Technical Note: An illustrative introduction to the domain dependence of spatial Principal Component patterns

Christian Lehr[1,2] and Tobias L. Hohenbrink[3]

[1] Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
[2] University of Potsdam, Institute for Environmental Sciences and Geography, Potsdam, Germany
[3] German Weather Service (DWD), Agrometeorological Research Centre, Braunschweig, Germany

Correspondence to: C. Lehr (christian.lehr.1@uni-potsdam.de)

Abstract

Principal Component Analysis (PCA) of synchronous time series of one variable, e.g. water level or discharge, measured at multiple locations, has been applied in a wide spectrum of hydrological analyses. Principal Components (PCs) were used for regionalisation and to identify dominant modes, signals, processes or other hydrological properties of the analysed system. The possibility that the PCs of such analysis can exhibit domain dependence (DD) found only little recognition in the hydrological PCA literature so far. DD describes the situation in which the spatial PC patterns are mainly determined by the spatial extent of the analysed data set (domain size) and the spatial arrangement of the data sets´ locations (domain shape). Thus, instead of the hydrological functioning of the analysed system, the spatial PC patterns rather reflect the functioning of the PCA within the context of the data set´s spatial domain. The effect is caused by homogeneous spatial autocorrelation in the analysed series. DD patterns are distinct, with strong gradients and contrasts. We show that it can come together with substantial accumulation of variance in the leading PCs. In addition, DD can cause effectively degenerate multiplets, i.e. PCs which are not well separable. All these features are highly suggestive and easily lead to wrong hydrological interpretations. Consequently, DD should be considered for any application in which the PCs are used to draw conclusions about spatially distinct properties of the analysed system. For most practical applications checking the first few leading PC patterns should be sufficient. Visual comparison of the spatial PC patterns from subdomains with markedly different shapes and/or sizes can serve as quick qualitative check. Reference patterns can be used to test whether spatial PC patterns differ significantly from pure DD patterns. We present two methods, one stochastic, one analytic, to calculate DD reference patterns for defined spatial correlation properties and arbitrary spatial domains. With a series of synthetic examples, we explore the DD effect with respect to a) domain shape, b) domain size and spatial correlation length and c) effectively degenerate multiplets. Particular focus is given to the effect of DD on the explained variance of the PCs and the contrasts of their spatial patterns. An application example with a precipitation raster data set is presented and different options to detect and diminish DD are discussed. Accompanying this technical note, R-scripts to (i) demonstrate and explore the DD effect, and (ii) perform the presented DD reference methods are provided.

1 Introduction

In hydrology, Principal Component Analysis (PCA), also known as Empirical Orthogonal Function (EOF) analysis or Karhunen–Loève Transform, is a popular tool to analyse spatio-temporal data sets. The analysed data can be structured in various ways (Richman, 1986; Demšar et al., 2013). Here, the focus is on PCA of data sets comprising synchronous time series of one observed variable, e.g. water level, with the time series (a) being distributed in space at multiple locations and (b) being used as variables for the PCA. This is known as S-mode PCA (Richman, 1986) or atmospheric science PCA (Demšar et al., 2013). In this setting, the covariance among the time series from the different locations is analysed (Richman, 1986; Isaak et al., 2018). For each PC there is a temporal and a spatial pattern. The PCs are series of the same length as the analysed time series and can be plotted against the common time index (temporal PC patterns). The eigenvector of each PC is associated with the complete set of locations and can be plotted against the locations´ coordinates (spatial PC patterns). All spatial patterns are orthogonal, all temporal patterns are mutually uncorrelated. With this, the leading PCs provide a compact description of the spatio-temporal variability of the data set. S-mode PCA can be applied to data from very different hydrological systems such as catchments or soil columns.

A non-exhaustive list of hydrological applications comprises S-mode PCA to describe the spatio-temporal variability of streamflow (Smirnov, 1973; Bartlein, 1982; Lins, 1985ab, 1997; Kalayci and Kahya, 2006), groundwater level (Winter et al., 2000; Longuevergne et al., 2007; Lehr and Lischeid, 2020), lake water level (Lischeid et al., 2010), soil moisture (Korres et al., 2010; Nied et al., 2013; Hohenbrink et al., 2016; Bieri et al., 2021), precipitation (Kumar and Duffy, 2009; Thomas et al., 2012; Bieri et al., 2021), drought (Karl and Koscielny, 1982; Santos et al., 2010; Ionita et al., 2015), atmospheric rivers (Li et al., 2023), or river water temperature (Isaak, et al., 2018). In stark contrast to its widespread use, the possibility that the PCs of such analysis can exhibit domain dependence (DD) is rather unknown in hydrological PCA literature.

DD describes the situation in which the spatial PC patterns from S-mode PCA are mainly determined by the size and shape of the analysed spatial domain, meaning the spatial extent of the data set and the spatial arrangement of its locations (Buell, 1975, 1979; Richman, 1986). If the spatial autocorrelation of the data set´s variable is homogeneous across the domain, its size and shape induce distinct sequences of spatial PC patterns due to the variance maximization of the PCs and the orthogonality constraint of the PCs` eigenvectors (Jolliffe, 2002; Wilks, 2006). Buell (1975) identified classical sequences for data sets with basic geometric domain shapes and isotropic spatial autocorrelation (e.g. Figure 1). The spatial pattern of PC 1 is a weighted spatial average emphasizing the centroid of the network ("mean behaviour"). The PC 2 pattern is a gradient depicting the variability along the axis of the longest extent of the domain. The PC 3 pattern covers the next largest spread of spatial variability orthogonal to the spatial patterns of PC 1 and PC 2, etc. Given the functioning of the PCA, the sequence simply reflects (a) that the covariance between the locations has its maximum in the centroid of the network because it is the point which is on average closest to all other locations, and (b) that the only structure in the variability of the data set is the homogeneous decay of covariance with distance (Dommenget, 2007). On a sphere the resulting spatial PC patterns of such a data set would be the spherical harmonics (North and Cahalan, 1981).

Ignorance about DD can easily lead to wrong interpretations of PCA results. DD patterns are distinct, with strong gradients and contrasts, and therefore highly suggestive to indicate physically meaningful drivers or properties of the analysed system. In the climatological literature DD was intensely discussed (Buell, 1975, 1979; Horel, 1981; Richman, 1986, 1987, 1993; Jolliffe, 1987; Legates, 1991, 1993). Apparently, the topic did not reach the hydrological community, even though the effect of size and shape of the network geometry on the results was observed in early hydrological S-mode PCA applications (Smirnov, 1973; Bartlein, 1982; Lins, 1985b). For that reason, we want to raise attention to the DD effect among PCA users in the hydrological community again to reduce the risk of drawing wrong hydrological conclusions from spatio-temporal PCA.

DD is one aspect in the general discussion on the physical interpretation of S-mode PCs. There are strongly diverging opinions, ranging from "never physically interpret any PCs" to "distinct processes can be meaningfully assigned to single

PCs". For physical processes or modes of geosystems, the S-mode PC properties orthogonality of spatial patterns, linear uncorrelatedness of temporal patterns and successive maximization of variance are heavy constraints (Buell, 1979; Jolliffe, 2002; von Storch and Zwiers, 2003; Hannachi et al., 2007; Monahan et al., 2009). By extracting maximal variance, different sources of variability can get pulled onto the first eigenvector, thereby mixing the sources (e.g. Figure 14A in Karl and Koscielny, 1982). The successive order of the PCs implies that they should not be interpreted isolated, but only with reference to the preceding PCs. The spatio-temporal patterns of the first PC set the reference for all subsequent PC patterns. Forced by the orthogonality constraint, prominent features of the first spatial PC pattern cascade down to the spatial patterns of the other PCs (Cahalan et al., 1996). The analysis is limited to linear relationships and assumes stationarity of mean and variance of the analysed variable. If single features are assigned to single PCs, this raises the question whether the hydrological features in the analysed system are expected to exhibit orthogonal spatial patterns, to be linearly uncorrelated in time and to successively maximize variance. If not, PCA is simply the wrong model (Jolliffe 1987; 2002).

Rotation of PCs can relax the aforementioned PCA constraints (Richman, 1986; Hannachi et al., 2007; Monahan et al., 2009). It is regularly used in atmospheric mode detection. Several studies found that rotated PCA performed better than unrotated PCA for this purpose, and that their spatial patterns were less prone to DD (Richman, 1986; Compagnucci and Richman, 2006; Huth and Beranova, 2021). Despite these findings, unrotated PCA is still often used (Huth and Beranova, 2021). Regardless of whether rotated or unrotated PCA is used, the physical interpretation depends on the spatial PC patterns and requires that they are not domain dependent. The knowledge which locations carry the most variance can already be helpful to improve the physical understanding of the analysed system (Monahan et al., 2009). In hydrology, unrotated PCA is to our knowledge much more common than rotated PCA. Therefore, we mainly focus on unrotated PCA here.

DD is import for any application in which a PCA of observed data is used to draw conclusions about spatially distinct properties of the analysed system. This concerns descriptive applications in which the spatial PC patterns are used to identify dominant hydrological modes (Smirnov, 1973; Bartlein, 1982; Lins, 1985ab, 1997; Kalayci and Kahya, 2006; Thomas et al., 2012; Ionita et al., 2015) or regions with similar hydrological behaviour (regionalisation) (Karl and Koscielny, 1982; Santos et al., 2010; Nied et al., 2013), as well as the interpretation that they represent the spatial variability of concrete hydrological signals (Longuevergne et al., 2007; Lewandowski et al., 2009), hydrological processes (Hohenbrink et al., 2016; Isaak et al., 2018; Scholz et al., 2024) or physical properties (Korres et al., 2010; Lischeid et al., 2010). For all those applications it is essential that there is a physical counterpart for the spatial PC patterns in the analysed system. Thus, DD touches the very basic question whether the applied combination of data set and data analysis method allows inference on the analysed system.

DD is critical in particular for any interpretation of the PCs based on correlation analysis with other variables (Korres et al., 2010; Lischeid et al., 2010; Hohenbrink et al., 2016; Isaak et al., 2018; Scholz et al., 2024). In case of “strong DD” the correlation between their spatial patterns depends mainly on the selected spatial domain. Consider for example a soil texture gradient in west-east direction and the classical Buell patterns in Figure 1. Depending on the selected domain the spatial patterns from different PCs would correlate strongly, moderately or not at all with the gradient. Consequently, those correlations would be neither useable for the interpretation of the PCs nor for the identification of predictors for their spatio-temporal patterns. Thus, spatial PC patterns should be checked for DD prior to any interpretation implying causal relationships.

When checking for DD, it has to be considered that DD patterns are original for every combination of spatial domain and spatial correlation properties of the analysed data set. Thus, the “classical Buell patterns” are DD patterns for the distinct combinations of size and shape of the domain, spatial covariance function and spatial correlation length used in Buell´s (1975) numerical experiments (e.g. Figure 1). Spatial PC patterns of real-world data sets can be expected to deviate from those archetypes due to possible differences in all these aspects. In addition, there might be a blurring effect of measurement errors. For spatially regular distributed data sets with strong homogeneous autocorrelation and domain boundaries similar to

one of Buell´s basic domains the DD patterns of the leading PCs are commonly visually easy to recognize as Buell-like. This is less clear for those of the PCs with smaller eigenvalues (low ranked PCs). They are more finely detailed and less robust against deviations from Buell´s settings. Furthermore, there might be intermixing of the variance structures when the eigenvalues from successive eigenvectors are of very similar size (North et al., 1982; Quadrelli et al., 2005). These PCs which are not well separated with the PCA are called effectively degenerated multiplets (North et al., 1982). For their separation, additional post-processing is required, e.g. rotation of eigenvectors (Richman, 1986; Jolliffe, 1989). DD patterns from data sets with more complex domain shapes and spatially irregular distributed locations, which is the common case in hydrology, can differ substantially from Buell´s archetypes. All in all, visual recognition by comparison with Buell patterns is rather limited. Comparison with DD patterns calculated for the analysed spatial domain overcome these limitations (Cahalan et al., 1996; Dommenget, 2007). They can be used as reference to test whether spatial PC patterns differ significantly from what has to be expected from DD alone.

The objective of this technical note is to introduce (i) the DD effect and (ii) the application of DD reference patterns to the hydrological community. We illustrate our introduction primarily with synthetic examples. This ensures that the statistical properties of the examples, in particular their spatial correlation properties and spatial domains, are strictly defined. It further clarifies that all observed effects are solely caused by the specified statistical properties. Another advantage is that series of examples with systematic differences can be constructed to study the effects of specific properties, e.g. spatial correlation length or spatial extent, on the PCA results.

Note that we aim for an illustrative introduction for PCA practitioners. For a mathematically rigid introduction to the DD phenomenon see Buell (1975, 1979) and North and Cahalan (1981). All the here presented analyses were performed in R (R Core Team, 2019). Scripts to reproduce the results, explore the DD effect and calculate DD reference patterns for defined spatial correlation properties and arbitrary spatial domains are provided (Lehr, 2024). After presenting two DD reference methods, a series of synthetic examples is used to explore the DD effect with respect to a) domain shape, b) domain size and spatial correlation length and c) effectively degenerate multiplets. Particular focus is given to the effect of DD on the explained variance of the PCs and the contrasts of their spatial patterns, both common indicators for the interpretation of PCA results. Finally, an application example with a precipitation raster data set is presented and different options to detect and diminish DD are discussed.

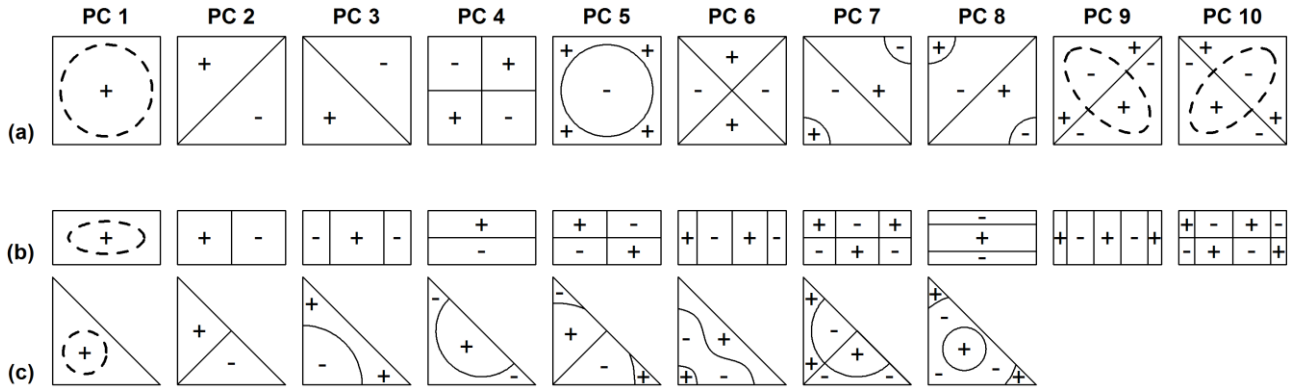

Figure 1 Exemplary reproduction of some "classical Buell patterns" for differently shaped domains of relatively similar size: (a) 6 × 6 square, (b) 5 × 10 rectangle and (c) 8 × 8 triangle (Figure 2, 5 and 4 adapted from Buell, 1975). The signs indicate positive and negative values of the spatial PC patterns. The patterns are for data exhibiting exponentially decaying spatially isotropic autocorrelation with spatial correlation length of 2 grid cells (function F1, scale parameter L = 2 in Buell (1975)). The spatial PC patterns of the rectangular shape are for the gaussian covariance function (function F2 in Buell (1975)) but Buell noted that the patterns of the exponential function are essentially the same. The dashed circle of PC 1 indicates that its pattern is of one sign in the entire domain with absolute values being highest within the circle and fading out towards the domain boundaries.

2 Data

2.1. Synthetic data

The synthetic data sets consist of synchronous spatially distributed time series exhibiting spatial but no temporal autocorrelation. Each data set is produced by concatenating realizations of a random field with identical spatial correlation properties (Figure 2). The grid cells (cells) of the random field represent the locations of a data set. The spatial autocorrelation is defined with a spatial covariance model. Each realization of the field represents one instant of time of a data set. Thus, at each location the respective time series consists of a sequence of random numbers. The number of field realizations gives the length of the simulated time series. The random fields were simulated with the "RandomFields" package (Schlather et al. 2015, 2020).

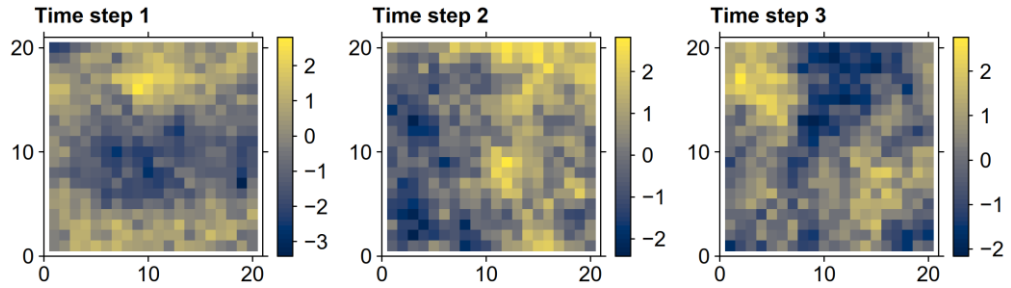

Figure 2 Three realizations of a 20 × 20 random field simulated with an isotropic exponential covariance model and spatial correlation length of 10 cells representing three instants of time of a synthetic data set.

2.2. Precipitation data

As an application example based on observed data we use time series of monthly precipitation sums from the years 1991–2020 out of a 200 km × 200 km square in northeast Germany (Figure 3). The precipitation series were selected from the 1 km × 1 km HYRAS-DE-PR precipitation grid provided by the German Weather Service (Deutscher Wetterdienst, 2025). Amongst others, the HYRAS-DE-PR precipitation product is suggested as input data for hydrological modeling (see the description file at Deutscher Wetterdienst (2025)). The monthly precipitation sums are based on daily measurements of precipitation height at the monitoring stations. The raster layers are interpolated by combining multiple linear regression considering topography with inverse distance weighting. The interpolation method preserves the measured precipitation values at the grid cells of the stations. For details, see Rauthe et al. (2013) and the description file of the data (Deutscher Wetterdienst, 2025). Except from z-scaling, no pre-processing of the precipitation series was applied.

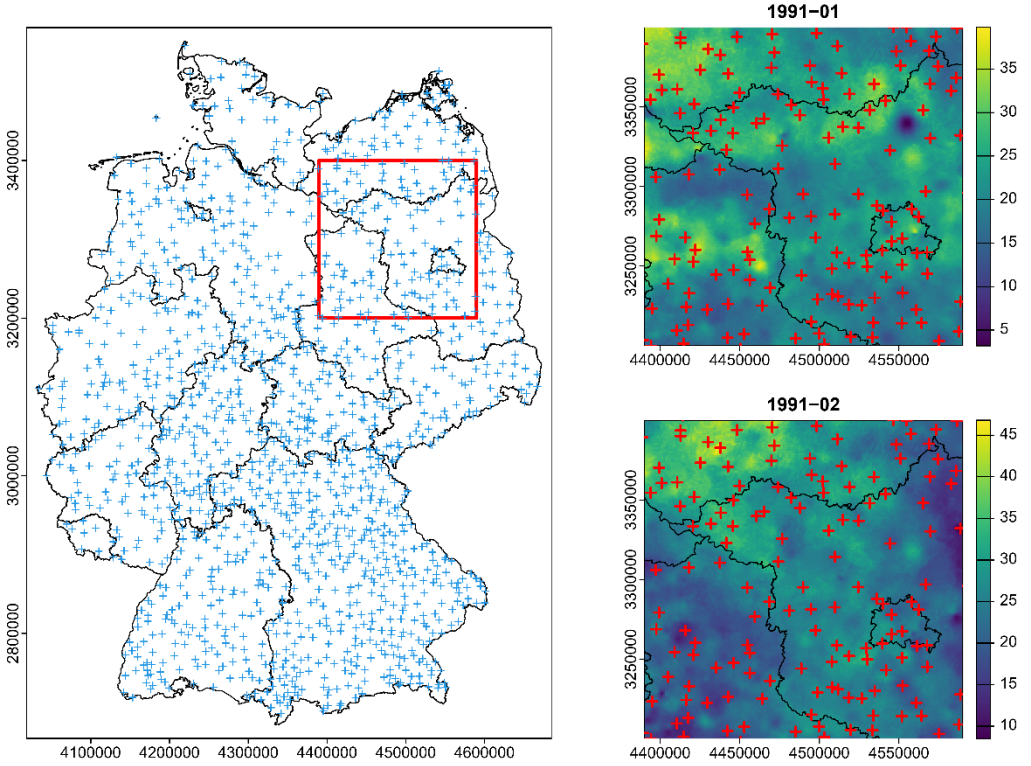

Figure 3 Maps of the precipitation data showing the permanent precipitation stations (crosses) that were used by the German Weather Service to produce raster of monthly precipitation sums in the monitoring period 1991–2020. Left panel: Federal States of Germany (black lines) and the domain selected for PCA (red square). Right panel: Sample raster from the first two months of the selected data set. The maps are in ETRS89 / LAEA Europe projection.

3 Methods

3.1. Principal Component Analysis

PCA maps an m × n data matrix X to n new linearly uncorrelated variables, the Principal Components (PCs), such that the PCs successively maximise represented fractions of the data set's variance (Wilks, 2006). The data set´s variance is defined as the sum of variances of the variables x. It equals the sum of the diagonal elements (trace) of its covariance matrix. PCA can be performed as eigenvalue decomposition of the variables' covariance matrix or as singular value decomposition of the variables' matrix with the variables being centred to their mean (Jolliffe, 2002). Unfortunately, the terminology is not used consistently throughout the literature. Here, we follow the terminology used by Jolliffe (2002) and Jolliffe and Cadima (2016) for the eigenvalue approach.

Each PC is associated with an eigenvalue λ and an eigenvector a. The values of a PC are termed scores. The variance of the scores of a PC equals its eigenvalue. The ratio of a PC eigenvalue to the sum of all PC eigenvalues gives the fraction of the data set's variance assigned to that PC. Each PC is calculated as linear combination of all n analysed variables x (non-locality).

$$pc_j = a_j X = \sum_{i=1}^{n} a_{ij}\, x_i \quad (1)$$

The coefficients a_{ij} in this linear combination are termed loadings. The loadings of a PC j are the n elements of the eigenvector a_j associated with that PC. The eigenvectors of all PCs define the orthogonal basis of the new ordination system into which the analysed data is projected (orthogonality constraint). Subject to the eigenvectors being orthogonal and the PCs being uncorrelated, the linear combinations of the PCs provide the optimal linear functions to successively maximise variance accounted for (variance maximization). The maximum variance that can be described by a linear combination of the

analysed variables is assigned to the first PC, the maximum of the remaining variance to the second PC, and so forth. Thus, the leading PCs provide a compact description of the data set´s variance. It is quite common that a few PCs suffice to summarize a major part of a data set´s variance.

For the synthetic data, PCA was performed with the function “prcomp” from the default “stats” package (R Core Team, 2019). For the precipitation data, a truncated PCA, calculating the first 20 PCs only, was performed with the "prcomp_irlba" from the "irlba" package to reduce computation time. The equivalence of the results of both PCA algorithms with respect to the leading PCs was confirmed by comparison of the results from smaller data sets.

3.1.1. S-mode PCA

In S-mode PCA, the analysed variables are synchronous time series distributed in space at multiple locations (**Figure S1**; Richman, 1986). Thus, the PCs are series of the same length as the analysed time series (temporal PC patterns) and the loadings yield values for each location (spatial PC patterns), describing the weighting of the analysed time series to calculate the PC scores. All temporal PC patterns are linearly uncorrelated with each other, each temporal PC pattern is associated with a spatial pattern and all spatial PC patterns are orthogonal to each other. Note that in this study, we perform S-mode PCA only.

3.1.2. Correlation matrix based PCA, correlation loadings and contrasts of spatial PC patterns

Normalizing the variables to zero mean and standard deviation one (z-scaling) prior applying PCA ensures equal weighting of the analysed variables. This is important if the range of values between the analysed variables differs substantially. A PCA with z-scaled variables is identical to an eigenvalue decomposition of the correlation matrix of the analysed variables. In hydrology, correlation matrix based PCA is to our knowledge more common than covariance matrix based PCA.

For the eigenvectors, different scaling conventions exist (Wilks, 2006). Here, the eigenvectors that are used to calculate the PCs are of unit length (Equation 1). In correlation matrix based PCA, normalizing the loadings from the unit length eigenvector a_j of a PC j by multiplying it with the square root of its eigenvalue λ_j is equivalent to the Pearson correlation between the scores of that PC pc_j and the analysed variables X.

$$c_j = a_j\sqrt{\lambda_j} = cor(pc_j, X) \tag{2}$$

Thus, the loadings are normalized to the commonly well-known Pearson correlation range from -1 to 1 which simplifies reading and interpretation of the PCA results. Here, we use the term “correlation loadings” for these normalized loadings c_j. We do so to prevent confusion with the coefficients that are used in the linear combination to calculate the PCs, which are not normalized to a common range (Equation 1). The sum of the squared correlation loadings c_j of a PC j equals its eigenvalue λ_j. Thus, they can be used to calculate the fractions of variance associated with the PCs. In the following, the spatial PC patterns are described with correlation loadings only.

For S-mode PCA, the normalization enables direct comparison of the contrasts of spatial patterns from different PCs or PCAs. Here, we define the contrast of a spatial PC pattern as the range between the minimum and maximum of the correlation loading values of that PC. Thus, the maximum contrast possible would be 2.

3.2. DD reference patterns

DD reference patterns are the DD patterns of a distinct combination of spatial domain and spatial correlation properties. They can be used as null hypothesis to test whether spatial PC patterns differ significantly from what has to be expected from DD alone.

3.2.1. Stochastic method

In the stochastic method, PCA is applied on synthetic data sets (Section 2.1) to derive DD reference patterns. As the data sets consist of spatially correlated white noise time series, their temporal PC patterns are white noise as well. The spatial PC patterns of the data sets are solely determined by the spatial domain and the spatial correlation properties defined in the simulation. The spatial PC patterns of data sets simulated with identically parameterized random fields differ due to the randomness in the simulations. Therefore, a three-step procedure is applied to get stable patterns (Figure 4).

Step 1: An ensemble of data sets with identical spatial domain and spatial correlation properties is simulated. Each of the data sets is analysed separately with a PCA, resulting in a PCA ensemble.

Step 2: The stability of the spatial PC patterns is assessed by pairwise correlating the spatial patterns of all possible combinations of PCs with identical ranks from the PCA ensemble. For each PC rank, the mean R^2 of the correlations is used to describe the overall similarity of the respective spatial PC patterns.

Step 3: For each PC rank (a) the mean spatial patterns from all PCAs of the ensemble and (b) their standard deviation patterns are calculated. They are calculated as the mean and standard deviation of the correlation loadings of PCs with identical rank from the PCA ensemble.

The mean spatial PC patterns are the DD reference patterns for data sets with the spatial domain and the spatial correlation properties defined in step 1. The standard deviation patterns serve as their spatially discrete uncertainty estimation. The variance represented with the DD reference patterns ("explained variance") is estimated with the mean and standard deviation of the explained variances of PCs with identical rank from the ensemble.

PCs with identical rank from different data sets of an ensemble might exhibit basically the same spatial pattern but with opposite signs due to the randomness of the field simulations, i.e. the pattern of one data set might be basically a negative version of another one. For the calculation of mean and standard deviation of the spatial PC patterns of an ensemble (step 3), the spatial patterns of PCs with identical rank are therefore harmonized such that they all are correlating positively. Thus, the correlation loadings of PCs that are correlating negatively with those of identically ranked PCs from the first data set are multiplied by -1 and therefore reversed.

Note that the suggested method requires the use of correlation loadings to describe the spatial PC patterns. Thus, it is restricted to correlation matrix-based S-mode PCA, meaning the analysed series have to be z-scaled (Sections 3.1.1 and 3.1.2). Furthermore, the mean spatial PC patterns are derived from a data set ensemble, not from a distinct single data set. Thus, they cannot be used to calculate PC scores.

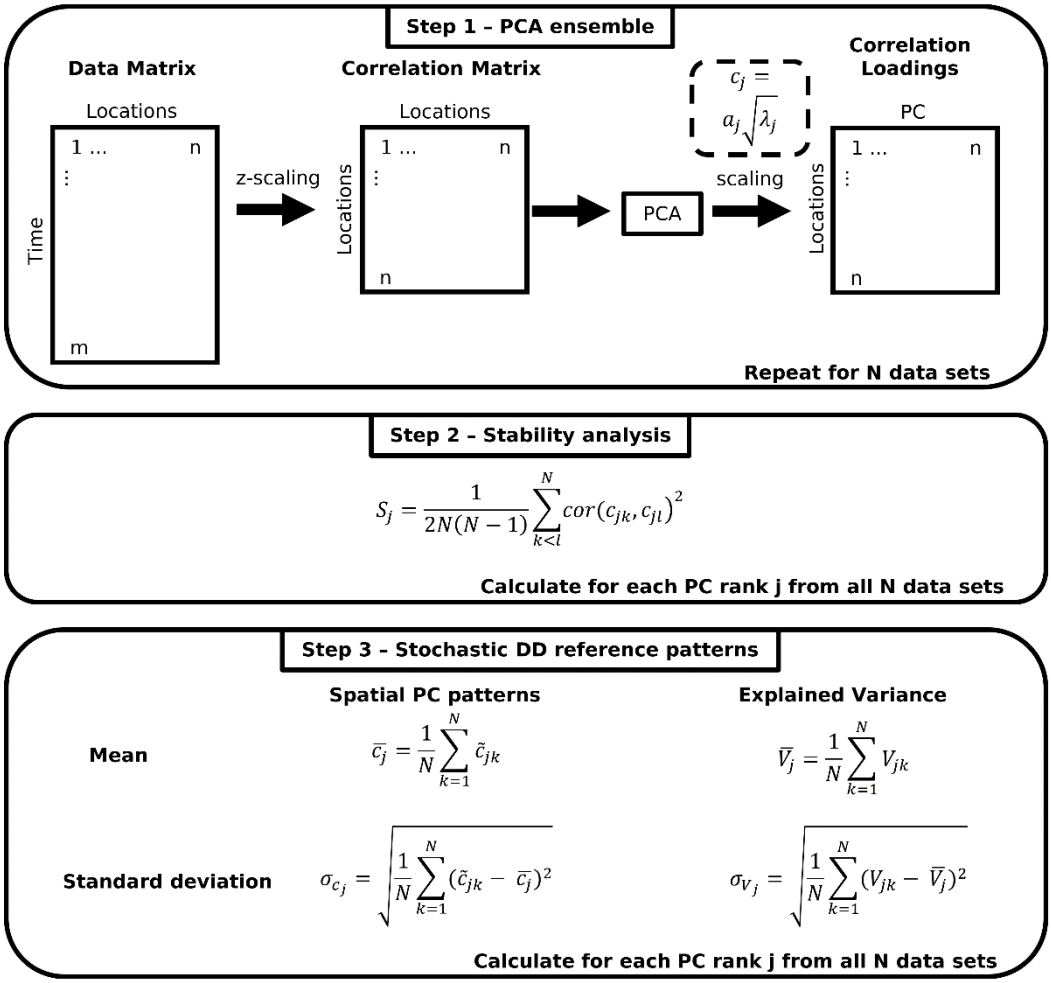

Figure 4 Stochastic DD reference method. n: number of locations, m: number of time steps, N: number of data sets, respectively PCAs, index j: PC rank, c: correlation loadings, a: loadings, λ: eigenvalue, S: stability, indices k, l: running indices for PCAs from the ensemble, $\tilde{c}$: harmonized correlation loadings, V: explained variance.

280

3.2.2. Analytic method

Another possibility to produce DD reference patterns is to perform a PCA with the “analytic”, or "exact", covariance matrix (North et al., 1982; Cahalan et al., 1996; Dommenget, 2007) of a spatially homogeneous covariance function (Figure 5). The analytic covariance matrix consists of the covariances between all of the data set´s locations calculated directly with their
285 interpoint distances from the function. For consistency with the stochastic method (Section 3.2.1), the eigenvectors (spatial patterns) were scaled to correlation loadings (Section 3.1.2). A brief review of different variants using the analytic covariance matrix to produce PCA reference patterns is given in Appendix A.

For the synthetic examples, the analytic method was performed as eigendecomposition of the analytic covariance matrix with the function “eigen” from the default “base” package (R Core Team, 2019). For the precipitation data, a truncated PCA,
290 calculating the first 20 PCs only, was performed with the function "eigs_sym" from the "RSpectra" package to reduce computation time. The equivalence of the results of both algorithms with respect to the leading PCs was confirmed by comparing the results from smaller data sets.

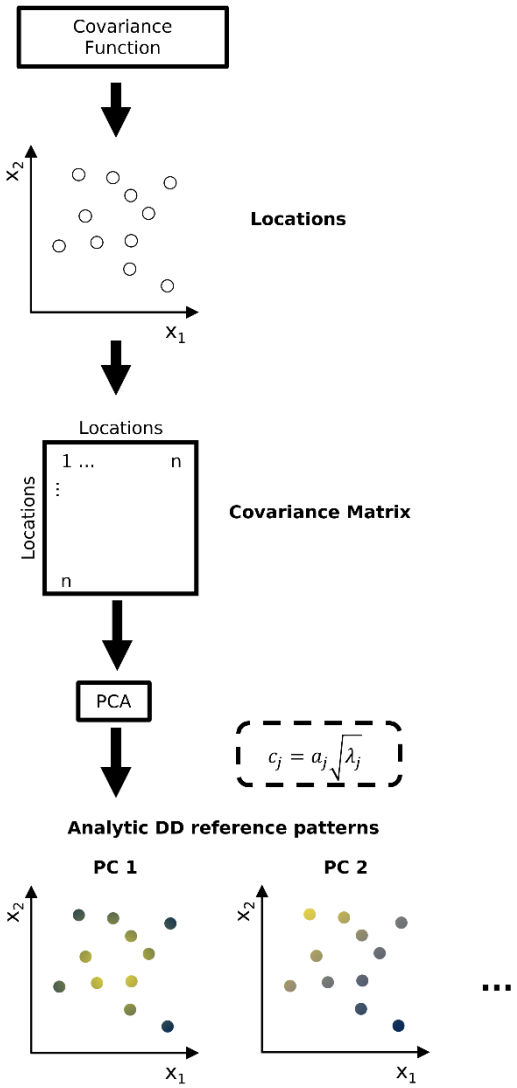

Figure 5 Analytic DD reference method.

3.3. Matching of spatial PC patterns

The matching of the spatial patterns from different PCAs was quantified with the congruence coefficient (Lorenzo-Seva and ten Berge, 2006) and Pearson correlation. The congruence coefficient φ is defined as the cosine of the angle between two vectors of component or factor loadings a_1 and a_2, both being based at the origin.

$$\varphi = \frac{\sum a_1 a_2}{\sqrt{\sum {a_1}^2 \sum {a_2}^2}} \tag{3}$$

In contrast, Pearson correlation gives the cosine of the angle between two vectors, both being based at the mean loading. Thus, the matching coefficient r in the following equation gives the Pearson correlation when b = mean(a_1), d = mean(a_2) and the congruence coefficient when b = d = 0 (see the help of R-function "factor.congruence").

$$r = \frac{\sum (a_1 - b)(a_2 - d)}{\sqrt{\sum (a_1 - b)^2 \sum (a_2 - d)^2}} \tag{4}$$

If the compared vectors have zero mean values (mean(a_1) = mean(a_2) = 0), both indices are identical. In all other cases, the results differ. The congruence coefficient is sensitive to the addition of constants, because the vector means are not removed (Lorenzo-Seva and ten Berge, 2006). Two eigenvectors with different means can be closely correlated even though their magnitude patterns differ substantially such that some variables load high on the one PC and low on the other (Richman, 1986; Lorenzo-Seva and ten Berge, 2006). In the S-mode PCA case, this means that two spatial PC patterns with different

means can be closely correlated even though some locations load high on the one PC and low on the other (thus, the maximum loadings of the two PCs could be in different locations). If that is the case, the congruence coefficient would be low, indicating the difference in magnitude patterns. Thus, in contrast to Pearson correlation it incorporates vector magnitudes in the comparison (Richman, 1986). This is desirable for the comparison of eigenvectors from PCA or factor analysis because the magnitude of the loadings is important for the interpretation of the components (Richman, 1986). Therefore, the congruence coefficient is recommended as matching coefficient over Pearson correlation for the comparison of eigenvectors from PCA or factor analysis (Richman, 1986). However, a major benefit of Pearson correlation is that it is well known and the results in terms of r or R^2 can easily be contextualized by the analyst.

Note that the stability analysis of the stochastic approach (step 2, Section 3.2.1) was performed with Pearson correlation only, because all compared PC patterns (i) were of identical rank and (ii) were based on synthetic data sets simulated with identical spatial correlation properties and identical domains. For this setting, we considered the effect of the pattern mean subtraction by Pearson correlation as negligible.

Both indices have a value range from -1 to 1, with 1 indicating a perfect match, 0 no relationship and -1 a perfect inverse match (Richman, 1986). Compared with Pearson correlation, the congruence coefficient is biased towards higher values (Richman, 1986). Several guidelines were suggested that assign specific ranges of absolute congruence coefficients (aCC) to categories of goodness-of-match, or specific thresholds as indication for the identity of components/factors (Richman, 1986; Lorenzo-Seva and ten Berge, 2006). Here, we follow Lorenzo-Seva and ten Berge (2006) who suggested that aCC values between 0.85 and 0.94 indicate fair similarity of the two components, values larger than 0.95 indicate that they can be considered equal and values below 0.85 should not be interpreted as indication for similar components.

The congruence coefficient was calculated with the function "factor.congruence" from the "psych" package. The statistical significance of the correlations was assessed with t-tests and the significance level 0.05 using the function “cor.test” from the default “stats” package (R Core Team, 2019).

3.4. North´s rule of thumb

Confidence limits to identify clearly separated eigenvalues and eigenvectors can be estimated e.g. with North´s rule of thumb (North et al., 1982; Hannachi et al., 2007) based on the data set´s effective sample size n^*, also known as number of independent observations in the sample or the number of degrees of freedom (Hannachi et al., 2007). The 95 % confidence interval of the eigenvalue λ_k is given by $\delta\lambda_k \sim \lambda_k\sqrt{2/n^*}$. In our case here, n^* equals the length of the analysed time series because the series do not exhibit temporal autocorrelation. The confidence interval for the associated eigenvector u_k can then be estimated with $\delta u_k \sim (\delta\,\lambda_k/\Delta\lambda)u_j$ where u_j is the eigenvector of λ_j, the closest eigenvalue to λ_k, and $\Delta\lambda$ the spacing $(\lambda_j - \lambda_k)$ between both eigenvalues.

3.5. Varimax rotation

Rotation aims at separating a subset of PCs more clearly such that the association between the eigenvectors and the PCs is more distinct. The goal is to reach a so called "simple structure" with the loadings being either close to zero or close to the maximum possible absolute values (Wilks, 2006). Thus, the magnitudes of the loadings are changed. The total variance of the rotated subspace is preserved, but the variance among the rotated PCs is redistributed more evenly (Jolliffe, 2002), potentially affecting which PCs are rated dominant. Different rotation methods are available (Richman, 1986). The rotation is performed by multiplication of the selected eigenvectors by a rotation matrix. If the rotation matrix is orthogonal, the

rotation is called orthogonal, otherwise oblique (Wilks, 2006). To support the interpretability of the results, the rotation matrix is chosen to optimize a simplicity criterion (Jolliffe and Cadima, 2016). Depending on the selected simplicity criterion, the rotation changes the properties of the eigenvectors and PCs. The results can depend on the number of eigenvectors that are rotated (Jolliffe, 2002; Wilks, 2006). This is different from standard PCA where the patterns and the associated variances from a set of PCs do not depend on the number of considered PCs. For example, in standard PCA the patterns and variance distributions of the first two PCs are identical, regardless of whether only the first two PCs are considered or, say, the first four PCs. Often the results are affected more by the choice of how many eigenvectors are rotated than by the choice of the simplicity criterion (Hannachi et al., 2006; Jolliffe and Cadima, 2016).

Here, we applied varimax rotation with Kaiser normalization (Kaiser, 1958). It is the most popular rotation method (Wilks, 2006). Varimax is an orthogonal rotation that maximizes the sum of the variances of the squared elements from the r selected eigenvectors b by iteratively rotating pairs of eigenvectors (Richman, 1986; Wilks, 2006). With the Kaiser normalization the eigenvectors b are normalized with the communalities h^2 of the n analysed variables (here the time series from the n different locations) prior rotation and renormalized afterwards. The communality h_i^2 of variable i is the fraction of variance from the variable that is depicted by the r rotated PCs. The normalized varimax criterion V can be calculated as

$$V = \sum_{j=1}^{r} \left\{ \left[n \sum_{i=1}^{n} (b_{ij}^2 / h_i^2)^2 - \left[\sum_{i=1}^{n} (b_{ij}^2 / h_i^2) \right]^2 \right] / n^2 \right\} \quad (5)$$

Note that the scaling of the eigenvectors that are rotated affects the varimax results (Jolliffe, 1995; Wilks, 2006). Either the orthogonality of the eigenvectors, the uncorrelatedness of the PCs or both get lost. The most popular scaling and the default in many software packages is to use eigenvectors scaled to the square root of their eigenvalues, derived from correlation matrix PCA (what we term correlation loadings here). In that case, the orthogonality of the eigenvectors and the uncorrelatedness of the PCs are lost. Other options are to use unit length eigenvectors which preserves the orthogonality of the eigenvectors, or to divide the unit length eigenvectors by the square root of their eigenvalues which preserves the uncorrelatedness of the PCs. For the introductory purpose we use the most popular variant and rotate correlation loadings only. Varimax rotation was performed with the function “varimax” from the default “stats” package (R Core Team, 2019).

4 Exploring the DD effect

4.1. Exploring Buell patterns and their stability

As a start, we estimated DD reference patterns for Buell´s (1975) three basic geometric domain shapes (Figure 1) using the stochastic method (Section 3.2.1). Ensembles of 100 data sets were simulated for each of the three shapes. The domain boundaries are shown in Figure 6. All cells within the boundaries were used. (Note that the shape of a domain means the spatial arrangement of the data set´s locations. It should not be confused with the shape of its boundary.) The sides of the square, the long side of the rectangle and the legs of the perpendicular triangle were 20 cells long, the short side of the rectangle was 10 cells long. Thus, the rectangular and the triangular domain were of half the size of the square. Each data set was simulated with a spatially isotropic exponential covariance model and a spatial correlation length of 10 cells.

For the reliability of the stochastic DD reference patterns, their stability is essential. Figure 7 summarizes the results of the stability analyses (step 2 of the stochastic method) from a series of ensembles with identical spatial domain and spatial correlation properties but different time series lengths. Thus, the plot shows for each PC rank the dependency of its spatial patterns’ stability from the time series length if all other parameters used in the simulation are identical. Based on that information it can be decided whether additional ensembles with longer time series shall be simulated to improve the estimation. Here, we considered a time series length of 10 000 sufficient for all three domains.

Note, that here and in the following we show the results for the first ten leading PCs. The decision was taken merely for the illustrative purpose. We found it to be a good balance between showing the DD pattern sequences and some degree of detail, but not too much detail that it is still visually easy to grasp. There was no other specific truncation criterion, e.g. based on eigenvalue magnitude or percent variance extracted, applied.

Figure 8 shows the mean spatial PC patterns of the ensembles. Those are the stochastic DD reference patterns. Most of them correspond to the Buell patterns shown in Figure 1. Some exhibit switches in the ranking, e.g. PC 3+4 of the rectangular domain or PC 7+8 of the square domain. The uncertainty estimation of the stochastic DD reference patterns, given by the standard deviation of the spatial PC patterns from the data set ensembles, is shown in Figure 9.

Exemplarily, the mean and standard deviation patterns of the square domain are shown in more detail (Figure S2). The scales provide information on the magnitude of both patterns. To make use of the standard deviation patterns (Figure S2b) as uncertainty estimation of the DD reference patterns, it is necessary to consider their magnitudes in relation to the contrast from the mean spatial patterns (Figure S2a). In addition, the fractions of variance assigned to the DD reference are given.

The stability of the DD patterns reflects their distinctness in the sequence of spatial PC patterns according to the PCA constraints. It depends on the specific combination of domain size and shape and spatial correlation properties of the data set. For example, for the properties here, PCs 8 to 10 of the triangle are more stable than the ones of the rectangle (Figure 7c+b). Generally, there is the tendency that the spatial patterns of low ranked PCs, which contain also more fine details, require longer times series to gain stability. It seems counter intuitive at first that PC 2 of the rectangle stabilizes faster than its PC 1 (Figure 7b). It indicates that for the properties of the simulated data the rectangular domain shape gives a clearer orientation for the spatial pattern of PC 2 than for the one of PC 1. Thus, especially for short time series the orientation of the gradient along the long side of the rectangle (PC 2) is more distinct than the position of the monopole in the centroid of the rectangle (PC 1) (Figure 8b). Similarly for the triangle, the orientation of the gradient patterns of PC 2 and 3 induced by its long side are more distinct than the position of its PC 1 monopole (Figure 7c and Figure 8c).

PCs with ambiguous orientation of spatial patterns are more likely to occur for symmetric domain shapes than for asymmetric ones (North et al., 1982). The basic geometric domain shapes used here exhibit rotational symmetry of order 4 (square), order 2 (rectangle) and order 1, i.e. no rotational symmetry, (triangle). Accordingly, within the range of the analysed time series lengths the number of PCs that exhibited unstable spatial patterns differed between the domain shapes (square: 8, rectangle: 5, triangle: 2 in Figure 7). Unstable spatial PC patterns are indicative for effectively degenerated multiplets and will be discussed in Section 4.4.

The stochastic reference script enables the production of catalogues of stability plots and DD patterns like in Figure 7 and Figure S2 for data sets with different spatial domains and spatial correlation properties (for sample catalogues see Lehr (2024)). Both plots in combination can be used to explore how the properties of a data set affect the DD patterns. Here, we neglect the effect of measurement errors. However, it can be simulated by adding noise to the realizations of the random field (Figure 2).

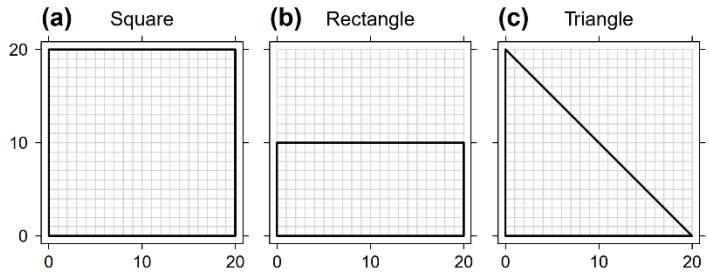

Figure 6 (a) Square, (b) rectangular and (c) triangular domain boundaries on the 20 × 20 grid. The grid cells represent locations from a data set.

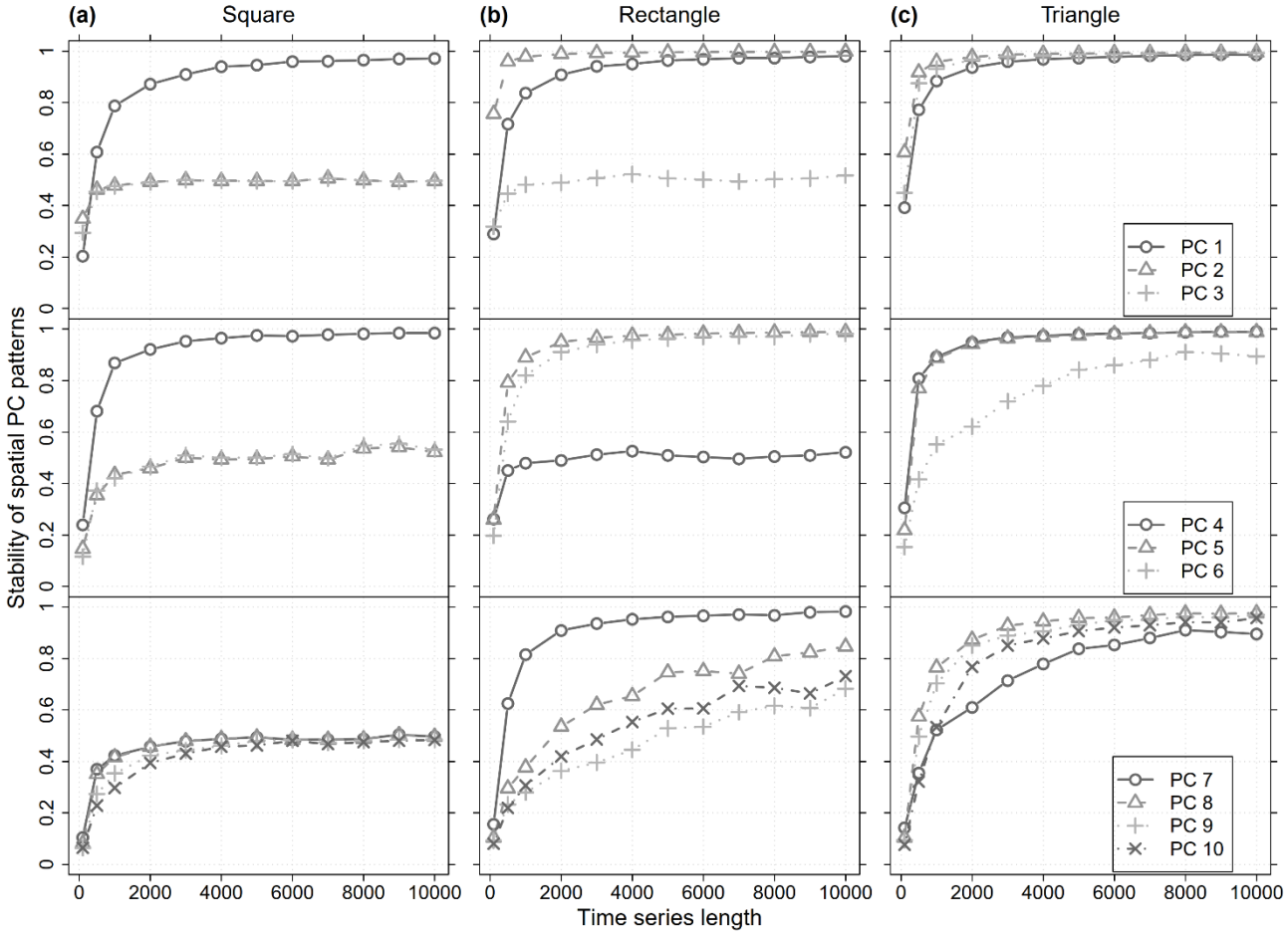

Figure 7 Stability of the spatial patterns from the leading ten PCs in relation to the time series length of the simulated data within the (a) square, (b) rectangular and (c) triangular domain boundaries of Figure 6. All cells within the boundaries were used. For each domain the results from 12 data set ensembles are shown. Each ensemble consists of 100 data sets simulated with identical time series length, an isotropic exponential covariance model and a spatial correlation length of 10 cells. Each simulated data set was analysed separately with PCA. Symbols depict the mean R^2 of the correlation between the spatial patterns of all PCs with identical rank derived from the respective ensemble. The legends in (c) apply also to (a) and (b) of the respective row.

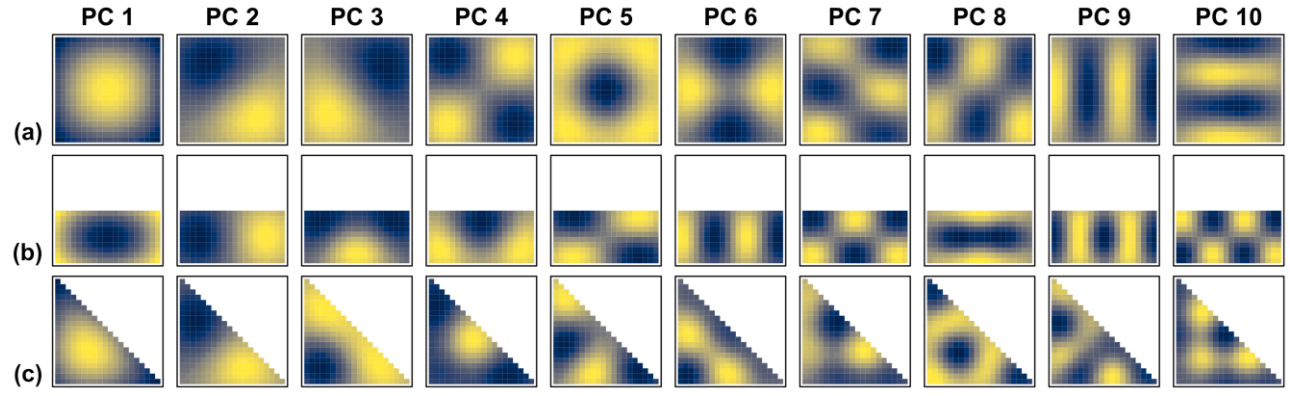

Figure 8 Overview of the leading ten mean spatial PC patterns (DD reference patterns), estimated with the stochastic method from the data set ensembles with time series length 10 000 shown in Figure 7. Instead of the +/- schemes used by Buell (1975) (Figure 1) we use colour gradients to picture the spatial patterns.

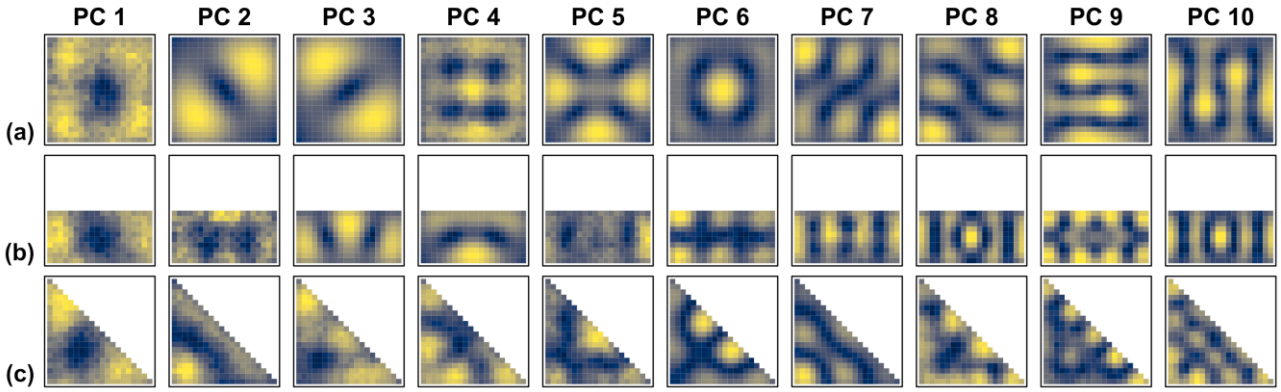

Figure 9 As in Figure 8 but for the standard deviation patterns (uncertainty estimation of the stochastic DD reference patterns). From blue to yellow the colour gradients depict increasing uncertainty.

4.2. Effects of the domain shape

For data sets with identical spatial correlation properties and similar domain size, the DD patterns are original for every domain shape. This is obvious for domains of such simple and clearly different shape like the three geometric shapes used so far. The sequence of their DD patterns is visually easy to recognize. For more complex shapes, the DD patterns are less predictable, a priori, and visual recognition is more limited.

For demonstration, we compared the DD patterns from data sets with identical spatial correlation properties in which all cells within the three geometric boundaries of Figure 6 were selected (Figure 8) with two variants in which only 40 % of the cells were randomly selected. In the first variant the subsampling was spatially homogeneous (Figure 10), in the second spatially heterogeneous (Figure 11). The domain of the second variant contained a subregion with higher sampling probability than the rest of the domain, i.e. within each domain there is one area in which the locations cluster. Clusters of locations have more weight in the calculation of the PCs analogue to the calculation of a weighted spatial mean (Karl et al., 1982). For the DD pattern of PC 1 the effect is obvious. Its monopole is placed in the centroid of the network. In comparison with the regular variant (Figure 1, Figure 8 and Figure 10) it is therefore shifted according to the density of the locations (Figure 11). The patterns of all other PCs are not predictable without calculating DD reference patterns.

Visually, the domains of the subsampling variants are still clearly of square, rectangular and triangular shape. Their leading DD patterns are recognizable as distinct spatial patterns. Most of those from the homogeneous subsampling variant (Figure 10) appear as noisy counterparts of the all cells patterns (Figure 8). In the heterogeneous case (Figure 11), the patterns of the square domain appear again relatively similar, whereas for the triangular and rectangular domain only a few PCs exhibit visually similar patterns, e.g. PC 2.

The similarity of patterns formed by congruent selections of cells from the different variants is of particular interest. It addresses the question whether the spatial PC patterns calculated from two different domains result in different relations between the values at locations with coincident coordinates. This is visually only poorly assessible. Therefore, we correlated the patterns of the subsampling variants with the patterns formed by the corresponding subsets from their all cells counterpart (that is, the all cells patterns clipped with the coordinates of the subsampling variant). For example, the patterns from the homogeneously subsampled square (Figure 10a) were correlated with the patterns from the all cells square (Figure 8a) clipped with the coordinates of the subsampled square.

For the spatial patterns of the homogeneous subsampling variant and the all cells variant, the correlation analysis confirmed the visual impression of overall similarity (Table 1). But it also showed that there are differences. The patterns of the subsampling variant can be:

1) simply noisy variants of the all cells patterns (e.g. PC 1 and 2 from all domains),

2) simply noisy variants of the all cells patterns but with different ranking (e.g. PC 3 and 4 from the rectangular domains),

3) a mix of all cells patterns (e.g. PC 4 and 5 from the square domains[1]), or

4) very different from the all cells patterns (e.g. PC 10 from all domains[2]).

Transitions between 3) and 4) are possible (e.g. PC 6 and 7 of the rectangular domain). Generally, the differences increase towards the low ranked PCs with the more detailed patterns. But, there are also substantial differences between the patterns from relatively high ranked PCs possible (e.g. PC 4 and 5 from the square domains). Thus, even for rather homogeneous subsampling, the DD patterns are not necessarily simply noisy variants of the classical Buell patterns. The comparison with the heterogeneous variant yielded substantially stronger deviations (Table 2). Thus, generally, visual recognition of Buell like patterns in S-mode PCA results is a concrete indication for DD. However, it is so in particular for the leading PC patterns from domains with rather homogeneous spatial arrangement of locations within boundaries similar to Buell´s archetypes. Even for domains of similar size and identical spatial correlation properties, deviations from strictly regular distribution of locations alone can result in DD patterns substantially deviating from what one might expect with the classical Buell patterns in mind.

Side note: The spatial PC patterns of the subsampling variants required shorter time series lengths to stabilize (Figure 12 and Figure S3) than the all cells variant (Figure 7). This indicates that the subsampling resulted in a more unbalanced arrangement of locations and therefore a more distinct orientation for the order of the orthogonal spatial PC patterns.

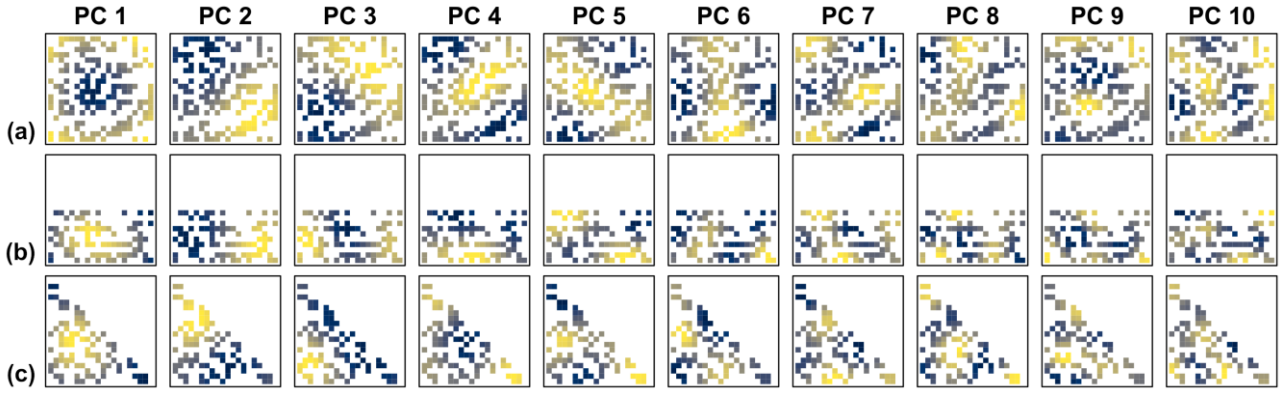

Figure 10 DD reference patterns as in Figure 8 but for a random selection of only 40 % from the cells within the three geometric domain boundaries of Figure 6. The sampling probability was homogeneous across the domain (spatially homogeneous case).

[1] In the all cells variant, PC 4 exhibits two maxima in the upper left and lower right corner and two minima in the lower left and upper right corner, PC 5 exhibits the maximum in the centre and four minima in the four corners. In the subsampling variant, PC 4 exhibits two maxima in the upper left and lower right corner and the minimum in the centre, PC 5 exhibits basically the same structure but rotated by 90°.

[2] For PC 10, the patterns of the all cells variant are for all domains already so fine structured that the subsampling results in quite different patterns.

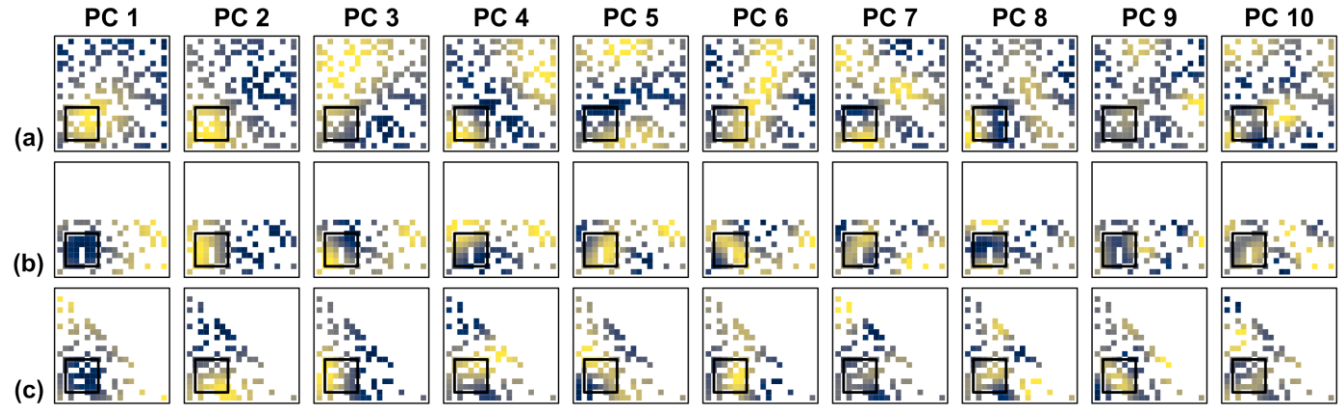

Figure 11 DD reference patterns as in Figure 8 but for a random selection of only 40 % from the cells within the three geometric domain boundaries of Figure 6. The sampling probability within the small square in the lower left was three times higher than in the rest of the domain (spatially heterogeneous case).

500

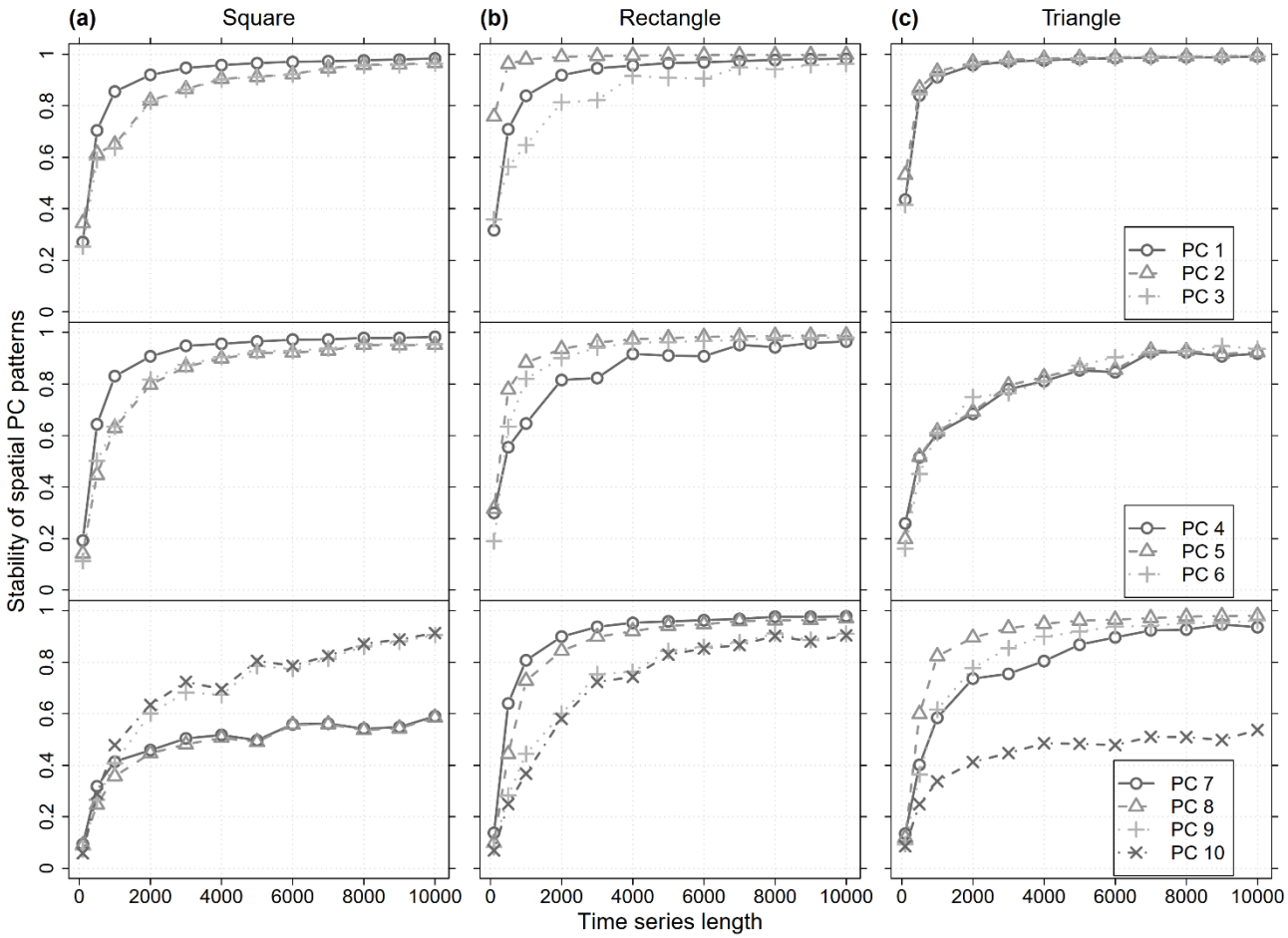

Figure 12 Stability of the spatial PC patterns as in Figure 7 but for the patterns of the homogeneous subsampling variant (Figure 10.

	Hom PC	1	2	3	4	5	6	7	8	9	10
Square	aCC	**1**	**0.99**	**0.96**	0.73 \5	0.64 \4	**0.88**	**0.85**	0.74	0.77 \10	0.62 \9
	R^2	0.94	0.97	0.93	0.52 \5	0.45 \4	0.79	0.73	0.55	0.59 \10	0.38 \9
Rectangle	aCC	**1**	**0.98**	**0.90 \4**	**0.88 \3**	**0.90**	0.72 \7	0.56	0.66 \9	0.64 \8	0.48
	R^2	0.87	0.97	0.80 \4	0.78 \3	0.83	0.52 \7	0.32	0.43 \9	0.40 \8	0.23
Triangle	aCC	**0.99**	**0.96**	**0.96**	**0.93**	**0.96**	**0.89**	**0.91**	**0.92**	0.65 \10	0.53 \9
	R^2	0.89	0.95	0.92	0.86	0.93	0.80	0.84	0.84	0.42 \10	0.28 \9

Table 1 Best matches between the DD patterns of the square, rectangular and triangular domains from the homogeneous subsampling variant (Figure 10) and the patterns formed by the corresponding subsets from their all cells counterpart (that is, the all cells patterns (Figure 8 (a) to (c)) clipped with the coordinates of the subsampling variant (Figure 10 (a) to (c))), quantified by the absolute values of the Congruence Coefficient (aCC) and R^2. Mostly, the best matches were of identical PC rank. If the best match was with an all cells pattern subset of different rank, that rank is given after the "\". Hom PC: PC ranks from the homogeneous subsampling variant. Bold aCC values indicate fairly similar PC patterns, grey shaded and bold aCC values PC patterns that can be considered equal (Section 3.3).

	Het PC	1	2	3	4	5	6	7	8	9	10
Square	aCC	**0.93**	**0.82 \3**	**0.97 \2**	0.67	0.77 \6	0.55 \5	0.65	0.66 \9	0.50	0.54
	R^2	0.56 \3	0.85 \3	0.97 \2	0.43	0.61 \6	0.32	0.44	0.43 \9	0.25	0.30
Rectangle	aCC	**0.91**	0.65	0.73 \4	**0.93 \3**	0.69 \6	0.60 \7	0.69 \10	0.61	0.75	0.43 \7
	R^2	0.85 \2	0.76	0.52 \4	0.85 \3	0.48 \6	0.37 \7	0.49 \10	0.39	0.56	0.18 \7
Triangle	aCC	**0.98**	0.80	0.61	0.71	**0.93 \6**	0.53 \8	0.64	0.62 \10	0.39 \10	0.70 \9
	R^2	0.73	0.64	0.43	0.49	0.86 \6	0.29 \8	0.42	0.39 \10	0.15 \10	0.49 \9

Table 2 As in Table 1 but for the heterogeneous subsampling variant (Figure 11).

4.3. Effects of the domain size and spatial correlation length

The ratio between domain size and the spatial correlation length affects the fractions of variance allocated to the PCs (Figure 13) as well as the contrasts of the spatial PC patterns (Figure 14). If there is no spatial correlation (spatial "white noise"), the spatial patterns of all PCs are white noise. All PCs represent the same fraction of variance, one divided by the total number of PCs. The magnitudes of the contrasts of their spatial patterns are small and on the same level. For spatial correlation length increasing from zero towards infinity, the data sets´ series from all locations get more and more similar, converging towards identity of all series (perfect correlation). If the latter is reached, there is no variance in the data that could be distributed and, consequently, there are no patterns or contrasts in the PC patterns. In between the two extremes, successive allocation of variance to the PCs and spatial PC patterns with distinct contrasts appear.

For the variance allocation, increasing correlation lengths result in increasing accumulation of variance in the leading PCs, converging towards accumulation of the total variance in PC 1.

For the contrasts, it is more complex. The maximum contrasts appear for correlation lengths in the order of magnitude of the domain size. The exact maximum is specific for the different PCs and depends on the particular domain shape. For example, for the triangular domain here (Figure 14c), the contrasts of the PC 1 patterns peak at a correlation length of 13 cells, the ones of the PC 2 patterns at a correlation length of 21 cells (not shown). The increase of the contrasts between zero correlation length and the correlation lengths of the maximum contrasts reflects the increasing fraction of covarying

locations that support the poles of the DD patterns. The decrease of the contrasts between the correlation lengths of the maximum contrasts and infinite correlation length reflects the increasing similarity of all locations which leads to smoother spatial PC patterns with contrasts converging towards zero.

Within a DD sequence, the magnitude of the contrasts differs between the PCs. Generally, they peak at PC 2 (Figure 14) and decay with decreasing PC order (Figure S4). In this sequence it is first the coarse structures with stronger contrasts that are described and then the more fined detailed structures which tend to be smoother (Figure 8 and Figure 14). The "spatial average" pattern of the PC 1 monopole generally exhibits contrasts on low to intermediate level compared with the "strongest contrast" pattern of the PC 2 dipole.

Substantial accumulation of variance in the leading PCs is commonly interpreted as indication for dominant processes or modes of the analysed system. In particular the combination with distinct PC patterns exhibiting strong contrasts is highly suggestive. The results demonstrate that both aspects are rather limited indicators and not sufficient for such interpretation. Quite the contrary, if spatially homogeneous autocorrelation is dominant in the data, both have to be expected.

Note also that the effect of the autocorrelation is spread over all PCs. Thus, for process identification etc., it is the question whether the features of interest cause signatures (spatio-temporal heterogeneities) distinct enough to be salient against the homogeneous background (Cahalan et al., 1996). Next question is whether they get clearly assigned to single PCs or whether they are as well smeared over several, if not all, PCs.

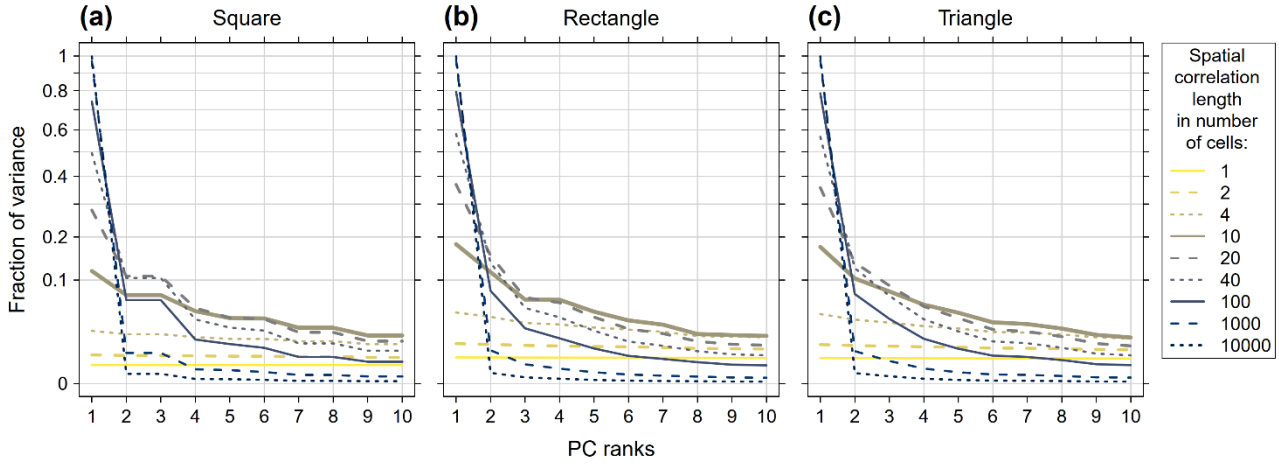

Figure 13 Variance representation of the ten leading PCs modelled with the analytic DD reference method using an isotropic exponential covariance model, nine different spatial correlation lengths and the domain boundaries (a) square, (b) rectangle and (c) triangle from Figure 6. All cells within the boundaries were used. The scale of the Y-axis is square root transformed for better readability.

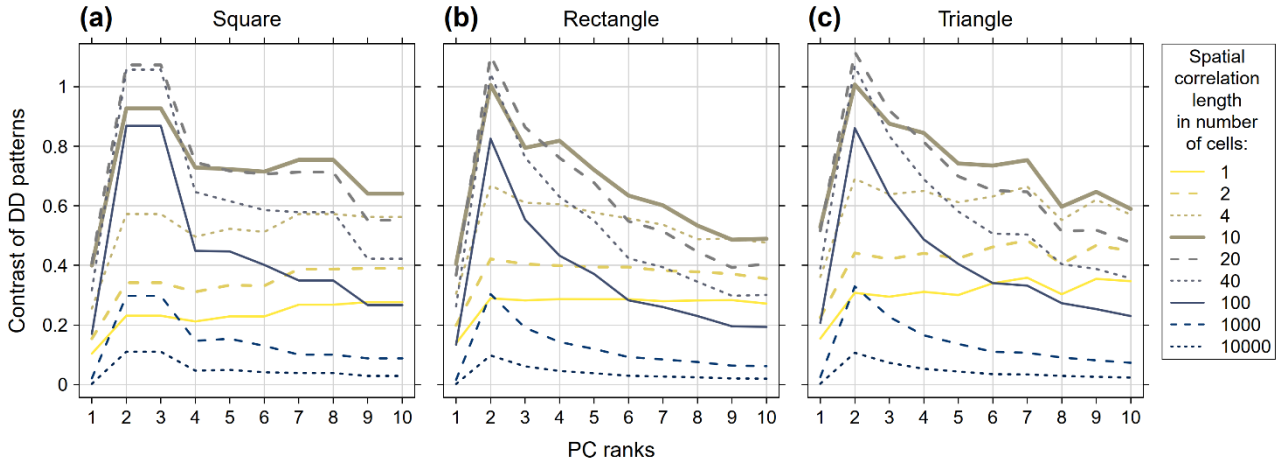

Figure 14 As in Figure 13 but for the contrasts of the DD patterns.

4.4. Effectively degenerate multiplets

Effectively degenerate multiplets are PCs with consecutive ranks, often PC pairs, which are not well separated by the PCA (North et al., 1982). They are indicated by noticeably similar eigenvalues (fractions of explained variance) considering their position in the ranking of the PCs, e.g. PC 2+3 in Figure 13a and PC 3+4 in Figure 13b, both for spatial correlation length of 10 cells. Within the subspace spanned by the multiplets' eigenvectors their rotation is arbitrary. All eigenvectors of the multiplet are needed to adequately describe the multiplets' subspace. Consequently, the multiplet should not be split for summarizing the data set, interpretation, further analysis (North et al., 1982) or rotation (Jolliffe, 1987, 1989). In particular, special care has to be taken that the truncation point of a PCA does not split a multiplet (North et al., 1982). The concept of effectively degenerate multiplets (short: effective multiplets, as in Wilks (2006)) is closely related to degeneracy of eigenvalues. For clarification we provide a brief introduction in Appendix B.

In S-Mode PCA, spatial and temporal patterns are associated to the PCs. Often the hope is that the leading PCs represent the dominant spatio-temporal features of the data set. In case of effective multiplets, one spatio-temporal feature is described by the two or more PCs forming the multiplet. This feature can be described with any linear combination of the spatio-temporal patterns of the involved PCs (Appendix B). For example, a degenerated PC pair could indicate "a signal that is propagating in space" (von Storch and Zwiers, 2003; Roundy et al., 2015) like the Madden-Julian Oscillation (Kessler, 2001). Note that such signal might be further modified by lower ranked PCs that are clearly separated from the degenerated pair (Kessler, 2001; Roundy et al., 2015). Thus, at first glance an effective multiplet could be considered indicative for a rather complex spatio-temporal feature. But as we showcase here, it might as well simply result from DD.

In the only spatial correlation case applied here, the temporal PC patterns are white noise (Section 3.2.1). Thus, the issue of one spatio-temporal feature being represented by two or more PCs is reduced to spatial features only. Effective multiplets are built by PCs of which the orientation of their eigenvectors, i.e. their spatial patterns, in the DD sequence is ambiguous. Therefore, their patterns are very sensitive to even small variations in the analysed data. All the multiplet members in combination describe a spatial feature of the data set. Thus, in case of a degenerated pair, a variation in the one pattern implies a complementary variation in the other. For the simple geometric shapes here, the pair´s spatial patterns from an ensemble of data sets simulated with identical spatial domain and spatial correlation properties will usually exhibit two predominant patterns with ambiguous ranking. Gradual variations of the predominant patterns and the switches in the ranking result simply from the randomness of the simulations.

For example, the two predominant spatial patterns of the degenerated pair formed by PC 3 and 4 from data sets simulated with rectangular domain (20 × 10 cells), isotropic spatial correlation with exponential decay and a correlation length of 10 cells (Figure 8b) randomly switch rank between distinct data sets (Figure 15). This results in the low stability of the PC 3 and 4 patterns from the respective ensemble (Figure 7b). For both PCs, the correlation of the ensemble´s patterns converge for long simulated time series around a mean R^2 of 0.5, indicating that the degeneracy of this pair cannot be resolved with longer time series. The complementarity of both parts of the pair is visible in the ensemble´s mean and standard deviation patterns. The standard deviation pattern of PC 3 reflects an absolute variant of the mean spatial pattern of PC 4, and vice versa (Figure 9b and Figure 8b). The R^2s of the correlation between the two patterns were 0.64 and 0.78, respectively. The aCCs of the two patterns were 0.95 and 0.96.

Note, however that degeneracy might cause domain dependent patterns that don´t seem to be DD patterns because they are intermixed into new patterns. For example, in Figure 15 the patterns of the multiplet pairs of simulations 1, 4 and 5 exhibit different patterns than those of simulations 2 and 3.

Symmetry of the domain shape triggers degeneracy (North et al., 1982). Thus, generally, it is recommendable to check spatial PCA results from data with symmetric domains for DD induced degeneracy. For example, the data sets simulated with the square domain yielded the four effectively degenerated PC pairs PC 2+3, PC 5+6, PC 7+8 and PC 9+10 (Figure 7a). Again, the complementarity within the pairs yields standard deviation patterns of the one PC reflecting an absolute version of the mean spatial patterns of the counterpart PC (Figure S2). The match of the respective two patterns was very close, with all R^2s being larger than 0.80 and all aCCs being larger than 0.95.

Asymmetrical distribution of locations diminishes the probability of DD induced degeneracy. In the subsampling variants (Figure 12 and Figure S3) most of the degenerated PC pairs of the all cells variant (Figure 7) disappeared. The subsampling reduced the symmetry of the domain shape, resulting in a less ambiguous orientation for the eigenvectors. Consequently, the order of the DD sequence is clearer defined.

Effective degeneracy depends not only on the spatial domain but also on the effective sample size of the series, which equals here the time series lengths (Section 3.4). For example, in the triangular domain the effective degeneracy of the PC pair 6+7 which is prominent at a time series length of 2000 gradually disappears with increasing time series length (Figure 7c). However, for very symmetric domains no sample size might be sufficient to resolve the degeneracy (e.g., see PC 2+3 and 5–10 of the square domain in Figure 7a or Table II in Richman, 1986).

Commonly, degenerated multiplets are detected qualitatively by checking for noticeably similar eigenvalues of PCs with adjacent ranks, forming steps in the sequence of the PC eigenvalues, or quantitatively with North´s rule of thumb (Figure S6). Analogue steps in the sequence of contrasts can serve as additional indication (Figure 14). With the stochastic DD reference method these steps are particularly pronounced, standing out as PCs of adjacent ranks with similar and rather low contrasts given their position in the DD sequence, e.g. PC 2+3, PC 5+6, PC 7+8 and PC 9+10 for most correlation lengths in Figure S5a, and PC 3+4 for spatial correlation length of 10 cells in Figure S5b. It is an effect of averaging patterns that switch ranks between the data sets from an ensemble. The magnitude of the drop depends on the specific patterns that are averaged.

Note also, that intermixing might be easier overlooked for the smaller eigenvalues that are more closely spaced. If the analysist selects PCs to separate noise from signal, this could possibly result in truncation within a multiplet and consequently intermixing of noise and signal in the last considered PCs. Here, we selected the first ten PC merely for the illustrative purpose (Section 4.1). If the goal would be to further analyse PC 10, it would be necessary to check its patterns for intermixing - also with the subsequent PCs, in particular PC 11. Indications for intermixing in the PC 10 pattern can be seen in the stability plots of Figures Figure 7a, Figure 12c, Figure S3a+c. In case of Figure 12c, PC 9 does not show sign of intermixing, thus, in this case the intermixing is probably with PC 11.

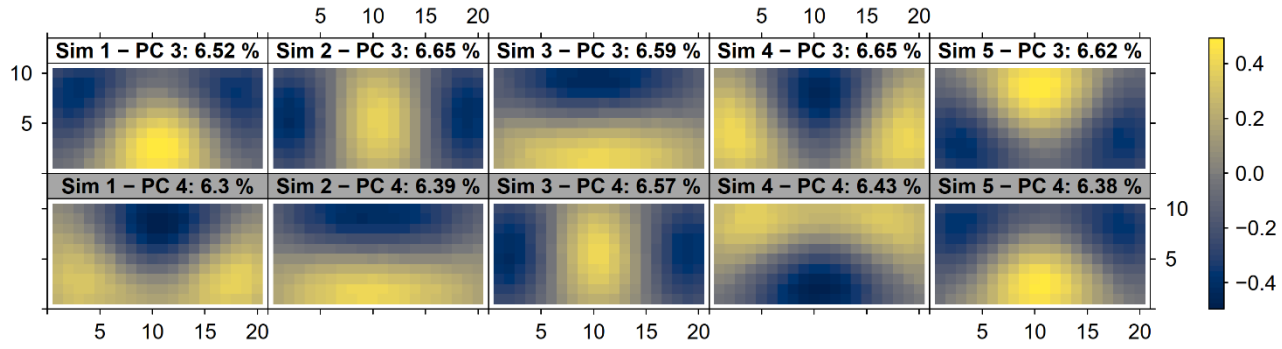

Figure 15 Spatial patterns of the degenerated PCs 3 and 4 from five distinct data sets, each simulated with rectangular domain (20 × 10 cells), isotropic exponentially decaying spatial autocorrelation of correlation length 10 cells and time series length 10 000. Identical properties were used to simulate the ensembles from Figure 7b, Figure 8b and Figure 9b. The spatial patterns that belong to the same data set are plotted above each other with PC 3 on the top (white panel titles) and PC 4 on the bottom (grey panel titles). The index of the simulated data set and the fraction of assigned variance is given in the panel titles.

5 Approaches to consider DD

The preceding section introduced different aspects of DD. It demonstrated the strong effect the domain´s size and shape can have on spatial PC patterns from S-mode PCA. In this section we continue with suggestions how to detect and diminish DD.

5.1. Detecting DD

5.1.1. Comparing spatial PC patterns from markedly different subdomains

The simplest way to check whether the spatial PC patterns of a data set are affected by DD is to visually compare the spatial PC patterns from sub-data sets with markedly different domains. Such a comparison can serve as quick qualitative check to detect cases in which DD is a prominent feature. We recommend partitioning of the original domain with basic geometric domain shapes like we used here. Thus, first take a subset with a square shaped domain, then taking from the square domain further subsets with rectangular and triangular domains of different orientation and compare the spatial PC patterns of these subsets. This proceeding is demonstrated in the associated demo scripts (Lehr, 2024).

A real-world data case is shown in Figure 16. Three sets of spatial PC patterns with square, rectangular and triangular domains were derived from raster of monthly precipitation sums from the years 1991 to 2020 in northeast Germany (Section 2.2). The square domain is the 200 km × 200 km square from the 1 km × 1 km precipitation grid in Figure 3. The rectangular and triangular domain were fitted in the square domain, analogue to the proceeding with the synthetic examples (Figure 6). Thus, the data sets consist of time series with 360 months length and 40 000 locations in case of the square domain, and 20 000 locations in case of the rectangular and triangular domains. The DD of the spatial PC patterns is clearly visible (Figure 16). Visually, the spatial PC patterns appear as noisy variants of the already well-known Buell patterns (Figure 1 and Figure 8). The very strong accumulation of variance in the centred monopole pattern of PC 1 (Table 3, Figure 16) is another indication for DD. Thus, in this case the quick check already clarifies the DD of the PCA results.

If the subdomains are of similar size, the focus is primarily on the domain shape aspect. Analogue, the PCA results can be checked for dependency from the selected domain size. However, we assume that commonly an analysis is focused on a specific scale and the domain size as well as the interpretation of results fit to that scale. Thus, usually the dependency from the domain´s shape should be more an issue than the dependency from its size.

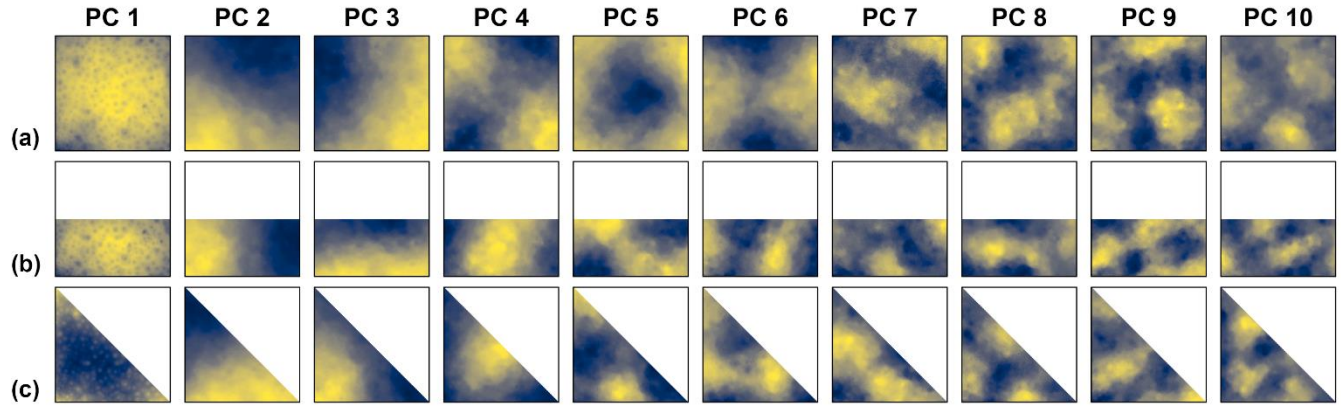

Figure 16 Overview of the leading ten spatial PC patterns from the PCAs of the precipitation data with the square, rectangular and triangular domain. The location of the square domain is marked in Figure 3. The two other domains are fit in the square domain.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
Square	79.96	4.93	3.69	1.58	1.16	0.90	0.55	0.44	0.44	0.37
Rectangle	83.22	5.06	2.49	1.30	0.92	0.67	0.61	0.42	0.36	0.32
Triangle	82.79	5.04	3.01	1.35	0.78	0.61	0.56	0.47	0.39	0.31

Table 3 Fractions of assigned variances in percent from the PCAs of the precipitation data with the square, rectangular and triangular domain (Figure 16).

5.1.2. Comparison with DD reference patterns

DD reference patterns can be tailored for defined spatial domain and spatial correlation properties of a data set. Spatial PC patterns can be visually compared against the reference or checked for significant deviations from the reference with the congruence coefficient or simple correlation analysis (Table 1). With the stability analysis of the stochastic method (Figure 7) or the confidence intervals of the analytic method (Figure S6) it can be identified for each PC rank which time series length is required to reach stable and clearly defined DD patterns. Consequently, PCA results from (observed) data sets with identical spatial domain and spatial correlation properties but shorter time series have to be interpreted with the reservation that the DD might be stronger than the comparison with the reference suggests.

As a real-world data case we look again at the precipitation PCAs (Figure 16). DD reference patterns were fitted for all three domains using an isotropic spherical covariance model and the analytic method (Figure S7). The spatial patterns of the leading precipitation PCs exhibited strong similarity with their DD reference counterparts (Table 4), clearly indicating DD. For the first four PCs, the main difference was the separation of PCs 2+3 from the square domain in the precipitation PCAs (Figure S8) which form a multiplet in the DD reference (Table S1). Meaning, while typical DD patterns occurred, the deviations of the precipitation PCA patterns from the pure theoretical DD case were strong enough to result in a clear ranking. In accordance with the findings of the synthetic experiments (Figure 13), the very large fraction of variance assigned to PC 1 of the precipitation PCAs (Table 3) is reflected in the very large theoretical correlation length of the DD reference (Figure S7), being substantially longer than the domain size.

That the patterns of the low ranked PCs exhibit stronger deviations from the DD reference than those of the leading PCs is no indication against DD. Recall that PCs should not be analysed in isolation, but only in reference to all PCs with preceding ranks. If the leading PCs exhibit DD, DD for the whole sequence of PC patterns can be concluded. It is not necessary to find DD reference patterns that perfectly fit to the patterns of the low ranked PCs for such conclusion. Because of the more finely structured spatial patterns, it can be expected that the patterns of the low ranked PCs from real-world data will deviate stronger from the DD reference than those of the leading PCs. Thus, the comparison with the DD reference confirmed the finding of strong DD from the visual comparison of the patterns from the three domains (Section 5.1.1).

We introduced building DD reference patterns for data sets exhibiting isotropic spatial but no temporal autocorrelation. It enables to test the null hypothesis that the spatial PC patterns from observed data merely result from simple isotropic spatial autocorrelation between random white noise time series. The main feature of the null hypothesis is the ratio of spatial correlation length to the domain size, in particular to the distances between the data set´s locations. To our knowledge, such test was suggested first by Cahalan et al. (1996). They fitted models to observed precipitation and temperature data and compared the eigenvalues and spatial PC patterns of observed and modelled data. Significant differences between the two eigenvalue spectra were considered to be “signal” and indicative for spatial anisotropies and inhomogeneities, “inhomogeneous processes”, combined space and time correlation, or (secular) trends.

However, DD is not restricted to the isotropic case (see “directional functions” in Buell (1975)). An anisotropic example for our three basic domains is given in the supplements. Compared with the "default" isotropic case, the DD patterns are distorted according to the direction and the ratio between longest and shortest spatial correlation length of the anisotropy (Figure S9 vs Figure 8). The spatial PC patterns tend to stabilize for shorter time series length (Figure S10 vs. Figure 7) and the PCs which form degenerated pairs are better separated (see the bigger differences between the fractions of assigned variance and the smaller magnitudes of the standard deviation patterns in Figure S11 vs. Figure S2). Both aspects reflect that the anisotropy gives a less ambiguous orientation for the DD sequence.

Elaborating on the DD of PC patterns from data sets with homogeneous autocorrelation in space and time is beyond the introductory scope here. However, spatially inhomogeneous temporal trends are indicative for distinct processes, modes or alike. They are likely to spread over more than one PC (Hannachi et al., 2007; Hannachi, 2007) and to affect the variance distribution among the PCs (Vejmelka et al., 2015). Thus, if the goal is not DD assessment but to construct reference patterns for the identification of distinct features, they should be considered.

DD reference patterns are rather well behaved. The main decisions for their construction are the choice between an isotropic or an anisotropic model, and the selection of the correlation length. The first primarily defines the typical patterns that appear (e.g., Figure 8 versus Figure S9), the second the variance distribution (Figure 13, Section 4.3). In comparison, the effects of different spatial covariance model types like exponential, gaussian or spherical are less important. For practical applications, the comparison with the spatial patterns is the main point rather than the exact reproduction of the variance distribution. A perfect fit is not required. The spatial patterns are very similar for a wide range of correlation lengths. This holds in particular for those of the leading PCs which are commonly used in practical applications.

	Precip PC	1	2	3	4	5	6	7	8	9	10
Square	aCC	**1**	**0.95 \3**	**0.95 \2**	**0.95**	**0.94**	**0.94**	0.81	0.84	0.74 \10	0.54 \9
	R^2	0.77	0.91 \3	0.90 \2	0.90	0.88	0.88	0.65	0.71	0.54 \10	0.29 \9
Rectangle	DDref PC	**1**	**0.99**	**0.98**	**0.94**	**0.91**	**0.85**	0.67	0.77	0.44 \7	0.35 \9
	R^2	0.73	0.98	0.96	0.88	0.83	0.72	0.45	0.59	0.19 \7	0.12 \9
Triangle	DDref PC	**1**	**0.95**	**0.93**	**0.96**	**0.91**	**0.86 \7**	**0.89 \6**	0.69 \9	0.77 \8	0.69
	R^2	0.76	0.90	0.86	0.93	0.82	0.74 \7	0.80 \6	0.47 \9	0.59 \8	0.48

Table 4 As in Table 1 but for the comparison of the spatial PC patterns from the precipitation data (Figure 16) and the corresponding DD reference patterns (Figure S7).

5.2. Approaches to diminish DD

5.2.1. Subsampling of domains

Analysing a subsampled data set with enlarged minimal distance between the locations can be used to diminish the DD of the PCA results. Reducing the symmetry of the domain can remove effective multiplets. Both can help to carve out features other than DD. On the other hand, informative local details might be filtered out together with the excluded locations. If there is still DD, the new DD patterns of the subsampled data set might be harder to recognize visually because of the smaller number of locations per area. The selected minimal distance, respectively the selection of locations, is critical for the analysis. Depending on the choice, different features in the results might stick out, get diminished or even disappear. In any case, the spatial resolution of the analysed data set has to be considered in the interpretation of the results. Also, only stable PC patterns should be used to draw conclusions on the analysed system. The stable PC patterns are those which are rather

insensitive to the specific selection of analysed locations. They can be identified by comparing the PCA results from different subsamples (Smirnov, 1973; Lins, 1985a; Lehr and Lischeid, 2020).

5.2.2. Rotation of PC eigenvectors

Another option that can diminish DD is to rotate the eigenvectors from the PCs of interest (Richman, 1986; Dommenget, 2007; Compagnucci and Richman, 2008). Often unrotated PCA results exhibit DD patterns, while rotated PCA seem to be less affected (Richman, 1986; Huth and Beranova, 2021). This finding is supported by experiments using synthetic data. Compagnucci and Richman (2008) analyzed different synthetic sequences of basic sea level pressure flow patterns ("plasmodes"). The unrotated S-mode patterns were systematically affected by DD. In the rotated variants the DD patterns vanished.

Exemplarily, we varimax rotated the leading spatial PC patterns of the precipitation PCAs (Figure 16) in three variants, using the first two PCs (2rPCs), the first three PCs (3rPCs) and the first four PCs (4rPCs) (Figure 17). The first four precipitation PCs were clearly separated in all three domains (Figure S7), thus, no multiplets were split by the rotations (Section 4.4). As expected, the variance distribution among the rotated PCs (Table 5) was much more even compared to the unrotated PCs (Table 3). Note, that the newly assigned fractions of variance do not any longer decrease continuously with the PC ranks in all cases. Note also, that the fractions of variance that are assigned to distinct patterns, for example to the diagonal gradient of the triangular domain, depend on the number of PCs that are rotated (Table 5). The magnitude of the pattern contrasts was more evenly distributed among the rotated PCs (Table S2) than among the unrotated PCs (Table S3). Most of the rotated patterns exhibited only positive or only negative loadings (Table S2), indicating a more "simple structure" (Section 3.5; Richman, 1986) than the unrotated patterns (Table S3).

In all three rotation variants, the patterns were clearly dependent on the domain geometries (Figure 17). For example, the patterns of the 2rPCs variant showed gradients from southwest to northeast in the square domain, from west to east in the rectangular domain and from north-west to south-east in the triangular domain. Thus, in our case here, varimax rotation was not successful in resolving DD. Instead, the patterns of the rotated PCs seemed to be the varimax way of displaying DD. The dominant PC 1 monopole of the unrotated PCA disappeared and the new dominant patterns are gradients reflecting the domain shape. For example, the gradients of the square domain from the 4rPCs variant reflect the rotational symmetry of the square (Figure 17a, right panel), or the gradients of the rectangular and triangular domain associated with the major fractions of variance (Table 5) depict in all three rotation variants the longest extent of the domain (Figure 17bc). The examples demonstrate that while rotated eigenvectors are generally considered to be less prone to DD (Richman, 1986; Wilks, 2006), there is no guarantee that rotation will remove or even diminish DD (NCAR, 2013).

Except from being less prone to DD, rotated PCA results were found to be more robust against spatial (Richman, 1986) and temporal (Cheng et al., 1995) subsampling and less sensitive to degeneracy (Richman, 1986). Rotation can support the interpretation of effective multiplets if the resulting PCA patterns are of more simple structure (Jolliffe, 1987; 1989). Rotating only multiplet members limits thereby the drawbacks of rotation (Section 3.5) to the multiplet (Jolliffe, 1989). Rotated PCA results were also found to be easier to interpret physically (Richman, 1986). Rotation can be used to systematically relax distinct PCA constraints that hamper physical interpretation (Hannachi et al., 2007; Monahan et al., 2009) by choosing between orthogonal and oblique rotation and selecting a simplicity criterion that suits best to the analysed system. In the aforementioned analysis of synthetic sea level pressure flow patterns, Compagnucci and Richman (2008) found the rotated PC patterns to be superior in depicting the "true" flow patterns. In a study using atmospheric reanalysis data, Huth and Beranova (2021) compared the spatial patterns from four PCA derived modes of climatic variability with autocorrelation maps of the analysed data to identify the true modes of climatic variability. Only the one mode based on

rotated PC patterns (North Atlantic Oscillation) corresponded well to underlying autocorrelation patterns, the modes based on unrotated PCA did not. However, these studies indicating that rotated PC patterns are more suitable for physical interpretation focused primarily on atmospheric mode detection.

For future work, we suggest to perform a study similar to Compagnucci and Richman (2006), but with a hydrological focus. Synthetic data from a hydrological simulation model could be analyzed, to test which hydrological features of the model can be uncovered by the patterns of the PCs. The test data could be, for example, spatially distributed groundwater level series simulated with a groundwater model. The experiments could be used to compare the performance in hydrological feature identification of unrotated versus rotated PCA and orthogonal versus oblique rotation, but also of S-mode versus T-mode PCA (Richman, 1986; Compagnucci and Richman, 2006; Isaak, et al., 2018) and different scaling of the eigenvectors (Jolliffe, 1995; Wilks, 2006).

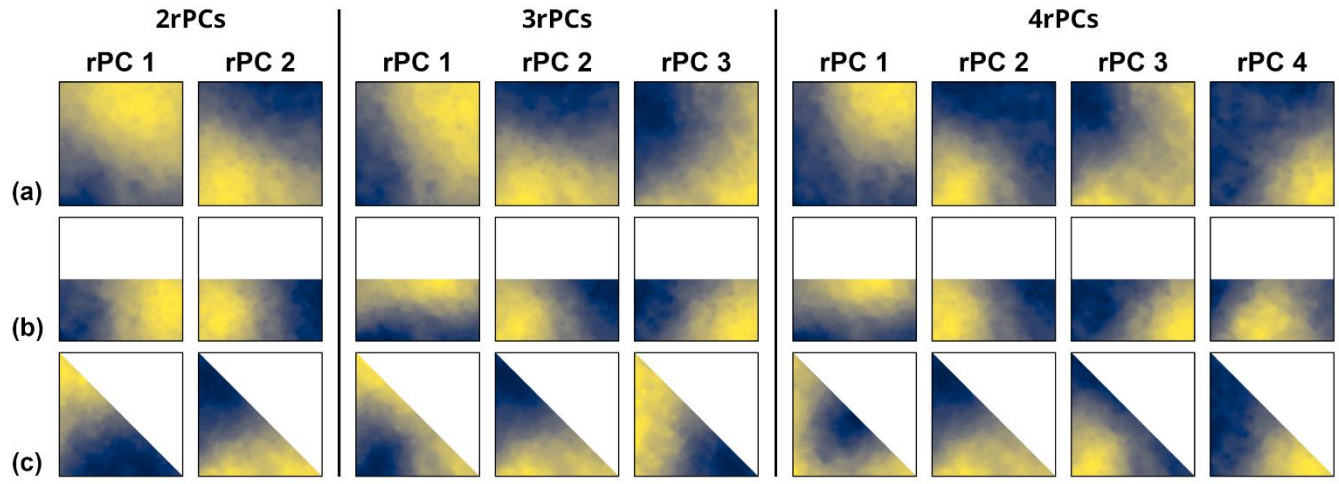

Figure 17 Leading varimax rotated spatial PC patterns from the PCAs of the precipitation data with the square, rectangular and triangular domain (Figure 16). The rotation was performed with the first two PCs (2rPCs), the first three PCs (3rPCs) or the first four PCs (4rPCs).

	2rPCs		3rPCs			4rPCs			
	rPC 1	rPC 2	rPC 1	rPC 2	rPC 3	rPC 1	rPC 2	rPC 3	rPC 4
Square	44.84	40.05	32.95	30.92	24.71	27.77	24.44	20.92	17.03
Rectangle	44.00	44.28	18.36	37.43	34.97	19.28	36.85	33.16	2.78
Triangle	46.69	41.14	26.61	34.26	29.96	4.78	32.93	26.36	28.11

Table 5 Fractions of assigned variances in percent from the varimax rotated spatial PC patterns from the precipitation PCAs (Figure 17).

6 Conclusion

Spatial patterns from S-mode PCA are regularly used for hydrological interpretations. In such analysis, homogeneous spatial correlation between the data sets` time series results in spatial PC patterns that are determined by the size and shape of the analysed spatial domain (domain dependence: DD). DD patterns are distinct, with strong gradients and contrasts. We showed that DD can come together with substantial accumulation of explained variance in the leading PCs. Thus, in contrast to what one might expect, neither distinct spatial PC patterns nor large fractions of explained variance in the leading PCs do necessarily indicate dominant hydrological processes or hydrologically meaningful properties. In addition, DD can induce effectively degenerated multiplets (effective multiplets). Without knowledge about DD, the multiplets can be misinterpreted

as indication for complex spatio-temporal features. Without knowledge about multiplets, the multiplet members can be mistaken as effects of independent hydrological processes. Without knowledge about the effects of multiplets, DD can be overlooked because the degeneracy can mask the expected DD patterns.

In summary, if DD is predominant, the spatial PC patterns do not reflect the hydrological functioning of the analysed system but rather the functioning of the PCA within the context of the data set´s spatial domain. Ignoring DD and effective multiplets easily leads to wrong hydrological interpretations. Consequently, DD should be considered for any application in which the PCs are used to draw conclusions about spatially distinct properties of the analysed system. In other words, it should be checked whether the spatial PC patterns differ significantly from patterns that result from the trivial case of nearby locations being homogeneously more related than those further apart.

Classical Buell patterns (PC 1: "mean behaviour", PC 2: gradient along the longest extent of the domain, lower ranking PCs: regular multipoles) and leading PCs with remarkably similar eigenvalues (effective multiplets) are an alert for DD. However, deviating patterns or clearly separated PCs are no contra-indication. DD patterns are original for every combination of spatial domain and spatial correlation properties. Thus, visual detection of DD is rather limited. Still, visual comparison of the spatial PC patterns from subdomains with markedly different shapes and/or sizes is practical as quick qualitative check.

To test whether spatial PC patterns differ significantly from DD patterns, reference patterns can be used as null hypothesis. For most practical applications checking the first few leading PC patterns should be sufficient. We presented two methods to produce DD reference patterns. For the introductory purpose, we focussed on the stochastic method. The comparison of data sets simulated with identical spatial domain and spatial correlation properties showed directly the ambiguity of the PC ranking within DD induced multiplets, including the variations of the predominant patterns. Furthermore, working with simulated data is less abstract than working with the analytic covariance matrix. For practical applications the analytic method is preferable. Its short computation time is a big advantage, especially when producing DD reference patterns for data sets with many locations.

Passing the check for DD and accounting for effective multiplets in the selection of the PCs are necessary but not sufficient conditions to assure physical meaningfulness. When single PCs, or combinations of PCs, are assigned to distinct hydrological features, it should be carefully considered whether the S-mode PCA constraints (i) successive maximization of variance on the PCs, (ii) orthogonality of spatial PC patterns and (iii) linear uncorrelatedness of temporal PC patterns support such interpretation. The spatio-temporal PC patterns should not only be checked for resemblance with the postulated features, but also the invariance of the spatial and temporal PC patterns against subsampling should be approved. Building on this study, a next research task could be to conduct systematic experiments with synthetic test data derived from hydrological simulation models to evaluate which PCA modes, rotation methods and scaling of the eigenvectors work best for hydrological feature identification.

Appendix A – PCA reference patterns based on the analytic covariance matrix

Deriving reference patterns with the analytic covariance matrix to evaluate PCA results was applied earlier by Cahalan et al. (1996) and Dommenget (2007). They modelled the evolution of a continuous meteorological field as stochastic spatially isotropic diffusion process, i.e. a spatial first order auto regressive (AR(1)) or "spatial red noise" process, and used the spatial PC patterns derived from the analytic covariance matrix of the model as null hypothesis for the spatial structure of climate variability.

Dommenget (2007) presented two adaptations of the analytic covariance matrix. In the first, for each pair of locations, the product of the standard deviations from the time series of the two locations is multiplied with their covariance calculated with the covariance function. The resulting spatial PC patterns provide the smooth pattern of the globally fitted covariance function weighted with the data set´s spatial distribution of covariance magnitude. In the second, the analytic covariance matrix is adapted to simulate the effect of areas with increased stochastic forcing. Areas with larger variance than the surrounding are defined and used for the weighting of the covariance matrix. In numerical experiments the effect of monopole, dipole or multipole structures in the data on the spatial PC patterns can be tested. Note that both variants are adaptations of the covariance matrix. Thus, other than in this study, the data must not be z-scaled prior PCA.

In addition, Dommenget (2007) suggested using the spatial PC patterns from an analytic covariance matrix as null hypothesis to find spatial PC patterns “that are most distinguished from those of the null hypothesis”. These so called Distinct Empirical Orthogonal Functions (DEOFs) are derived by rotating the eigenvectors of the observed data to maximum difference in explained variance between the EOFs of observed data and those of the analytic covariance matrix. A Matlab script to perform DEOF analysis is available as supplementary material to Dommenget (2007). The DEOFs were suggested as starting point to identify teleconnections patterns or physical processes. Even though not in focus, DD patterns of the null hypothesis were observed and described as hierarchy of multipoles, “starting with a monopole as EOF-1, followed by a dipole, and then by higher order multi poles”. In analogy to the spectrum of time series the DD sequence was interpreted as reflection of different spatial scales. The DEOF approach can be also used to compare the spatial variability modes from different data sets (Bayr and Dommenget, 2013). For data sets exhibiting temporal trends detrending prior applying DEOF is recommended (Hannachi and Dommenget, 2009).

Appendix B – Effectively degenerated multiplets

An eigenvalue is called degenerate if it is associated with more than one linearly independent eigenvector. That is, the eigenvalue is repeated (non-distinct), its multiplicity is larger than one. In the PCA case, the algebraic multiplicity of an eigenvalue (the multiplicity of the eigenvalue as a root of the characteristic polynomial) equals always its geometric multiplicity (the dimension of its eigenspace) (Hefferon, 2020; Meyer, 2000) because PCA performs an eigenvalue decomposition of a symmetric matrix (see “spectral theorem for symmetric matrices”, e.g. in Lay (2016) or "real spectral theorem" e.g. in Larson and Falvo (2009)). A degenerate eigenvalue together with its eigenvectors is called degenerate multiplet. The eigenvectors span the subspace of the degenerate multiplet. Within this subspace their orientation is not uniquely defined and they can be arbitrarily rotated (von Storch and Zwiers, 2003). Any linear combination of the eigenvectors from the multiplet is as well an eigenvector of the eigenvalue (North et al., 1982; Hefferon, 2020).

In real-world data sets, perfectly symmetric distribution of variance such that degeneracy in the strict sense appears is unlikely to happen. However, if the eigenvalues of the “true population” are of very similar size, the sampling variability and errors can lead to “effective degeneracy” (North et al., 1982) with eigenvalues that are “indistinguishable within their uncertainties” (Hannachi et al., 2007) and eigenvectors that are random mixtures of the true population´s eigenvectors (North et al., 1982).

This shall be illustrated with a slightly extended variation of an illustration given by Wilks (2006). We start with degenerated multiplets in the strict sense, i.e. an eigenvalue with more than one eigenvector. Consider a 3D point cloud with perfectly spheroid shape (idealized rugby ball). It has one long axis and two short axes of identical size. The cloud´s first eigenvector is aligned with the long axis and its eigenvalue depicts the variance of the cloud in this direction. The second and the third eigenvector can be any pair of orthogonal vectors that are orthogonal to the long axis. They share a common eigenvalue. Thus, the variance representation is split in equal parts in the plane orthogonal to the first eigenvector. If we compare the eigenvectors from random subsamples of this data set, the orientation of the first one would be very stable, while the orientation of the second and third would exhibit large sampling variability. This correctly reflects the ambiguous orientation of the second and third true population eigenvector.

In the "effective degeneracy" case the eigenvalues are merely of very similar size. Consider again a spheroid shaped cloud but this time with the two shorter axes being of slightly different size (a slightly deflated rugby ball squeezed perpendicular to its long axis). Now the orientation of the second and third true population eigenvectors is distinct and both have distinct eigenvalues. Their share to the variance representation differs. If we compare again the eigenvectors from random subsamples of the data set, the question is whether the sampling is accurate enough to detect the slight difference in size of the two shorter axes and whether the detection of the difference is stable among the subsamples? If this is not the case, the second and third sample eigenvalues are "effectively degenerate". Together with their eigenvectors they build an "effective degenerate multiplet". Thus, again the orientation of the second and third eigenvectors exhibits large sampling variability but this time because of the limited sampling accuracy. Due to the ambiguity of their orientation the pair is a potentially arbitrary mixture of the unknown true population eigenvectors (Wilks, 2006). Within the "accuracy range" determined by the subsampling, the fraction of the cloud´s variance depicted by the plane orthogonal to the first eigenvector is approximated with the ratio of the sum of the multiplets´ eigenvalues to the sum of all three eigenvalues.

Author contribution

CL developed the stochastic method and the scripts, designed the experiments and carried them out. CL prepared the manuscript with contributions from TH.

Competing interests

The authors declare that they have no conflict of interest.

Code availability

A selection of scripts accompanying this technical note is freely available at https://doi.org/10.5281/zenodo.11213430 (Lehr, 2024). It contains: (1) a demo in which the DD of PCs is demonstrated by visual examination of the spatial PC patterns from single simulated data sets, (2) an implementation of the stochastic DD reference method (Section 3.2.1), and (3) an implementation of the analytic method (Section 3.2.2) based on Dommenget (2007) and the associated Matlab scripts. The user can define domains with distinct sizes and shapes, and the spatial correlation properties. The scripts and their documentation can directly be used for educational purposes. We recommend going first step by step through the demo to get into the functioning and logic of the scripts. For the demo and the stochastic reference script, it is best to start with the pdf documentation which includes a formatted version of the script, extra annotations and sample results. All scripts are written in R (R Core Team, 2019).

Data availability

Data sets containing the spatial domains and spatial correlation properties used in this technical note can be produced with the associated scripts. The grids of the monthly precipitation sums are freely available at the German Weather Service (https://opendata.dwd.de/climate_environment/CDC/grids_germany/monthly/hyras_de/precipitation/). Here, the version HYRAS-DE-PR v6.0 was used.

Acknowledgements

We thank Philipp Rauneker and Katharina Brüser from the Leibniz Centre for Agricultural Landscape Research (ZALF) and Marcus Fahle from the German Federal Institute for Geosciences and Natural Resources (BGR) for the fruitful discussions. Furthermore, we thank Gunnar Lischeid from ZALF for the in-depth discussions and detailed comments.

Financial support

The authors thank the ZALF and the University of Potsdam for the possibility to use the institutional resources.

References

Bartlein, P. J.: Streamflow anomaly patterns in the U.S.A. and southern Canada — 1951–1970, Journal of Hydrology, 57, 49–63, https://doi.org/10.1016/0022-1694(82)90102-0, 1982.

Bayr, T. and Dommenget, D.: Comparing the spatial structure of variability in two datasets against each other on the basis of EOF-modes, Climate Dynamics, 42, 1631–1648, https://doi.org/10.1007/s00382-013-1708-x, 2013.

Buell, C. E.: The topography of the empirical orthogonal functions, Preprints, Fourth Conference on Probability and Statistics in Atmospheric Sciences, 18–21 November 1975, Tallahassee, Florida, USA, American Meteorological Society, 188–193, 1975.

Bieri, C. A., Dominguez, F., and Lawrence, D. M.: Impacts of Large-Scale Soil Moisture Anomalies on the Hydroclimate of Southeastern South America, Journal of Hydrometeorology, 22, 657–669, https://doi.org/10.1175/JHM-D-20-0116.1, 2021.

Buell, C. E.: On the physical interpretation of empirical orthogonal functions, Preprints, Sixth Conference on Probability and Statistics in Atmospheric Sciences, 9–12 October 1979, Banff, Alberta, Canada, American Meteorological Society, 112–117, 1979.

Cahalan, R. F., Wharton, L. E., and Wu, M.-L.: Empirical orthogonal functions of monthly precipitation and temperature over the United States and homogeneous stochastic models, Journal of Geophysical Research: Atmospheres, 101, 26309–26318, https://doi.org/10.1029/96jd01611, 1996.

Cheng, X., Nitsche, G., and Wallace, J. M.: Robustness of Low-Frequency Circulation Patterns Derived from EOF and Rotated EOF Analyses, Journal of Climate, 8, 1709–1713, https://doi.org/10.1175/1520-0442(1995)008<1709:ROLFCP>2.0.CO;2, 1995.

Compagnucci, R. H. and Richman, M. B.: Can principal component analysis provide atmospheric circulation or teleconnection patterns?, International Journal of Climatology, 28, 703–726, https://doi.org/https://doi.org/10.1002/joc.1574, 2008.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., and McLoone, S.: Principal Component Analysis on Spatial Data: An Overview, Annals of the Association of American Geographers, 103, 106–128, https://doi.org/10.1080/00045608.2012.689236, 2013.

Deutscher Wetterdienst, Raster data set precipitation sums in mm for Germany - HYRAS-DE-PR, Version v6.0, https://opendata.dwd.de/climate_environment/CDC/grids_germany/monthly/hyras_de/precipitation/, 2025, last.access: 26 March 2025.

Dommenget, D.: Evaluating EOF modes against a stochastic null hypothesis, Climate Dynamics, 28, 517–531, https://doi.org/10.1007/s00382-006-0195-8, 2007.

Hannachi, A., Jolliffe, I. T., Stephenson, D. B., and Trendafilov, N.: In search of simple structures in climate: simplifying EOFs, International Journal of Climatology, 26, 7–28, https://doi.org/https://doi.org/10.1002/joc.1243, 2006.

Hannachi, A.: Pattern hunting in climate: a new method for finding trends in gridded climate data, International Journal of Climatology, 27, 1–15, https://doi.org/10.1002/joc.1375, 2007.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B.: Empirical orthogonal functions and related techniques in atmospheric science: A review, International Journal of Climatology, 27, 1119–1152, https://doi.org/10.1002/joc.1499, 2007.

Hannachi, A. and Dommenget, D.: Is the Indian Ocean SST variability a homogeneous diffusion process?, Climate Dynamics, 33, 535–547, https://doi.org/10.1007/s00382-008-0512-5, 2009.

Hefferon, J.: Linear Algebra, 4th ed., http://joshua.smcvt.edu/linearalgebra, 2020.

Hohenbrink, T. L., Lischeid, G., Schindler, U., and Hufnagel, J.: Disentangling the Effects of Land Management and Soil Heterogeneity on Soil Moisture Dynamics, Vadose Zone Journal, 15, https://doi.org/10.2136/vzj2015.07.0107, 2016.

Horel, J. D.: A Rotated Principal Component Analysis of the Interannual Variability of the Northern Hemisphere 500 mb Height Field, Monthly Weather Review, 109(10), 2080–2092, https://doi.org/10.1175/1520-0493(1981)109<2080:ARPCAO>2.0.CO;2, 1981.

Huth, R. and Beranová, R.: How to Recognize a True Mode of Atmospheric Circulation Variability, Earth and Space Science, 8, e2020EA001275, https://doi.org/https://doi.org/10.1029/2020EA001275, 2021.

Ionita, M., Scholz, P., and Chelcea, S.: Spatio-temporal variability of dryness/wetness in the Danube River Basin, Hydrological Processes, 29, 4483–4497, https://doi.org/10.1002/hyp.10514, 2015.

Isaak, D. J., Luce, C. H., Chandler, G. L., Horan, D. L., and Wollrab, S. P.: Principal components of thermal regimes in mountain river networks, Hydrology and Earth System Sciences, 22, 6225–6240, https://doi.org/10.5194/hess-22-6225-2018, 2018.

Jolliffe, I. T.: Principal Component Analysis, 2nd ed., New York, Springer, 2002.

Jolliffe, I. T.: Rotation of principal components: Some comments, Journal of Climatology, 7, 507-510, https://doi.org/10.1002/joc.3370070506, 1987.

Jolliffe, I. T.: Rotation of Ill-Defined Principal Components, Applied Statistics, 38, 139–147, https://doi.org/10.2307/2347688, 1989.

Jolliffe, I. T.: Rotation of principal components: choice of normalization constraints, Journal of Applied Statistics, 22, 29–35, https://doi.org/10.1080/757584395, 1995.

Jolliffe, I. T. and Cadima, J.: Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374, 20150202, https://doi.org/10.1098/rsta.2015.0202, 2016.

Kaiser, H. F.: The Varimax Criterion for Analytic Rotation in Factor Analysis, Psychometrika, 23, 187–200, https://doi.org/10.1007/BF02289233, 1958.

Kalayci, S. and Kahya, E.: Assessment of streamflow variability modes in Turkey: 1964–1994, 324, 163–177, https://doi.org/10.1016/j.jhydrol.2005.10.002, 2006.

Karl, T. R. and Koscielny, A. J.: Drought in the United States: 1895–1981, Journal of Climatology, 2, 313–329, https://doi.org/https://doi.org/10.1002/joc.3370020402, 1982.

Karl, T. R., Koscielny, A. J., and Diaz, H. F.: Potential Errors in the Application of Principal Component (Eigenvector) Analysis to Geophysical Data, Journal of Applied Meteorology, 21, 1183–1186, https://doi.org/10.1175/1520-0450(1982)021<1183:PEITAO>2.0.CO;2, 1982.

Korres, W., Koyama, C. N., Fiener, P., and Schneider, K.: Analysis of surface soil moisture patterns in agricultural landscapes using Empirical Orthogonal Functions, Hydrology and Earth System Sciences, 14, 751–764, https://doi.org/doi:10.5194/hess-14-751-2010, 2010.

Lay, D. C., Lay, S. R., and McDonald, J. J.: Linear Algebra and its applications, 5th ed., Pearson, 2016.

Legates, D. R.: The effect of domain shape on principal components analyses, International Journal of Climatology, 11, 135–146, https://doi.org/10.1002/joc.3370110203, 1991.

Kumar, M. and Duffy, C. J.: Detecting hydroclimatic change using spatio-temporal analysis of time series in Colorado River Basin, Journal of Hydrology, 374, 1–15, https://doi.org/10.1016/j.jhydrol.2009.03.039, 2009.

Larson, R. and Falvo, D.C.: Elementary Linear Algebra, 6th ed., Boston, Houghton Mifflin Harcourt Publishing Company, 2009.

Legates, D. R.: The effect of domain shape on principal components analyses: A reply, International Journal of Climatology, 13, 219–228, https://doi.org/10.1002/joc.3370130207, 1993.

Lehr, C. and Lischeid, G.: Efficient screening of groundwater head monitoring data for anthropogenic effects and measurement errors, Hydrology and Earth System Sciences, 24, 501–513, https://doi.org/10.5194/hess-24-501-2020, 2020.

Lehr, C.: R-scripts to (i) explore the domain dependence (DD) of spatial Principal Components and (ii) calculate DD reference patterns. Zenodo. https://doi.org/10.5281/zenodo.11213430, 2024.

Li, C., Mei, W., and Kamae, Y.: Variability and predictability of cold-season North Atlantic atmospheric river occurrence frequency in a set of high-resolution atmospheric simulations, Climate Dynamics, 58, 2485–2500, https://doi.org/10.1007/s00382-021-06017-y, 2022.

Lins, H. F.: Interannual streamflow variability in the United States based on principal components, Water Resources Research, 21, 691–701, https://doi.org/10.1029/WR021i005p00691, 1985a.

Lins, H. F.: Streamflow Variability in the United States: 1931–78, Journal of Climate and Applied Meteorology, 24, 463–471, https://doi.org/10.1175/1520-0450(1985)024<0463:SVITUS>2.0.CO;2, 1985b.

Lins, H. F.: Regional streamflow regimes and hydroclimatology of the United States, Water Resources Research, 33, 1655–1667, https://doi.org/10.1029/97WR00615, 1997.

Lischeid, G., Natkhin, M., Steidl, J., Dietrich, O., Dannowski, R., and Merz, C.: Assessing coupling between lakes and layered aquifers in a complex Pleistocene landscape based on water level dynamics, Advances in Water Resources, 33, 1331–1339, https://doi.org/10.1016/j.advwatres.2010.08.002, 2010.

Longuevergne, L., Florsch, N., and Elsass, P.: Extracting coherent regional information from local measurements with Karhunen-Loève transform: Case study of an alluvial aquifer (Rhine valley, France and Germany), Water Resources Research, 43, WO4430, https://doi.org/10.1029/2006WR005000, 2007.

Lorenzo-Seva, U. and ten Berge, J. M. F.: Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity, Methodology, 2, 57–64, https://doi.org/https://doi.org/10.1027/1614-2241.2.2.57, 2006.

Meyer, C.D.: Matrix analysis and applied linear algebra, SIAM, 2000.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B., and North, G. R.: Empirical Orthogonal Functions: The Medium is the Message, Journal of Climate, 22, 6501–6514, https://doi.org/10.1175/2009JCLI3062.1, 2009.

NCAR, National Center for Atmospheric Research Staff (Eds): The Climate Data Guide: Empirical Orthogonal Function (EOF) Analysis and Rotated EOF Analysis., https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis, last access: 18 April 2025, last modified 22 July 2013.

Nied, M., Hundecha, Y., and Merz, B.: Flood-initiating catchment conditions: a spatio-temporal analysis of large-scale soil moisture patterns in the Elbe River basin, 17, 1401–1414, https://doi.org/10.5194/hess-17-1401-2013, 2013.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J.: Sampling Errors in the Estimation of Empirical Orthogonal Functions, Monthly Weather Review, American Meteorological Society, 110, 699–706, https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2, 1982.

Quadrelli, R., Bretherton, C. S., and Wallace, J. M.: On Sampling Errors in Empirical Orthogonal Functions, Journal of Climate, 18, 3704–3710, https://doi.org/10.1175/JCLI3500.1, 2005.

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., and Gratzki, A.: A Central European precipitation climatology? Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS), Meteorologische Zeitschrift, 22, 235–256, https://doi.org/10.1127/0941-2948/2013/0436, 2013.

R Core Team: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.R-project.org/, last access: 12 December 2019.

Richman, M. B.: Rotation of principal components, Journal of Climatology, 6, 293–335, https://doi.org/10.1002/joc.3370060305, 1986.

Richman, M. L. B.: Rotation of principal components: A reply, Journal of Climatology, 7, 511–520, https://doi.org/10.1002/joc.3370070507, 1987.

Richman, M. B.: Comments on: 'The effect of domain shape on principal components analyses', International Journal of Climatology, 13, 203–218, https://doi.org/10.1002/joc.3370130206, 1993.

Roundy, P. E.: On the Interpretation of EOF Analysis of ENSO, Atmospheric Kelvin Waves, and the MJO, Journal of Climate, 28, 1148–1165, https://doi.org/10.1175/jcli-d-14-00398.1, 2015.

Santos, J. F., Pulido-Calvo, I., and Portela, M. M.: Spatial and temporal variability of droughts in Portugal, Water Resources Research, 46, https://doi.org/10.1029/2009wr008071, 2010.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K.: Analysis, Simulation and Prediction of Multivariate Random Fields with PackageRandomFields, Journal of Statistical Software, 63, 1–25, https://doi.org/10.18637/jss.v063.i08, 2015.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Auel, J., Menck, P. J., Gross, S., Ober, U., Ribeiro, P., Ripley, B. D., Singleton, R., Pfaff, B.: R Core Team: RandomFields: Simulation and Analysis of Random Fields. R package version 3.3.8, https://cran.r-project.org/package=RandomFields, 2020, last access: 17 March 2021.

Scholz, H., Lischeid, G., Ribbe, L., Hernandez Ochoa, I., and Grahmann, K.: Differentiating between crop and soil effects on soil moisture dynamics, Hydrology and Earth System Sciences, 28, 2401–2419, https://doi.org/10.5194/hess-28-2401-2024, 2024.

Smirnov, N.: Asynchronous long-period streamflow fluctuations in the European U.S.S.R, Soviet Hydrology: Selected Papers, American Geophysical Union, 46–49, 1972.

Smirnov, N. P.: Spatial patterns of long-period streamflow fluctuations in the European U.S.S.R, Soviet Hydrology: Selected Papers, American Geophysical Union, 112–122, 1973.

von Storch, H. and Zwiers, F. W.: Statistical Analysis in Climate Research, netLibrary ed., Cambridge University Press, 2003.

Thomas, B., Lischeid, G., Steidl, J., and Dannowski, R.: Regional catchment classification with respect to low flow risk in a Pleistocene landscape, Journal of Hydrology, 475, 392–402, https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2012.10.020, 2012.

Upton, K. A. and Jackson, C. R.: Simulation of the spatio-temporal extent of groundwater flooding using statistical methods of hydrograph classification and lumped parameter models, Hydrological Processes, 25, 1949–1963, https://doi.org/10.1002/hyp.7951, 2011.

Vejmelka, M., Pokorná, L., Hlinka, J., Hartman, D., Jajcay, N., and Paluš, M.: Non-random correlation structures and dimensionality reduction in multivariate climate data, Climate Dynamics, 44, 2663–2682, https://doi.org/10.1007/s00382-014-2244-z, 2014.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd ed., Elsevier, 2006.

Winter, T. C., Mallory, S. E., Allen, T. R., and Rosenberry, D. O.: The Use of Principal Component Analysis for Interpreting Ground Water Hydrographs, Groundwater, 38, 234–246, https://doi.org/10.1111/j.1745-6584.2000.tb00335.x, 2000.