere. This is depicted by a slightly convex shape of the curves. It also seems that the DWD localisation works better for crosscorrelations than for selfcorrelations. This can be seen especially in the correlation of temperature to itself (Fig.17a). The worse results of the DWD in comparison to the optimised Gaspari Cohn function emerge from the use of too small localisation scales (see Fig. 12b). The worse results of DWD for levels above 300 hPa are due to the applied damping, this lowers the peak of the Gaspari Cohn function below one which falsely depicts the weighting of selfcorrelations, because a correlation with itself at the correlation level is always one.

Summarising, the optimisation of the current methods brings improvement, especially the use of an



(a) T & T

(b) QV & T

(c) T & QV

(d) QV & QV

Figure 17: Comparison of RMSE between localisation and 1000-member ensemble at 20 correlation levels for three different Methods (optimised GC, GC used by DWD, optimal weighting function) and the mean RMSE to the ***correlations of 40-member ensembles****, show are four correlations: (a) T & T, (b) QV & T, (c) T & QV, (d) QV & QV*

optimal weighting function. But also splitting the Gaspari Cohn function into multiple variable dependent functions improves the localisation generally and notably for selfcorrelations.

4.3 Random forest

After exploring and optimising distance-based methods, a new approach for localisation that does not rely on a single weighting function is tested. Machine learning has a huge variety of methods that all try to learn about the connections between parameters to use them to predict wanted parameters. The method chosen to explore localisation with machine learning in this thesis is the random forest (see section 2.3). For all further results, a random forest with 10 trees is applied. The random forest is always tested on a single 40 member subset from one day. The training data is built by single subsets of the four other available days. Correlations of the 1000-member ensemble serve as target values as the goal is to predict correlations as close as possible to the truth. In the following the different steps taken to explore and find an optimal working random forest are shown.

4.3.1 First try

As a first test of the random forest, a simple configuration, the basis random forest, is chosen (see sec. 3.4.1). It only exploits two predictors. The first predictor is always the 40-member correlation of the respective correlation. The second predictor is the parameter correlated from every level in the profile with the parameter at the correlation level. These two predictors are chosen because they are believed to bring the most important information, as the 40-member correlation is the correlation to be corrected, and the second predictor is the physical parameter with the highest impact on the correlation because it is directly connected to it on every level.

The basis random forest is tested on four correlations of temperature and specific humidity in three levels (900hPa, 500hPa, 300hPa). Figure 18 shows the relative improvement of the RMSE between the localised and the 1000-member correlation relative to the RMSE of the 40-member correlation, averaged over three levels and 5 test days. It is compared to the previously discussed methods.

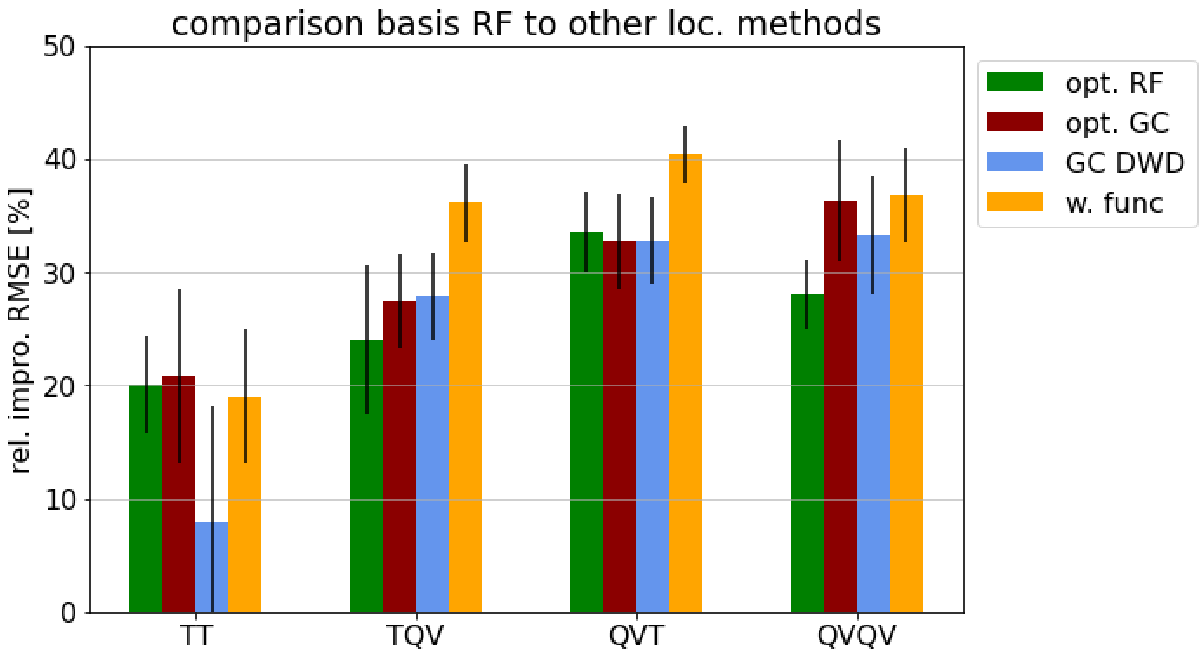


Figure 18: Comparison of relative improvement to 40-member ensemble between basis Random forest, optimal variable dependent GC, DWD GC and optimal weighting function improvement to 40-member ensemble. Mean of 5 days and 3 correlation levels (900hPa, 500hPa, 300hPa); the error bars show the standard deviation of the averaged data

Although the basis random forest is a simple approach, the results show that it is a well-working method to correct the 40-member ensemble correlations, as for all four correlations improvements from 20% to about 33% are achieved. On the other hand in comparison to the other techniques, the random forest underachieves, except for the selfcorrelation of temperature (TT), where it is nearly as good as the optimal Gaspari Cohn function. It also works well for the correlation of specific humidity with temperature (QVT) where it is on the same level as the localisation with the Gaspari Cohn function, only the optimal weighting function performs better. These results show that the random forest may be a good possibility in order to improve localisation, but there is also some room for improvement, as there are many possibilities in the configuration of the random forest.

4.3.2 Optimise random forest

In the following, two approaches of tuning the random forest are discussed. These methods try to find the right predictors and optimise the training data. Both methods are described in detail in Section 3.4.2.

Parameter test

The first approach is to include more predictors in the random forest to get more physical information of the data and therefore better predictions. The difficulty of choosing the right predictors is that not all predictors improve the quality of a random forest. Some even lead to a decrease in quality, because they bring information to the random forest that may be linked to the target value statistically but have no physical core. As it is not easy to theoretically filter which predictors bring useful information a so-called parameter test was developed for this thesis to test all available parameters on their improvement compared to the basis random forest.
The parameter test works by applying a third additional predictor to the random forest and then calculating the difference between the basis random forest and the new one. This is done for multiple different predictors (see tab.1 & 2).

The averaged results of the parameter test for the correlations of temperature and humidity in three levels (900hPa, 500hPa, 300hPa) can be seen in Figure 19. Shown is the difference in the relative improvement of the RMSE between the basis random forest and the random forest with three predictors. The parameters are sorted left to right from the most improving to the least improving. The shortcuts of the parameters are explained in tables 1 & 2.

From eleven tested parameters, just four bring improvement. They are the hydrometeor mixing ratio (qhydro), vertical wind (w), hourly precipitation (prec) and radar reflectivity (dbz). These parameters bring information about convective events, furthermore about rainfall and clouds. As a high vertical wind indicates convection in the atmosphere, a high hydrometeor mixing ratio indicates clouds and precipitation and also the radar reflectivity brings information about rain and snow. By combining these parameters it is possible to get a very clear overview of where in the domain rain or furthermore thunderstorms happen. Therefore, these parameters are called convective parameters in the following. It shows that it is important for the random forest to be able to differentiate between varying weather situations like rain, thunderstorms, or just cloudiness. With this knowledge, it concludes that there has to be a difference in the vertical correlations for convective and non-convective events. The second major point derived from the results in Figure 19 is that many parameters used as predictors decrease the quality of the random forest, therefore it is essential to explore the available predictors in order to choose the improving ones. This big number of bad predictors, on the other hand, could be an indicator for an unstable machine learning method, because a perfect stable technique would ignore information that results in a decrease in prediction quality. However, the reason for the

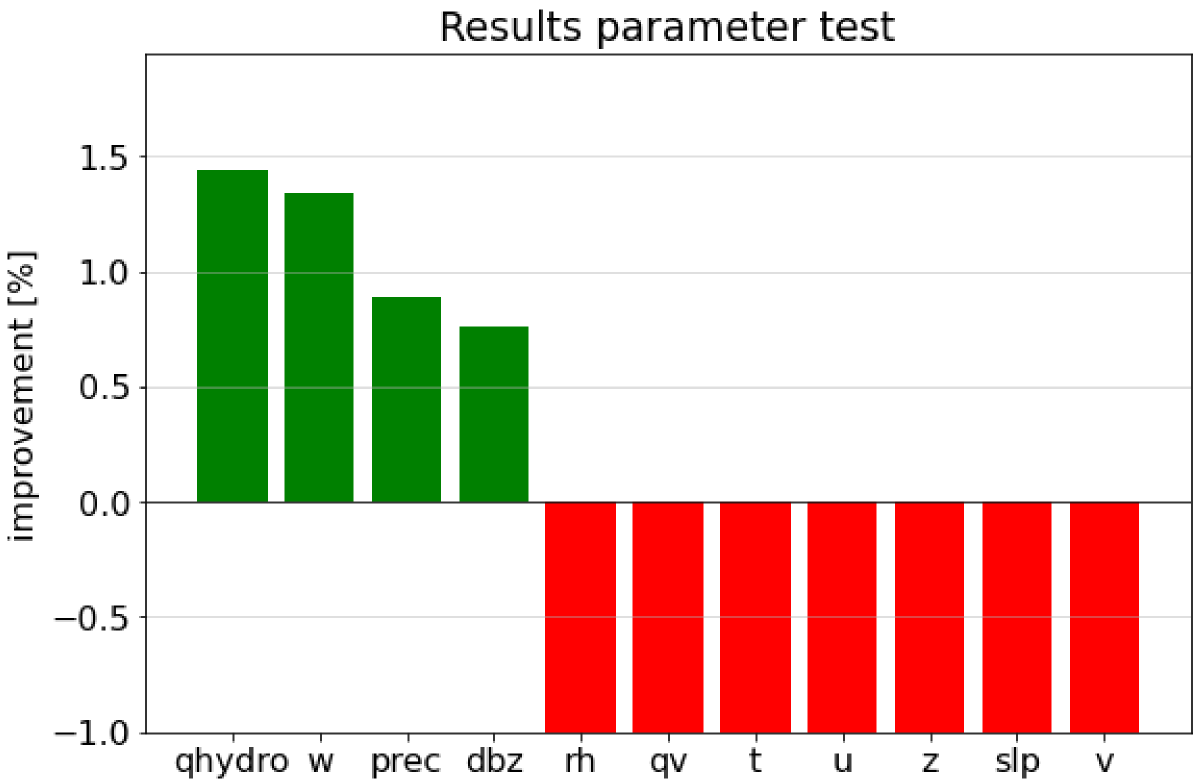


Figure 19: Change of relative improvement, a third predictor (detailed info in Tab.1 & Tab.2) is added to the basis random forest, shown is the mean of self- and crosscorrelations between temperature and specific humidity, at three correlation levels (900hPa, 500hPa, 300hPa); mean of 5 days

bad predictors could not only be an unstable random forest but a too small size of data used to train so that the flaws in the data get much more weight than in a bigger sized data. As the exploration of the stability of the random forest and the data would exceed the volume of this thesis, it is not further studied.

With the results of the parameter test, the basis random forest is optimised by adding the four convective parameters as predictors. The now optimised random forest consists of the following six predictors:

1. 40 member correlation
2. parameter correlated from every level with correlation level
3. mixing ration of hydrometeors
4. vertical wind
5. hourly precipitation
6. radar reflectivity

Data filtering

As the optimal combination of predictors is now known, the next step to optimise the random forest is to optimise the training data. The more accurate the training data is, the better the quality of the built decision trees is. The idea is to filter the data before training the random forest to exclude values or grid points that deliver information highly diverging from the assumed true information. To clean the data from harmful values, the same filtering as for the optimal weighting function is applied (see section 3.2). This filter excludes all grid points where the deviation of the 1000-member and 40-member correlations is below or above the 2% or 98% percentile of all weights. The filtered data is utilised to train the random forest.

Figure 20 shows the relative improvement compared to the 40-member ensemble correlation for different configurations of the random forest: first the basis random forest, secondly the forest after the parameter test with the added convective predictors and thirdly the random forest after the parameter test and with filtered training data. The results are shown for four correlations (TT, QVT, TQV, QVQ) averaged over three levels and 5 days.

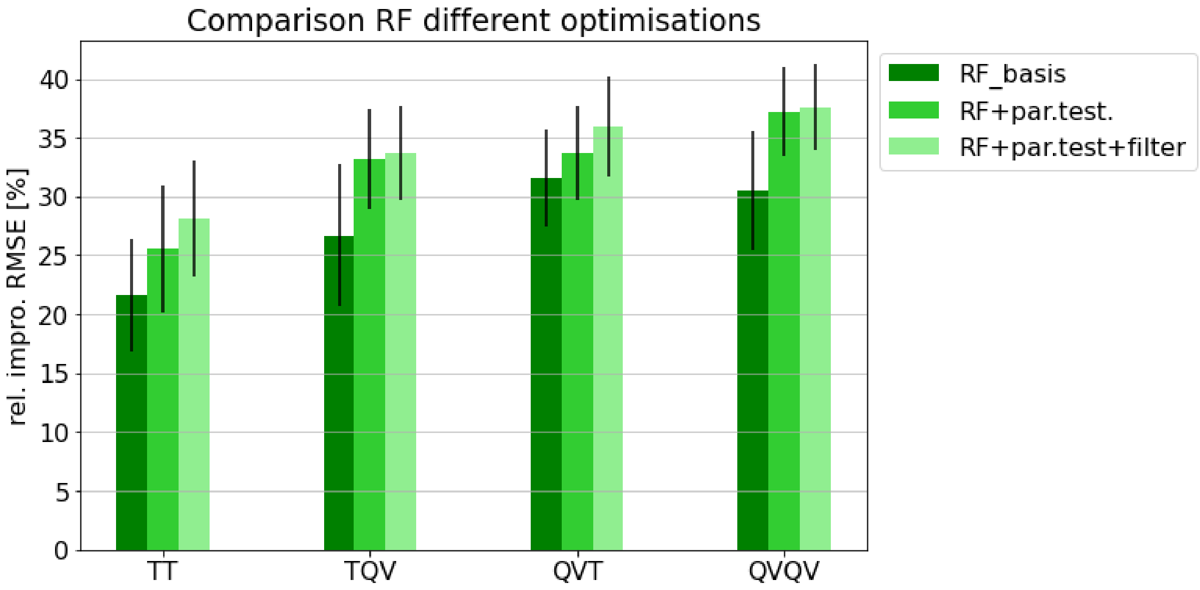


Figure 20: Relative improvement of three versions of random forest, mean of 5 days and 3 Levels (900hPa, 500hPa, 300hPa), the error bars show the standard deviation of the averaged data

The results show that both methods improve the random forest clearly with an increase of about 5% - 7%. Which method works better seems to depend on the type of correlation. While the filtering of the data brings little improvement for the correlations of specific humidity (TQV, QVQV), it brings the same improvement as the choice of predictors for temperature correlations (TT, QVT). Although the optimisation methods shown here are rather simple, they still show that the random forest has much more potential than shown in this thesis and can be improved further with more exploration of possible optimisation methods.

4.3.3 Endresults

As an optimised random forest is explored, it can be compared to the previously discussed localisation methods of the Gaspari Cohn and the optimal weighting function. In Figure 21 the comparison between the optimised random forest (opt. RF), the optimal Gaspari Cohn (opt. GC), the localisation of the DWD (GC DWD), and the optimal weighting function (w. func) is displayed. The relative improvement compared to the 40-member ensemble for the correlations of temperature (T) and specific humidity (QV) is shown.

Different to the basis random forest (see Fig.18) the optimised random forest beats the method of the DWD for all four correlations by 2-3% except for the selfcorrelation of temperature (TT) where the random forest is about 15% better. The optimised random forest also has better results than the optimised Gaspari Cohn function except for the correlation of humidity to itself, where they have

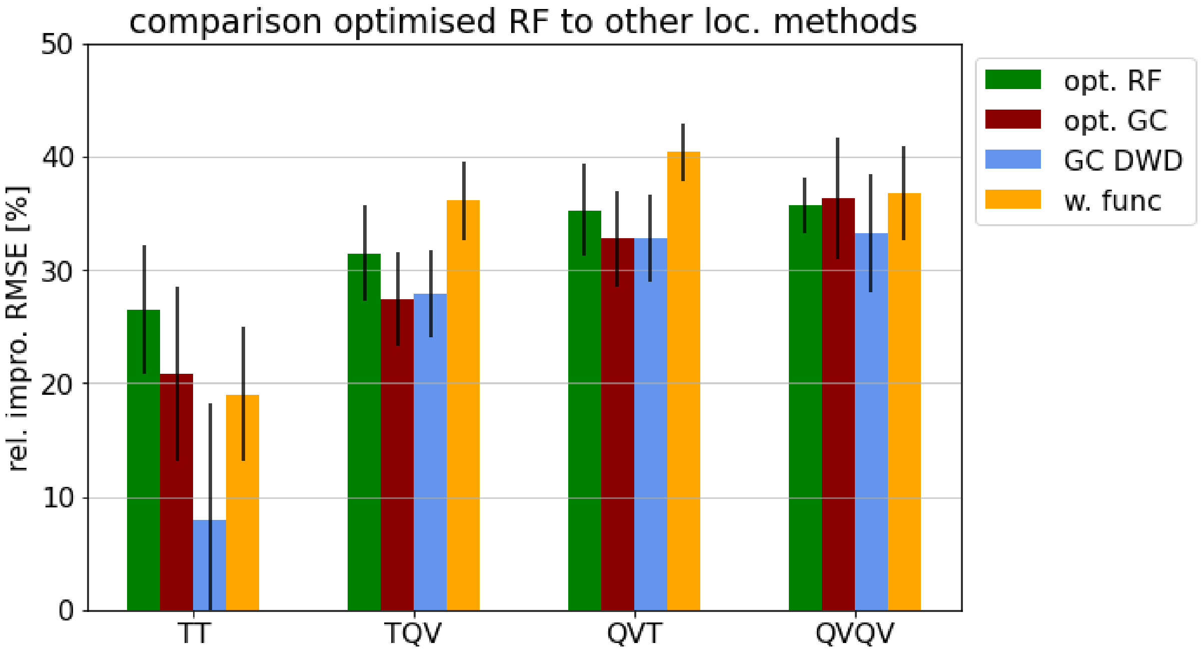


Figure 21: Comparison of relative improvement to 40-member ensemble between optimised Random forest, optimal variable dependent GC, DWD GC and optimal weighting function improvement to 40-member ensemble. Mean of 5 days and 3 correlation levels (900hPa, 500hPa, 300hPa), the error bars show the standard deviation of the averaged data

nearly equal results. Only the optimal weighting function performs mostly better except the temperature selfcorrelation, where all three distance-based methods perform significantly worse than for the other correlations.
These results show that the random forest as a localisation method has great potential as it can not only compete with current methods, even when they are optimised but performs even better in most cases. With the knowledge that the configuration of the random forest in this thesis is a rather simple one, it is definitely possible to increase the quality of localisation with random forest even further. This could be done by exploring more parameters as possible predictors or even creating new variables, especially for the random forest. In addition, an increase or better choice of training data can improve the random forest or the structure of the trees and their building can be tuned in many possible ways. This all could lead to much better random forests but needs a lot of further exploration.

One of the main goals of this thesis is to improve vertical localisation in data assimilation compared to currently used methods. The four methods of this thesis are compared to the method of the DWD, to see how much the previously described optimised or new methods improve the current localisation methods. The methods are the variable dependent optimal Gaspari Cohn function (opt.GC), the optimal Gaspari Cohn function plus the Sampling Error Correction (opt.GC+SEC), the use of a random forest (RF) and the optimal weighting function (opt.w.func.). The relative differences between the RMSE of the DWD localisation and the RMSE of the optimised methods are calculated. Figure 22 shows the relative improvement as the mean of the data used for the random forest calculations. These are the correlations between temperature and humidity (TT, QVT, TQV, QVQV) in three levels (900hPa, 500hPa, 300hPa) for one subsample at 5 days.

The method with the highest improvement is the optimal weighting function with about 9.5%. The second best method is using a random forest for localisation, it achieves an improvement of over 8%. The random forest applied in this thesis has a rather simple configuration and nevertheless achieves

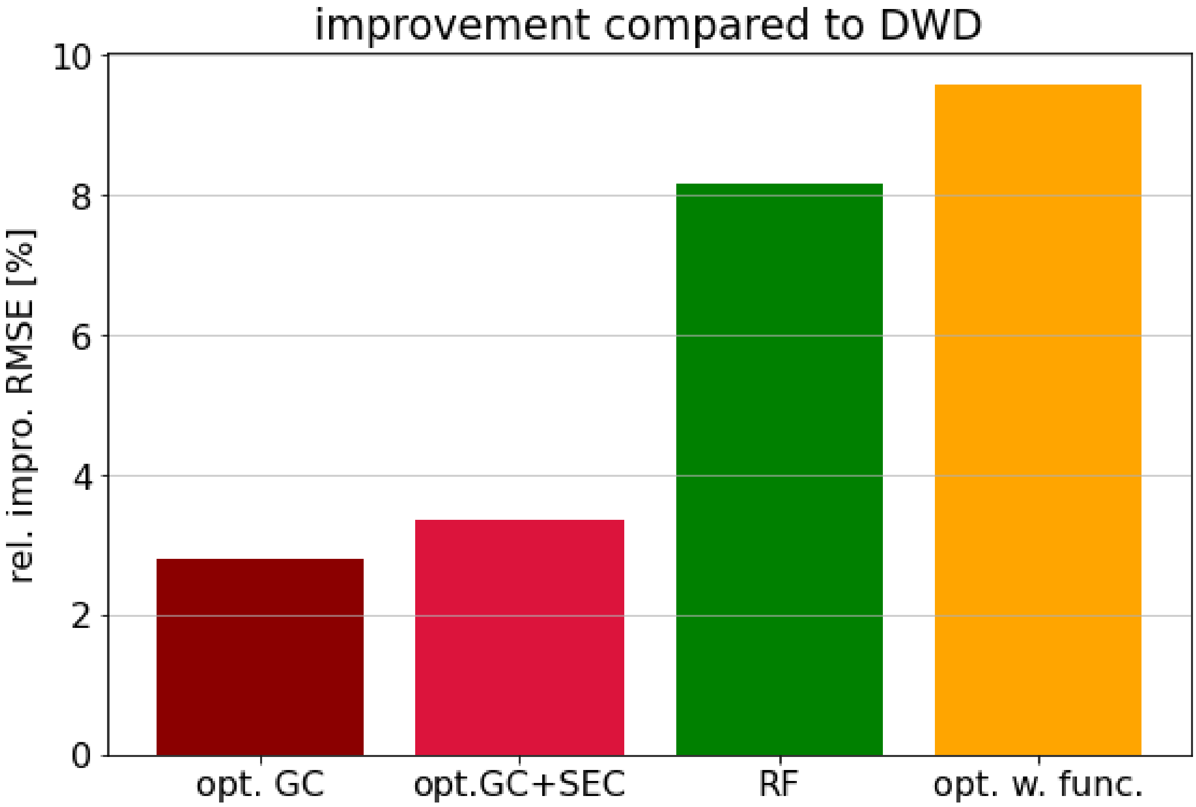


Figure 22: Improvement of localisation methods compared to the DWD; mean of 5 days and 3 correlation levels (900hPa, 500hPa, 300hPa) for the correlations of temperature (T) and specific humidity (QV)

much better results than the DWD and the optimised Gaspari Cohn function. This indicates the possible high potential that random forests or machine learning, in general, has in localisation. The optimised Gaspari Cohn function with and without the SEC, despite being half as good as the newly developed methods, also brings a clear improvement of about 3% compared to the DWD. This shows that even the current used methods can be optimised further when made variable dependent and combined with statistical methods, with the advantage of needing much less computational effort than especially the random forest.

5. Summary and conclusion

Four different approaches of vertical localisation in model space were enhanced, explored, and tested to optimise localisation. To test the methods, a convective-scale 1000-member ensemble with a domain of 200 x 200 points positioned over Germany was used. This ensemble is assumed to show the true forecast error correlations and serves as a target for all localisations. The test data is built by 40-member subsamples drawn from the 1000-member ensemble. Further, all developed optimised localisation methods are compared to the configuration of the Gaspari Cohn function utilised by the DWD, applied to the data of this thesis.

Firstly, the current method of the Gaspari Cohn function is optimised by finding the optimal localisation scale. This is done for a non-variable and variable dependent function. It shows that there are apparent differences in the optimal localisation scale between different correlations and that there is minor variation in the localisation scale between different days. Compared to the DWD, it shows that, especially at the bottom levels, the localisation scale applied by the DWD is chosen too small. This results in the biggest improvement in this area, whereas in the middle levels the DWD and the optimised Gaspari Cohn function are mostly very similar. The use of an optimised variable dependent Gaspari Cohn function leads to an improvement compared to the DWD of about 3%.

Secondly, as a small further development, a combination of a distance-based and a statistical method is explored. Therefore, the variable dependent optimal localisation scale is sought for data corrected by the Sampling Error Correction. The improvement achieved by the combination of these two methods is just slightly better than when using only the optimal Gaspari Cohn function. However, detailed exploration shows that this method only works better for crosscorrelations, therefore to work for both types of correlations, the sampling error correction needs to be adapted. It also shows higher optimal localisation scales, which has the advantage that when a slightly different localisation scale is applied for sampling error corrected data, the result would not decrease as much as without Sampling Error Correction. Furthermore, it has to be considered that the assumptions in the calculation of the SEC tables are chosen generally. It may be possible to improve this method with more specific SEC tables that take the real distributions of the correlations into account.

Thirdly, as the first newly developed method, an optimal weighting function is explored. The goal is to find the function that corrects the correlations of the 40-member ensemble to be the 1000-member correlations. After filtering out the biggest outliers, the averaged shape of this function can be utilised for localisation or even more important for learning about the flaws of the currently used functions.
It shows that there are major differences, between self- and crosscorrelations, regarding the peaks of the functions. As for selfcorrelations, in contrast to the single peak of the Gaspari Cohn function, a plateau-like peak which weights levels near to the correlation point equally would be closer to the perfect weighting function. For crosscorrelations, the single peak seems to be the right choice, but the amplitude of it should be lowered with having the maximum between weights around 0,5 to 0,75. Besides learning from the shape of the function and adapting the currently used functions, it can also be applied directly for localisation. This results in an improvement of around 10% compared to the DWD.

As the last method, a new technique of using a random forest for correcting the correlations is explored as an example for machine learning. For this, different configurations and methods to improve the random forest were tested. The optimising techniques focus on picking the right predictors and cleaning up the training data before using it.

The results show that especially the choice of the right predictors is essential because some parameters will even decrease the quality of the random forest. In the case of this thesis, convective parameters additional to the 40-member correlation and the correlated parameter turn out to be the right choice. This shows that for a random forest it is important to know about rainfall, clouds, and thunderstorms. It indicates that correlations in such situations are different as in dry and clear sky situations and therefore need to be localised differently. The testing of different predictors also shows that the random forest may be unstable towards false information, as some parameters cause terrible results. By using the best predictors and applying the same filtering as for the optimal weighting function, the still rather simple random forest although achieves an improvement of about 8% compared to the DWD. This is better than the optimised versions of the Gaspari Cohn function and slightly worse than the optimal weighting function.

With all the results mentioned previously, it has to be considered that the data used only depicts five consecutive days in summer. Therefore, all methods should be explored for different seasons as well to determine if there is a difference, especially in winter, where for example, inversion in the lower levels plays a bigger role than in summer. Furthermore, the assumption of the 1000-member ensemble as the truth is only an approximation, as, besides the high number of members, the ensemble also has some sort of noise and errors that have to be considered. Furthermore, does this thesis show theoretical approaches to optimise localisation, assuming no restrictions in computational resources or other forms of limitations. This means when using any of the described methods, the existing restrictions depending on the respective data assimilation algorithm have to be considered.

In conclusion, there is potential in improving localisation as it is utilised at the moment. Some minor improvements as using a variable dependent function and combining it with statistical methods could be applied rather easily and would lead to a clear improvement. However, the development of new methods leads to great progress. Especially, the development of the optimal weighting function brings important information that can be used to adapt current localisation functions. This might be a possibility with lesser effort than to apply the function directly. Especially because this thesis only delivers a first insight into how the optimal weighting function looks like. There is probably much more hidden information, and it needs more exploration, particularly differentiating between different weather situations as there seem to be differences in the correlations. Besides this, with the rapid increase in computer performance, machine learning brings great potential to localisation as it does not need a fixed function and rather depends on knowing about physical connections between parameters and correlations as exactly as possible. To use machine learning as a reliable source in operational data assimilation, it needs a lot more exploration and computer resources. In the near future, the best solution may be a hybrid method between different techniques, as it would need less computational power but nevertheless would lead to a clear improvement.

All in all, to improve localisation, it is important to consider many possible methods and learn about their strengths and flaws. Then it is possible to choose the best approach regarding the circumstances it is applied. Besides selecting the best available localisation method, it is even more important to know about the correlations themselves and how they are influenced by localisations. Because only with enough knowledge, it is possible to adapt and develop localisation methods that secure the most truthful initial states for weather predictions.

Acknowledgement

First of all huge thanks to my supervisors Martin Weissman and Tobias Necker, who helped me with great passion whenever they could and for pushing me to work. Without that, I probably would have needed much longer to finish this thesis.
Thanks to my sister for correcting and improving the linguistic part of this thesis.
Lastly, I thank my parents for supplying me with food, a home and emotional support during the last 25 years, without them this thesis would not have been possible.

Bibliography

Anderson, J. L. (2007), Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear Phenomena* **230**(1-2), 99–111. doi:10.1016/j.physd.2006.02.011.

Anderson, J. L. (2012), Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review* **140**(7). doi:10.1175/MWR-D-11-00013.1.

Anderson, J. L., B. Wyman, S. Zhang, and T. Hoar (2005), Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *Journal of the Atmospheric Sciences* **62**(8), 2925–2938. doi:10.1175/JAS3510.1.

Breiman, L. (1996), Bagging predictors. *Machine Learning* **24**(2), 123–140. doi:10.1007/bf00058655.

Breiman, L. (2001), Random Forests. *Machine Learning* **45**(1), 5–32. doi:10.1023/A:1010933404324.

Buehner, M. and A. Shlyaeva (2015), Scale-dependent background-error covariance localisation. *Tellus, Series A: Dynamic Meteorology and Oceanography* **6**(1). doi:10.3402/tellusa.v67.28027.

Campbell, W. F., C. H. Bishop, and D. Hodyss (2010), Vertical covariance localization for satellite radiances in ensemble Kalman filters. *Monthly Weather Review* **138**(1), 282–290. doi:10.1175/2009MWR3017.1.

Destouches, M., T. Montmerle, Y. Michel, and B. Ménétrier (2021), Estimating optimal localization for sampled background-error covariances of hydrometeor variables. *Quarterly Journal of the Royal Meteorological Society* **147**(734), 74–93. doi:10.1002/qj.3906.

Ensemble-Data-Assimilation (n.d.), Why is Ensemble Data Assimilation important. URL: `https://www.dwd.de/EN/research/weatherforecasting/num_modelling/04_ensemble_methods/ensemble_data_assimilation/ensemble_data_assimilation.html`.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* **99**(C5). doi:10.1029/94jc00572.

Gaspari, G. and S. E. Cohn (1999), Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society* **125**(554). doi:10.1256/smsqj.55416.

Ho, T. K. (1995), Random decision forests. In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. 1, pp. 278–282. doi:10.1109/ICDAR.1995.598994.

Houtekamer, P. L. and H. L. Mitchell (1998), Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review* **126**(3), 796–811. doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Houtekamer, P. L. and H. L. Mitchell (2001), A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review* **129**(1). doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen (2005), Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Monthly Weather Review* **133**(3), 604–620. doi:10.1175/MWR-2864.1.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh (2007), Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena* **230**(1-2), 112–126. arXiv:0511236, doi:10.1016/j.physd.2006.11.008.

Kalman, R. E. and R. S. Bucy (1961), New results in linear filtering and prediction theory. *Journal of Fluids Engineering, Transactions of the ASME* **83**(1), 95–108. doi:10.1115/1.3658902.

Le Dimet, F. X. and O. Talagrand (1986), Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus, Series A* **38 A**(2), 97–110. doi:10.3402/tellusa.v38i2.11706.

Lei, L. and J. L. Anderson (2014), Comparisons of empirical localization techniques for serial ensemble kalman filters in a simple atmospheric general circulation model. *Monthly Weather Review* **142**(2), 739–754. doi:10.1175/MWR-D-13-00152.1.

Lei, L. and J. S. Whitaker (2015), Model space localization is not always better than observation space localization for assimilation of satellite radiances. *Monthly Weather Review* **143**(10). doi:10.1175/MWR-D-14-00413.1.

Lorenc, A. C. (1986), Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **112**(474), 1177–1194. doi:10.1002/qj.49711247414.

Lorenc, A. C. (2003), The potential of the ensemble Kalman filter for NWP - A comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society* **129**(595 PART B), 3183–3203. doi:10.1256/qj.02.132.

Miyoshi, T. and Y. Sato (2007), Assimilating satellite radiances with a local ensemble transform Kalman filter (LETKF) applied to the JMA global model (GSM). *Scientific Online Letters on the Atmosphere* **3**, 37–40. doi:10.2151/sola.2007-010.

Moosavi, A., A. Attia, and A. Sandu (2019), Tuning Covariance Localization Using Machine Learning. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11539 LNCS, pp. 199–212. doi:10.1007/978-3-030-22747-0_16.

Necker, T. (2019), *The impact of observations in convective-scale numerical weather prediction.* Ph.D. thesis, Ludwig-Maximilians-Universität München. URL: `http://nbn-resolving.de/urn:nbn:de:bvb:19-246537`.

Necker, T., S. Geiss, M. Weissmann, J. Ruiz, T. Miyoshi, and G. Y. Lien (2020), A convective-scale 1,000-member ensemble simulation and potential applications. *Quarterly Journal of the Royal Meteorological Society* **146**(728). doi:10.1002/qj.3744.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Petrie, R. E. and S. L. Dance (2010), Ensemble-based data assimilation and the localisation problem. *Weather* **65**(3), 65–69. doi:10.1002/wea.505.

Potthast, R. (2019), Documentation of the Data Assimilation Coding Environment (DACE). Tech. rep., Deutscher Wetterdienst (DWD).

Reinert, D., F. Prill, H. Frank, M. Denhard, M. Baldauf, C. Schraff, C. Gebhardt, C. Marsigli, and G. Zängl (2021), DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System. Tech. Rep. Research and Development at DWD DWD. URL: `www.dwd.de`, doi:10.5676/DWD _ pub/nwv/icon _ 2.1.4.

Roh, S., M. Jun, I. Szunyogh, and M. G. Genton (2015), Multivariate localization methods for ensemble Kalman filtering. *Nonlinear Processes in Geophysics* **22**(6), 723–735. doi:10.5194/npg-22-723-2015.

Saha, S., S. Moorthi, H. L. Pan, X. Wu, J. Wang, S. Nadiga, P. Tripp, R. Kistler, J. Woollen, D. Behringer, H. Liu, D. Stokes, R. Grumbine, G. Gayno, J. Wang, Y. T. Hou, H. Y. Chuang, H. M. H. Juang, J. Sela, M. Iredell, R. Treadon, D. Kleist, P. Van Delst, D. Keyser, J. Derber, M. Ek, J. Meng, H. Wei, R. Yang, S. Lord, H. Van Den Dool, A. Kumar, W. Wang, C. Long, M. Chelliah, Y. Xue, B. Huang, J. K. Schemm, W. Ebisuzaki, R. Lin, P. Xie, M. Chen, S. Zhou, W. Higgins, C. Z. Zou, Q. Liu, Y. Chen, Y. Han, L. Cucurull, R. W. Reynolds, G. Rutledge, and M. Goldberg (2010), The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society* **91**(8), 1015–1057. doi:10.1175/2010BAMS3001.1.

Schur, J. (1911), Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal fur die Reine und Angewandte Mathematik* **1911**(140). doi:10.1515/crll.1911.140.1.

Whitaker, J. S. and T. M. Hamill (2002), Ensemble data assimilation without perturbed observations. *Monthly Weather Review* **130**(7), 1913–1924. doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

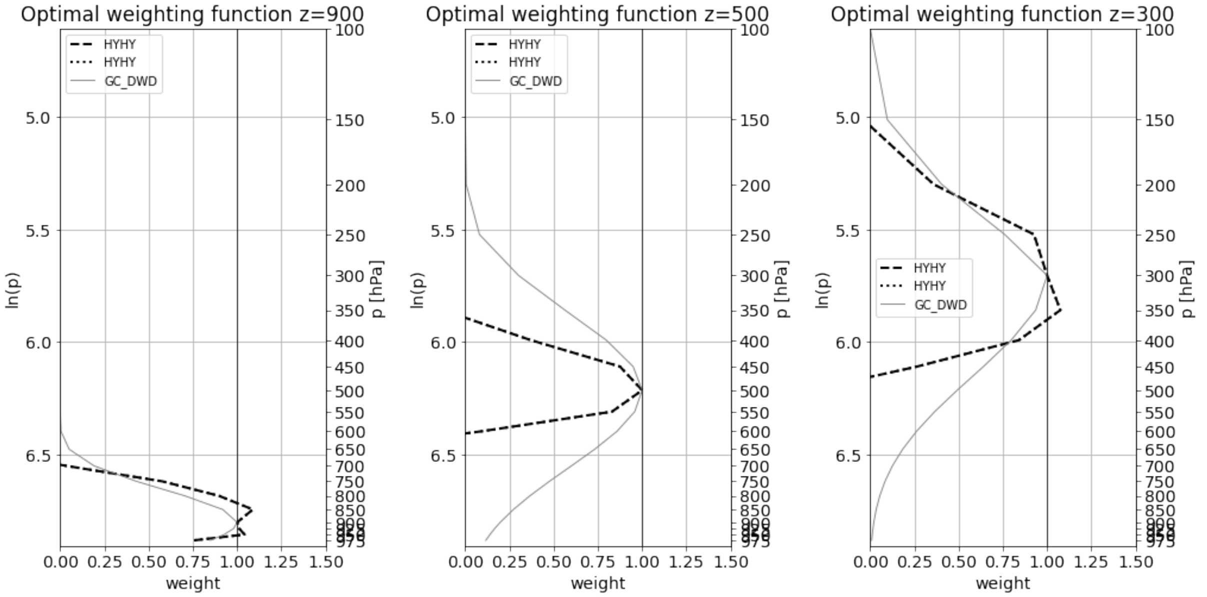
List of Tables

List of Figures

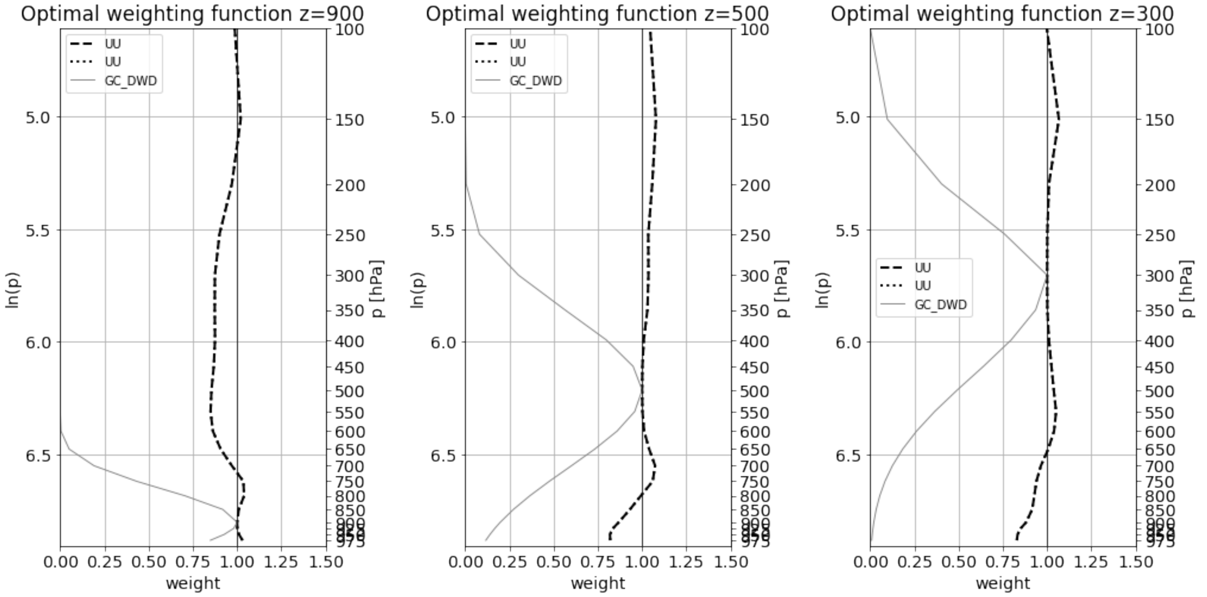
Appendix

In the following shown are the optimal weighting functions for other correlations than in the thesis discussed. All weighting functions are calculated like the functions shown in Section 4.2.

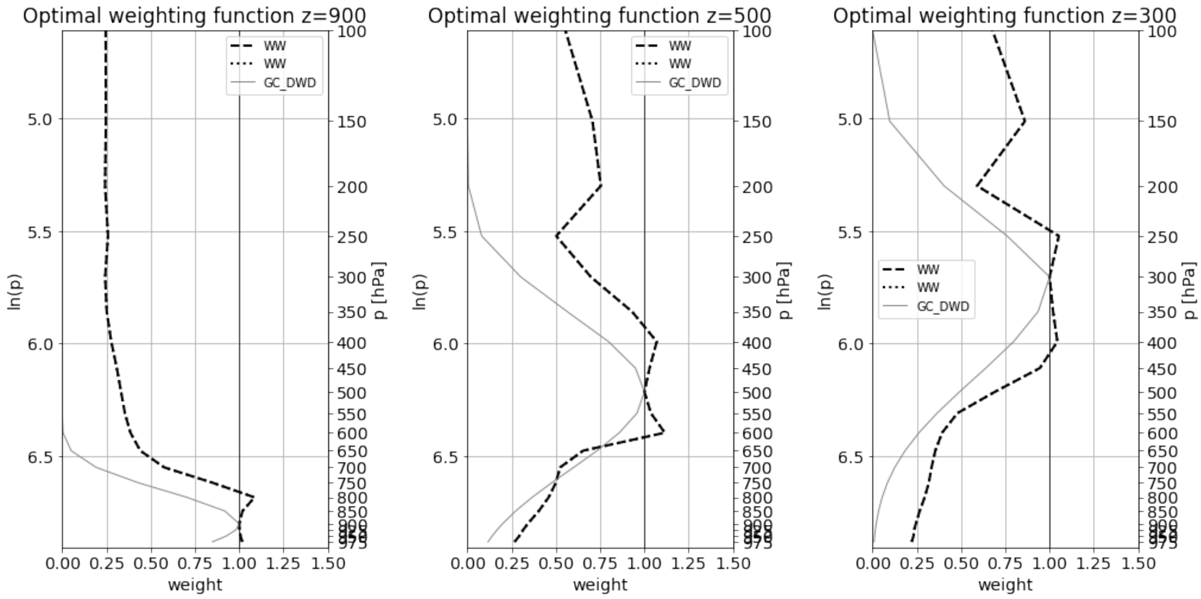
selfcorrelations



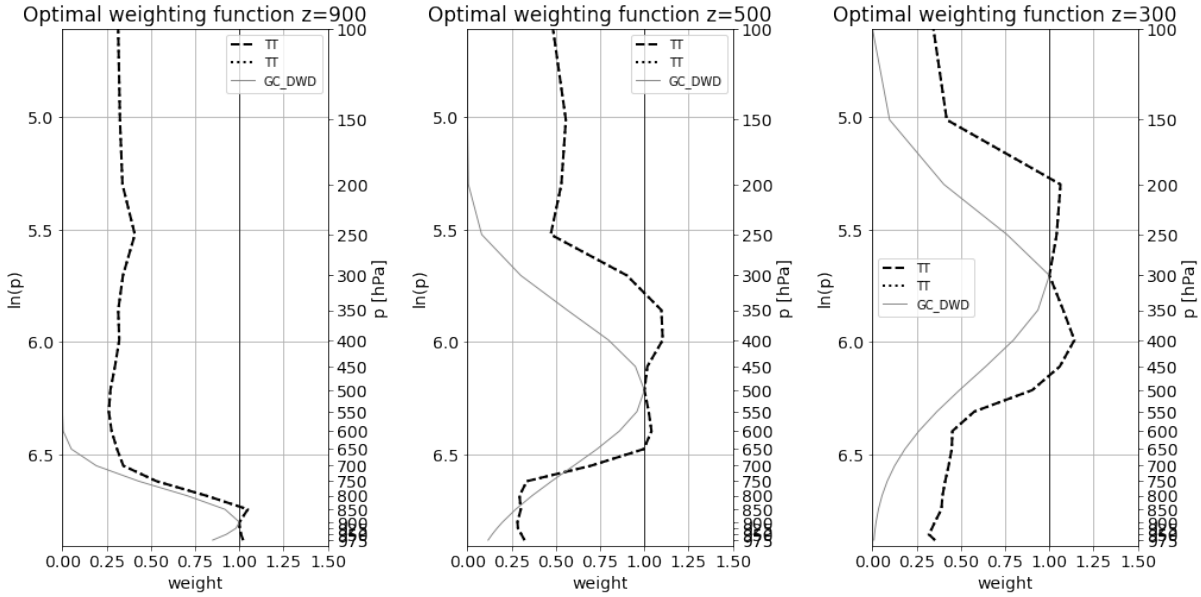
mean optimal weighting function of selfcorrelations of hydrometeor mixing ratio (HY), compared to the GC of the DWD (grey line)



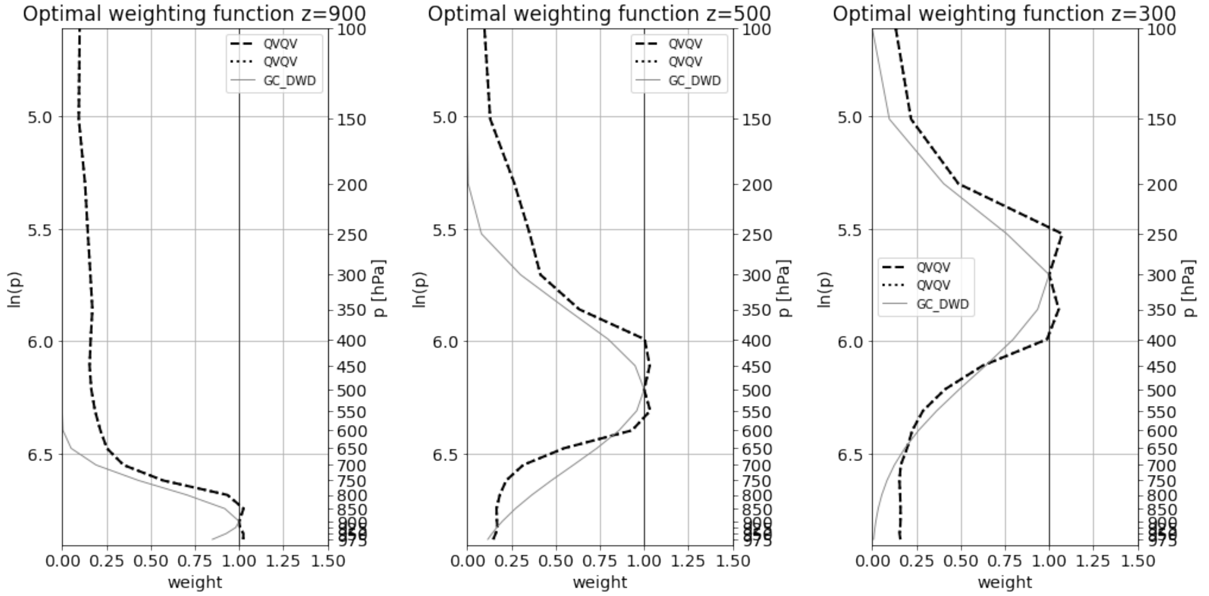
mean optimal weighting function of selfcorrelations of u-wind (U), compared to the GC of the DWD (grey line)



mean optimal weighting function of selfcorrelations of vertical wind (W), compared to the GC of the DWD (grey line)



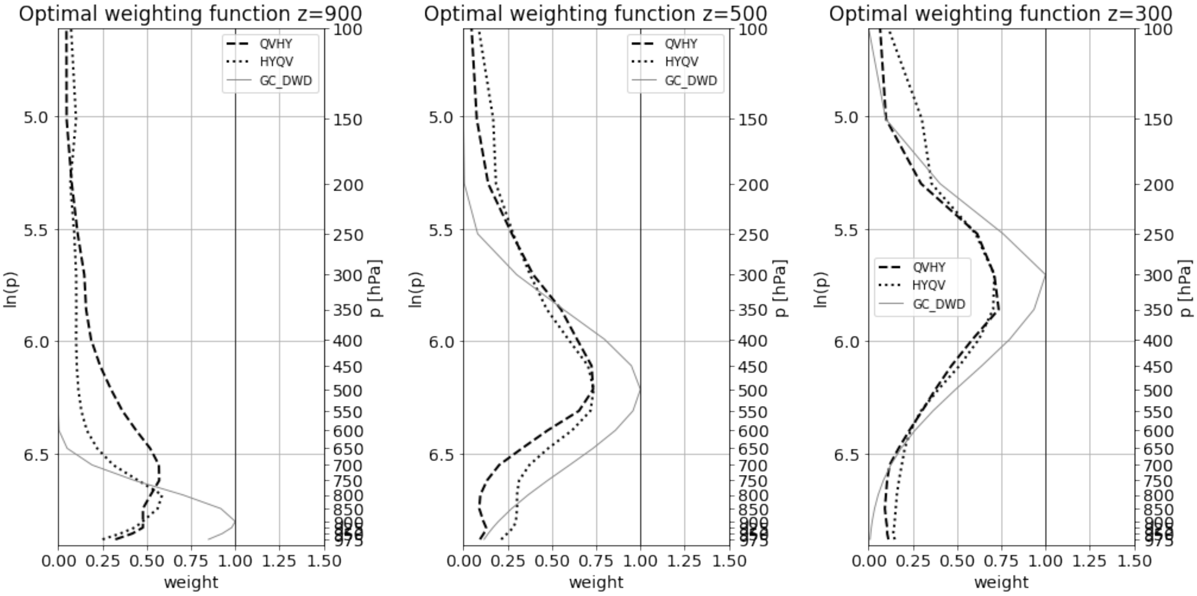
mean optimal weighting function of selfcorrelations of temperature (T), compared to the GC of the DWD (grey line)



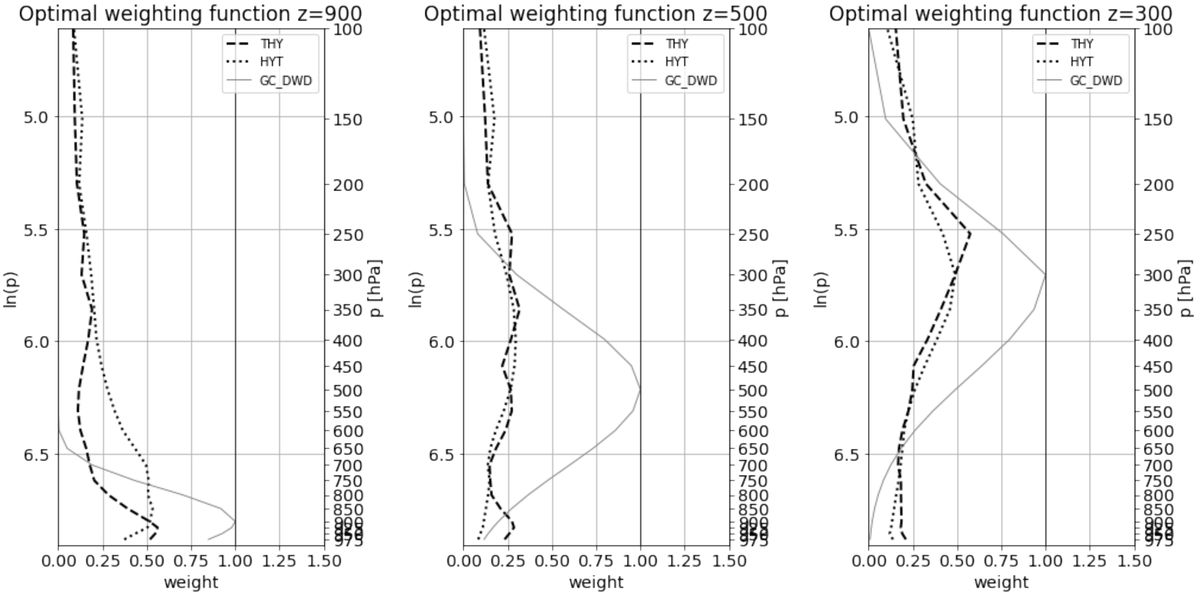
mean optimal weighting function of selfcorrelations of specific humidity (QV), compared to the GC of the DWD (grey line)

crosscorrelations

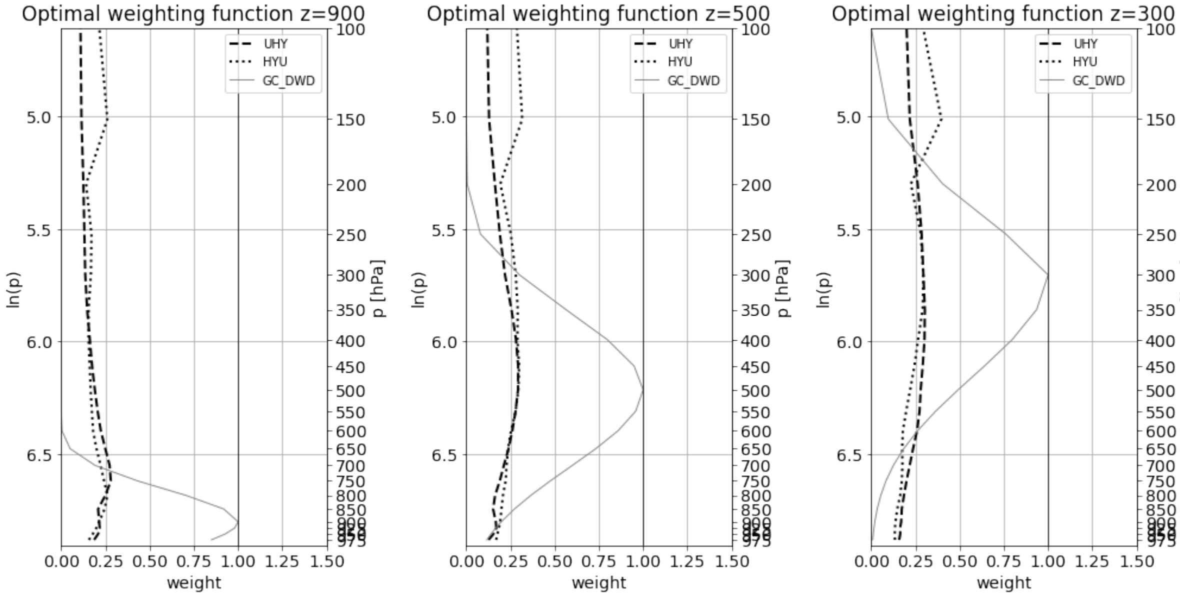
In the shortcuts of the correlations, the parameter named first is the parameter correlated from every level with the one on the correlation level.



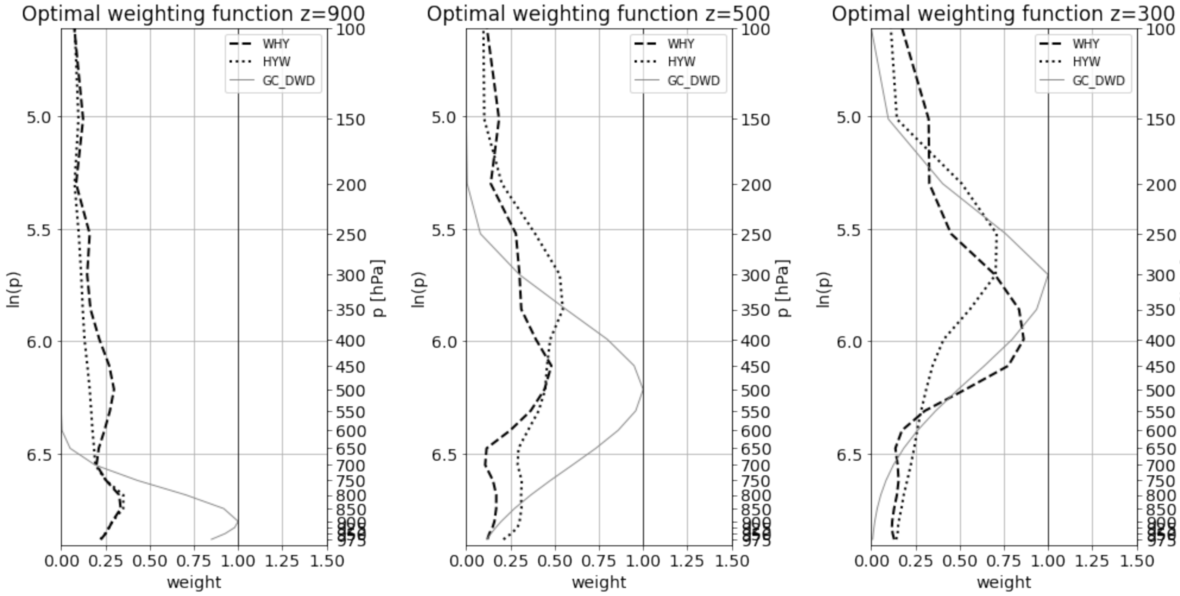
mean optimal weighting function of crosscorrelations between hydro meteor mixing ratio and specific humidity: QVHY (dashed), QVHY (dotted); compared to the GC of the DWD (grey line)



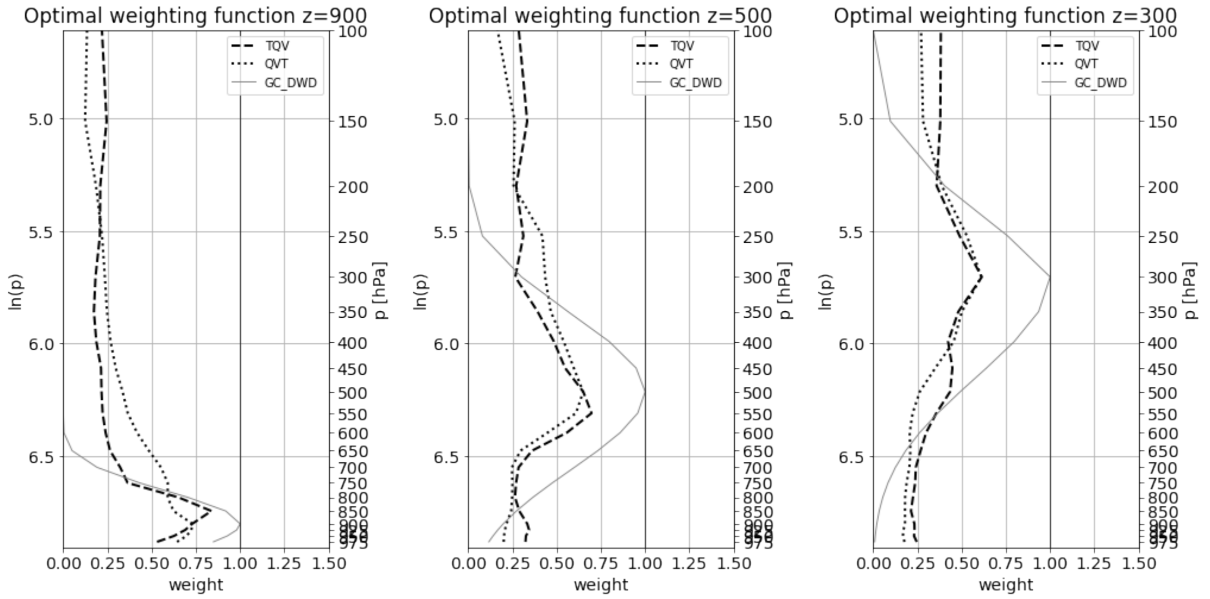
mean optimal weighting function of crosscorrelations between hydro meteor mixing ratio and temperature: THY (dashed), HYT (dotted); compared to the GC of the DWD (grey line)



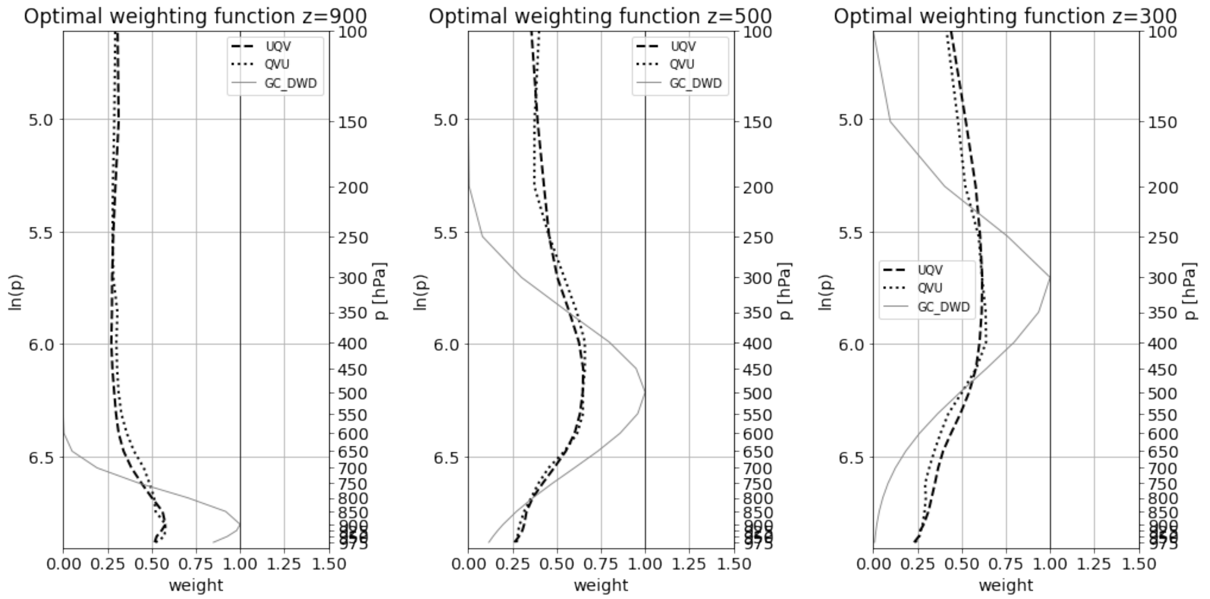
mean optimal weighting function of crosscorrelations between hydro meteor mixing ratio and u-wind: UHY (dashed), HYU (dotted); compared to the GC of the DWD (grey line)



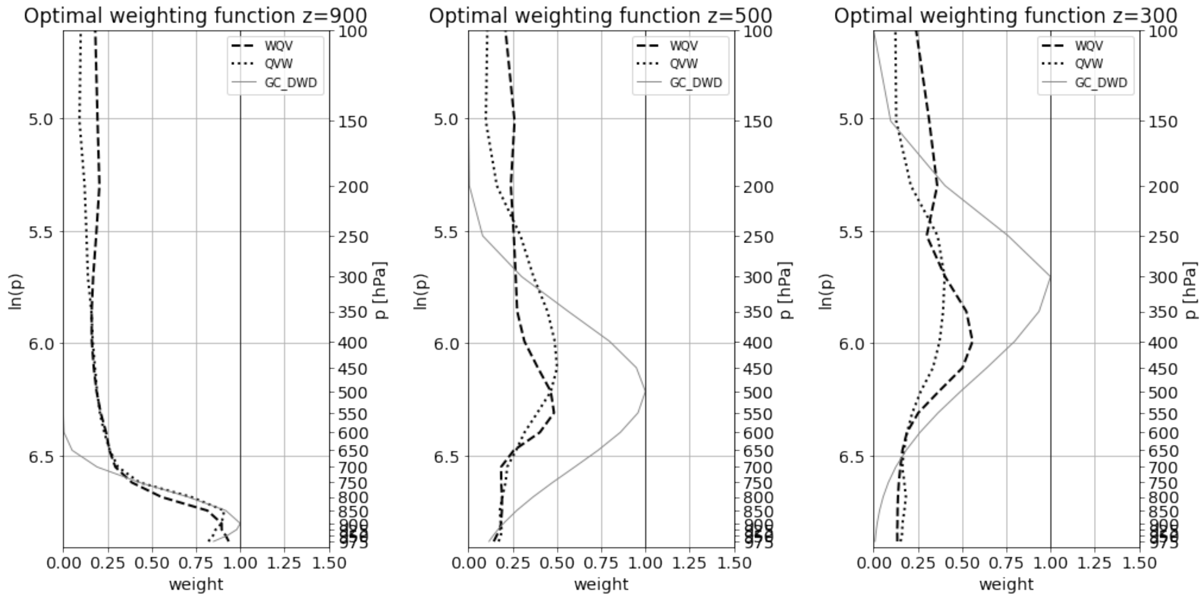
mean optimal weighting function of crosscorrelations between hydro meteor mixing ratio and vertical wind: WHY (dashed), HYW (dotted); compared to the GC of the DWD (grey line)



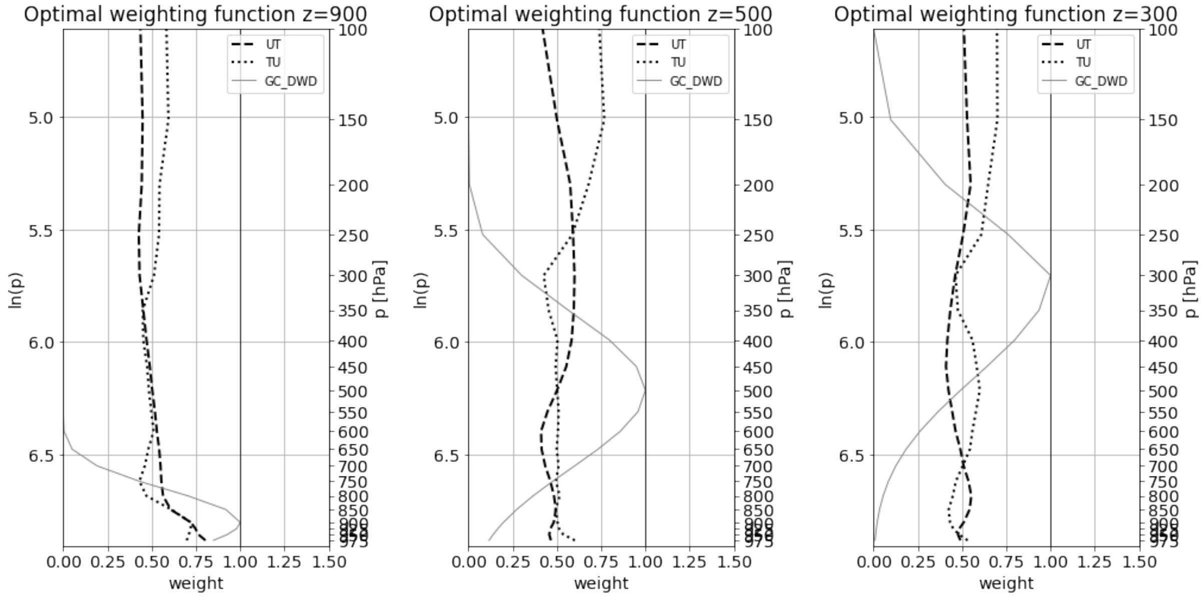
mean optimal weighting function of crosscorrelations between specific humidity and temperature: TQV (dashed), QVT (dotted); compared to the GC of the DWD (grey line)



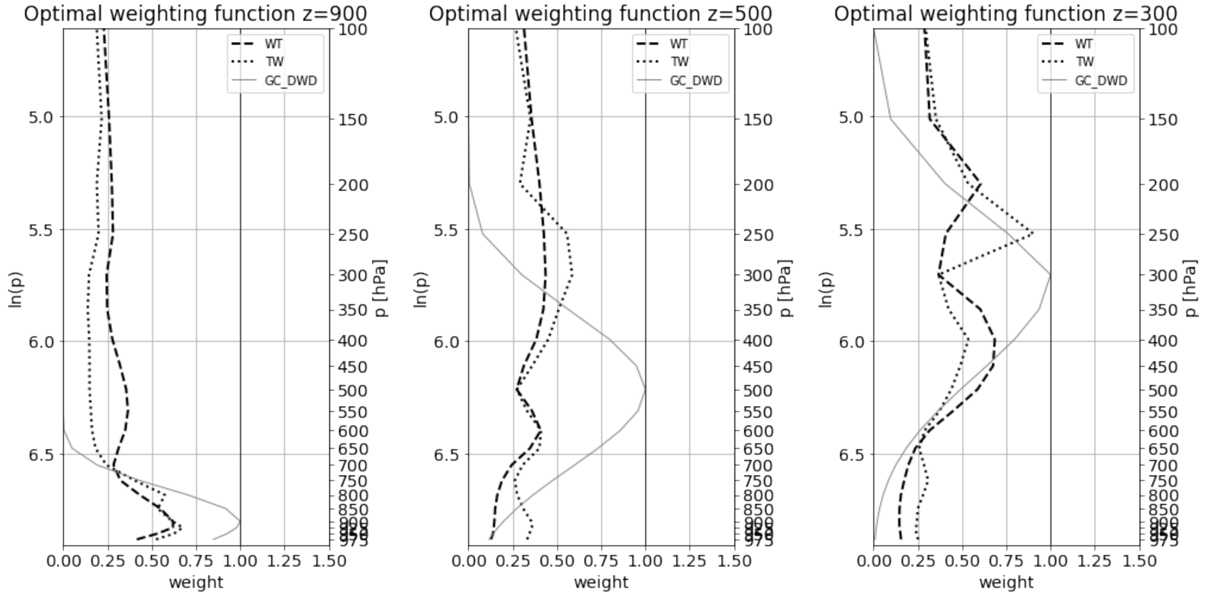
mean optimal weighting function of crosscorrelations between specific humidity and u-wind: UQV (dashed), QVU (dotted); compared to the GC of the DWD (grey line)



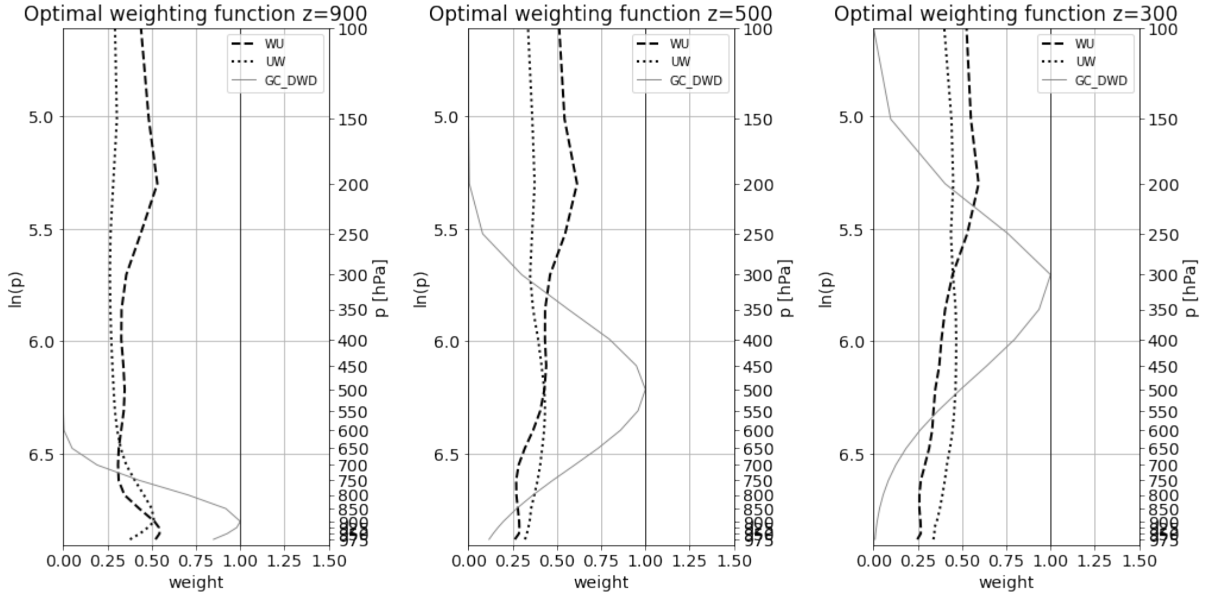
mean optimal weighting function of crosscorrelations between specific humidity and vertical wind: WQV (dashed), QVW (dotted); compared to the GC of the DWD (grey line)



mean optimal weighting function of crosscorrelations between temperature and u-wind: UT (dashed), TU (dotted); compared to the GC of the DWD (grey line)



mean optimal weighting function of crosscorrelations between temperature and vertical wind: WT (dashed), TW (dotted); compared to the GC of the DWD (grey line)



mean optimal weighting function of crosscorrelations between u-wind and vertical wind: WU (dashed), UW (dotted); compared to the GC of the DWD (grey line)