Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR025728

On the Optimal Spatial Design for Groundwater Level Monitoring Networks

M. Ohmer[1], T. Liesch[1], and N. Goldscheider[1]

[1]Institute of Applied Geosciences, Division of Hydrogeology, Karlsruhe Institute of Technology, Karlsruhe, Germany

Key Points:
- A new low-discrepancy method (R_2 method) for an extensible groundwater level monitoring network design is proposed
- We define a range of groundwater level monitoring network densities with an optimized information/cost ratio
- We show that global cross-validation error parameters are not suitable for the comparative assessment of different sampling designs

Supporting Information:
- Supporting Information S1
- Figure S1
- Table S1

Correspondence to:
M. Ohmer,
marc.ohmer@kit.edu

Citation:
Ohmer, M., Liesch, T., & Goldscheider, N. (2019). On the Optimal Spatial Design for Groundwater Level Monitoring Networks. *Water Resources Research*, *55*. https://doi.org/10.1029/2019WR025728

Received 6 JUN 2019
Accepted 30 OCT 2019
Accepted article online 16 NOV 2019

Abstract Effective groundwater monitoring networks are important, as systematic data collected at observation wells provide a crucial understanding of the dynamics of hydrogeological systems as well as the basis for many other applications. This study investigates the influence of six groundwater level monitoring network (GLMN) sampling designs (random, grid, spatial coverage, and geostatistical) with varying densities on the accuracy of spatially interpolated groundwater surfaces. To obtain spatially continuous prediction errors (in contrast to point cross-validation errors), we used nine potentiometric groundwater surfaces from three regional MODFLOW groundwater flow models with different resolutions as a priori references. To assess the suitability of frequently-used cross-validation error statistics (MAE, RMSE, RMSSE, ASE, and NSE), we compared them with the actual prediction errors (APE). Additionally, we defined upper and lower thresholds for an appropriate spatial density of monitoring wells. Below the lower threshold, the observation density appears insufficient, and additional wells lead to a significant improvement of the results. Above the upper threshold, additional wells lead to only minor and inefficient improvements. According to the APE, systematic sampling lead to the best results but is often not suited for GLMN due to its nonprogressive characteristic. Geostatistical and spatial coverage sampling are considerable alternatives, which are in contrast progressive and allow evenly spaced and, in the case of spatial coverage sampling, yet reproducible coverage with accurate results. We found that the global cross-validation error statistics are not suitable to compare the performance of different sampling designs, although they allow rough conclusions about the quality of the GLMN.

1. Introduction

Groundwater is an important, yet spatially extensive, concealed, and inaccessible resource. Therefore, an effective groundwater monitoring network (GMN) is important, as systematic data collected at observation wells provide a crucial understanding of the dynamics and quality of the hydrogeological system. A GMN is defined by a spatial arrangement of monitoring sites and a temporal sampling frequency (Loaiciga et al., 1992). Economic considerations most strongly influence the number and location of monitoring wells. Designing an optimal GMN is, therefore, a task of balancing prediction accuracy with cost minimization (Krivoruchko, 2011). Since a high spatial resolution is usually associated with disproportionate costs, often only domains of high water management importance are adequately monitored. The design, that is, the selection of the location and number of the monitoring wells, is a vital part of any study involving modeling and prediction based on spatial data. A groundwater model can only be as good as the model input data available. Therefore, poorly distributed monitoring wells can lead to wrong assumptions or to a bias of the regional image. Unsuitable interpolation methods can yield drastic overestimates or underestimates of the groundwater level in areas with a low monitoring density, as the parts with a high monitoring density are disproportionately weighted (Ohmer et al., 2017). GW-quality observations are subject to the same problem. In this study, however, we focus on the regional groundwater level as a monitoring parameter.

Varieties of studies dealing with the optimization of GMN have been published in the last 20 years. The literature focuses on the optimization of GMN design to observe the groundwater quality (GQMN), and the number of studies dealing with groundwater level (GLMN) is limited. The majority of approaches for GLMN design optimization are based on geostatistical analysis and therefore on minimizing uncertainty in parameter estimation. Several studies apply undifferentiated kriging (Prakash & Singh, 2000; Theodossiou & Latinopoulos, 2006), ordinary kriging (OK; Nunes et al., 2004; Yang et al., 2008), universal kriging (UK; Kambhamettu et al., 2011; Kumar et al., 2005; Olea, 1984), OK and UK (Ahmadi & Sedghamiz, 2007), OK and co-kriging (CG; Ma et al., 1999), or indicator kriging (IK; Cameron & Hunter,

2002) to interpolate the groundwater surface and use either the mean or maximum kriging variance to determine where additional observation wells should be built and/or identify well redundancy. Some recent works use the kriging variance as a part of a multicriteria decision-making analysis (MCA). Chandan and Yashwant (2017) considered multiple parameters in addition to the kriging variance such as groundwater level fluctuation, land use, hydrology, and recharge lineament density, to optimize an existing GLMN. In Uddameri and Andruss (2014), the kriging variance was linked to a monitoring priority index (MPI) calculated from a weighted average of several criteria (GW-variability, recharge, surface/GW interaction, and GW fluxes across district boundaries). A similar approach can be found in Zhou et al. (2013) and Wang (2011). Esquivel et al. (2015) used for their MCA a weighted linear combination that takes GW fluctuations, rates of decline, observation density, hydraulic gradients, mountains, and water bodies, for example, into account. Additional studies use information theory (entropy estimation) to evaluate the spatial location and temporal measuring frequency of monitoring wells (Alfonso et al., 2014; Leach et al., 2016; Masoumi & Kerachian, 2008; Mogheir et al., 2006; Mogheir et al., 2009). Further studies apply kriging-based genetic algorithms (Babbar-Sebens & Minsker, 2010; Dhar & Patil, 2012; Kollat & Reed, 2006; Luo et al., 2016; Reed et al., 2007; Yeh et al., 2006), algorithms based on Kalman filter with space-time covariance matrix as input (Júnez-Ferreira & Herrera, 2013; Wu, 2004; Zhang et al., 2005), and artificial neural networks (ANNs; Giustolisi & Simeone, 2010).

Some authors also take the temporal component of monitoring networks into account (i.e., frequency of measurements based on groundwater fluctuations; e.g., Ahmadi & Sedghamiz, 2007; Cameron & Hunter, 2002; Chandan & Yashwant, 2017; Kambhamettu et al., 2011; Nunes et al., 2004; Theodossiou & Latinopoulos, 2006). However, the results are always a compromise between the spatial and temporal component. Moreover, the influence of the individual components is difficult to quantify. Since the temporal component (i.e., measuring interval) can be easily varied, whereas the repositioning of sampling points represents a disproportionately greater effort, a focus solely on the spatial arrangement seems justified. Therefore, in our study we focus on the spatial component and assume a steady-state groundwater surface to find the optimal spatial arrangement of sampling points regardless of the temporal component.

What most of the studies mentioned here have in common is that a large amount of auxiliary data for the respective study areas have been incorporated into the optimizations. Therefore, the results cannot be easily transferred to other areas where these data may not be available. One aim of this study is to find out if there is a universally applicable design approach that can achieve the best possible results without a priori knowledge of the hydrogeological situation. For this reason, we compared six design strategies that were as generally applicable as possible and analyzed their results in nine areas of investigation. We assumed "real" groundwater surfaces to be known as they are taken from large-scale numerical groundwater models for this simulation experiment. The idea behind this approach is to compare the interpolated surfaces resulting from the respective GLMN design to the "real" surface and thereby compute the "real" error in addition to the cross-validation (CV) error.

In detail, the research objectives are to answer the following questions:

1. Is there an extensible and transferable GLMN design that allows reliable spatial estimates of groundwater level with a minimum number of monitoring wells?
2. What quality differences result from the use of different GLMN design approaches?
3. At what observation well density a reasonable information/cost ratio emerges?
4. Which is the most suitable CV error statistic (MAE, RMSE, RMSSE, ASE, or NSE) to evaluate the quality of interpolated groundwater surfaces?

To answer these questions, we applied six different design strategies on nine different groundwater surfaces, starting with initial 10 observation wells each. Based on the groundwater levels of the observation wells, a groundwater surface was interpolated and the deviation from the real groundwater surface calculated, along with global CV error statistics. The monitoring networks were then gradually densified up to 500 observations wells, the interpolation and error calculation being repeated after each step. The design strategies, which contain random components, were repeated 10 times in order to include the influences caused by them.

2. Methods

2.1. Interpolation Method

In a previous study, we examined and compared nine different interpolation methods (inverse distance weighting, radial basis function, simple, ordinary, universal, empirical Bayesian and CG, and local and global polynomial interpolation) to find the most suitable method to interpolate a continuous groundwater surface from observed groundwater levels (Ohmer et al., 2017). The best results, based on global CV error statistics, were achieved with co-OK. In this type of kriging, additional correlated secondary variables (e.g., DEM, springs, and wetlands) are used to improve the prediction. Since secondary data are generally not sufficiently available everywhere, we decided not to consider this method here and instead use the second best method, OK, which is one of the most frequently used geostatistical estimators (Siska et al., 2005; Wackernagel, 1995, *i.a.*).

Geostatistics is based on the work of Krige (1951) and was further developed by Matheron (1963) with his theory of regionalized variables. Kriging is a generic name for a group (e.g., simple kriging, OK, and UK) of generalized least squares regression algorithms (Li, 2008). Before the prediction, the spatial correlation of the regionalized data is assessed by a semivariogram analysis. The semivariance γ of Z between 2 points $x_i x_o$ separated by distance h is defined as

$$\gamma(x_i x_o) = \gamma(h) = \frac{1}{2} var\,[Z(x_i) - Z(x_0)]. \tag{1}$$

The empirical semivariogram is a graphical representation of the semivariance ($\gamma(h)$ vs. h) and represents the spatial autocorrelation of the data points. It quantifies the assumption that nearby data points tend to be more similar than more remote points (First Law of Geography, according to Tobler, 1970). This empirical semivariogram is used as the first estimate of the theoretical semivariogram that is needed for the spatial interpolation. Important features of the semivariogram are the nugget, the range, and the sill/partial sill. The nugget effect is a positive value of $\gamma(h)$ at h close to 0. It allows for the variogram to assume a nonzero value for two observations having a separation distance that is less than the minimum bin size, and accounts for a sum of measurement error and microscale irregularities. A method that produces an estimate equal to the observed value at the sample points is called exact; all others are called inexact. The sill is the semivariance value at which the semivariogram levels off for stationary data sets (Bohling, 2005). Partial sill results from the difference between sill and nugget. The range is a value of distance at which the sill is reached. Points further away than the range are regarded as spatially independent (Li, 2008). OK is robust and straightforward and therefore probably the most widely used kriging technique (Heuvelink & Pebesma, 2002). Each of the different kriging methods (e.g., simple kriging and OK) is based on the following basic equation:

$$\hat{z}(x_0) - \mu(x_0) = \sum_{i=1}^{n} \lambda_i [Z(x_i) - \mu(x_i)], \tag{2}$$

where $\hat{z}$ is the estimated value at a point of interest x_0, n is the total number of measured groundwater levels, and $Z(x_i)$ is the measured groundwater level at well x_i. λ_i are the kriging weights derived from a covariance function or semivariogram; μ is in the case of OK the Lagrange multiplier constant that has to be estimated and is considered to be constant over the area to be interpolated (Li & Heap, 2008):

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i^{OK}(x_o) Z(x_i) \; with \; \sum_{i=1}^{n} \lambda_i^{OK}(x_o) = 1. \tag{3}$$

We used an omnidirectional Gaussian semivariogram model, which is flexible, and a good candidate for a default model (Krivoruchko, 2011). With its parabolic behavior at the origin, it represents very smoothly varying properties (Bohling, 2005). The associated parameters partial sill, range, search neighborhood, and specific search distance were optimized by using automated CV diagnostics minimizing the RMSE for each individual case. To ensure the stability of the resulting kriging matrices, a systematically small constant nugget of 0.05 m was used (Johnston, 2004; Yarus & Chambers, 1994).

It should be noted that the use of a single variogram model (Gaussian) might not be the optimal way to quantify spatial correlation, especially for nonstationary data. However, it is a necessary simplification owed to

the automation process, which still allows comparability of the spatial design methods, while the best possible interpolation result is not the focus of this study.

2.2. Data and Data Processing

The spatial and temporal variability of the groundwater surface is generally unknown except for wells, springs, wetlands, and interacting surface waters. Therefore, the accuracy assessment of an interpolated/predictive groundwater surfaces can only take place at these measured locations using CV and error statistics (e.g., ME, MAE, RMSE, NSE, and RMSSE). The expected level of fit of these results is therefore primarily dependent on the number and distribution of the monitoring locations.

To quantitatively determine and compare the effects of the different GLMN designs on the accuracy of the predicted groundwater surfaces, we used nine potentiometric groundwater surfaces, extracted from simulation results of three regional MODFLOW groundwater flow models as an a priori reference. The model data are publicly accessible from the USGS (DeSimone et al., 2002; Parsen et al., 2016; Sepulveda & Painter, 2017). The idea behind this approach was to compare the interpolated surfaces to the "real" surface and thereby compute the "real" error in addition to the CV error. This has been done with completely artificial surfaces to compare different network designs (Aguilar et al., 2005; Heuvelink & Pebesma, 2002; Romero et al., 2005; Wild, 2009), but artificial surfaces may not have the same properties as typical groundwater surfaces in terms of variability, roughness, gradients, and so on. To use surfaces computed by numerical groundwater models, which incorporate the hydraulic properties of the aquifer, are based on physical processes of groundwater flow, and therefore produce not "real" but at least "realistic" surfaces, seems to be the best compromise. The resulting groundwater surfaces each consist of 100 × 100 pixels with pixel sizes of 100 m × 100 m, 200 m × 200 m, or 500 m × 500 m. The pixel size corresponds approximately to the element size of the groundwater models.

The different resolutions of the surfaces were chosen to assess if and how the resolution affects the results. The resolution can, for example, influence elevation and slope values (Chunmei et al., 2013), as small-scale variabilities below the pixel size are eliminated. However, this does not allow comparing the observation density and resulting errors with each other directly. A detailed overview of essential parameters of the surfaces is given in Table S1 in the supporting information, groundwater contour maps of the surfaces are shown in Figure 1.

For the simulation experiment, we used the following automated workflow for all surfaces:

1. Ten initial monitoring points were distributed randomly on every surface from a random number generator (exceptions are the systematic sampling and low discrepancy sampling methods; see section 2.3).
2. Based on the groundwater level of the available observation points, an empirical variogram was computed, and an optimized (by CV) Gaussian semivariogram model was fitted to the data.
3. Based on the semivariogram, the GWS was interpolated with OK.
4. The prediction error (the difference between the real and the predicted surface), as well as different CV error statistics, was calculated.
5. The location for an additional monitoring point was computed, depending on the used design method (see section 2.3),
6. Steps (ii) to (iv) were repeated until the monitoring network included 500 points. Steps (i) to (vi) were repeated 10 times for the methods that include random components, and the results were averaged to consider errors caused by random initialization or random addition of point locations.

For OK stationarity (a constant mean and variance of the values across the area) generally must be assumed, which is not the case for all tested surfaces (Figure 1). Nevertheless, we have chosen a uniform procedure in order to ensure comparability of the sampling designs and to exclude the influence of different interpolation methods.

2.3. Sampling Designs

When planning a new or extending an existing GLMN, one of two fundamentally different strategies must be chosen. One is the *design-based sampling* approach, which is *based* on *classical sampling theory*; the other is the model-*based approach*, *based* on geostatistics. The main difference between the two is how they deal with the randomness they use to give the inference a stochastic structure (Särndal, 1978). The additional

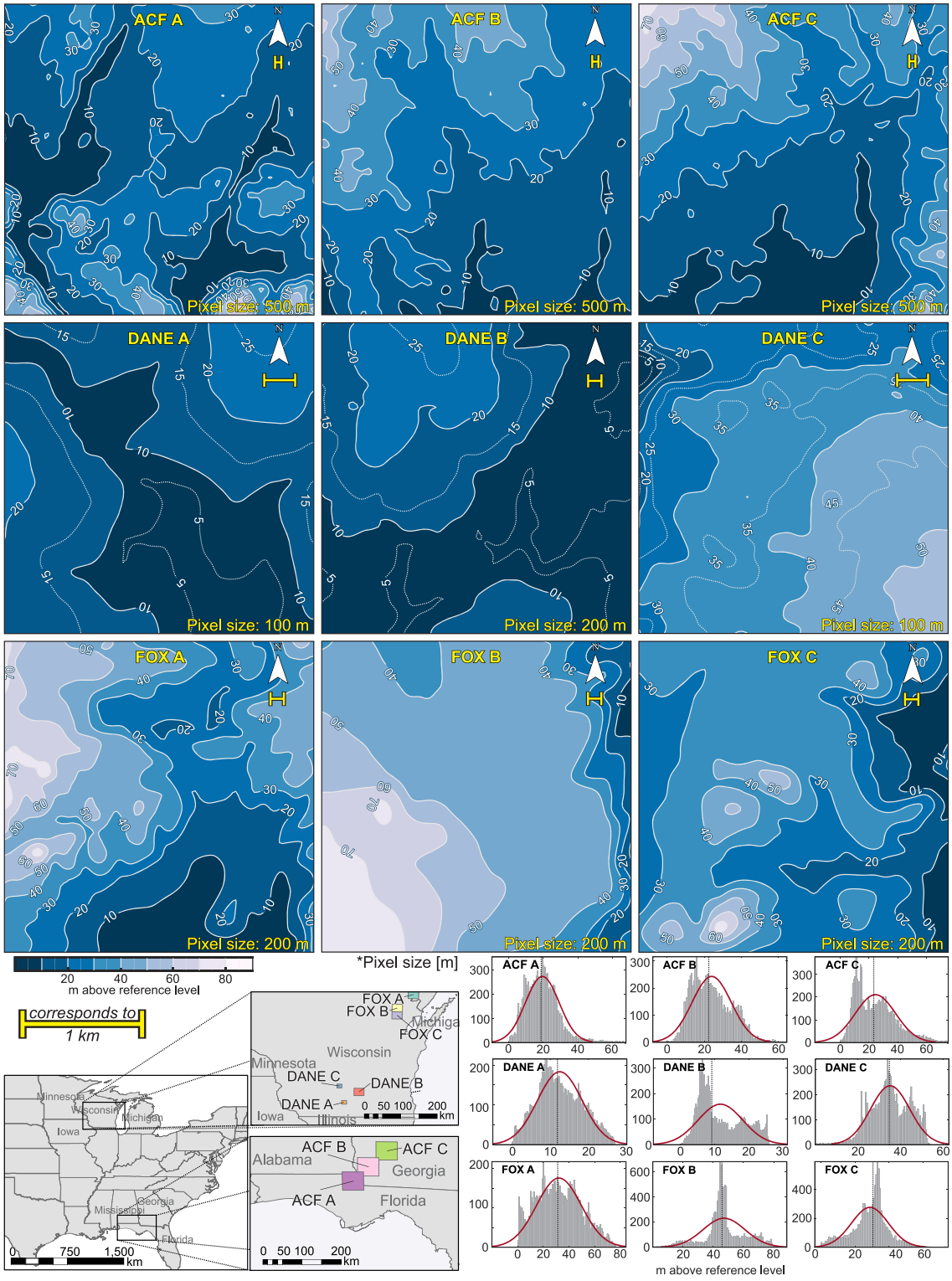


Figure 1. Groundwater contour maps and histograms extracted from three MODFLOW models as a priori reference to evaluate the investigated observation network designs qualitatively and quantitatively (DeSimone et al., 2002; Parsen et al., 2016; Sepulveda & Painter, 2017).

monitoring wells (or, in general, samples) in a design-based observation network are selected in such a way that each location within the study area has the same probability of being chosen. Another term for design-based sampling is, therefore, probability sampling. These probabilities provide the foundation for statistical

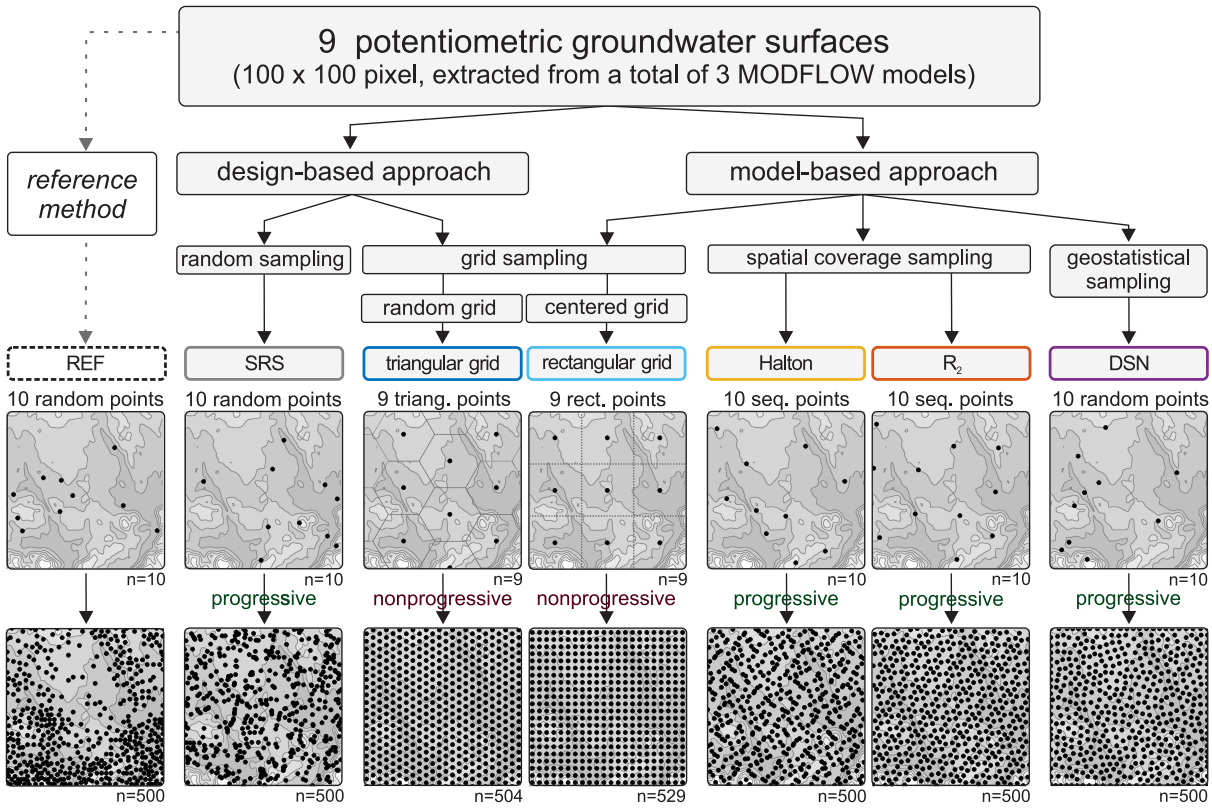


Figure 2. Schematic overview of the observation network designs compared in this study.

inference from the observations (Gruijter et al., 2006). In a model-based network, a theoretical construction (model) is used to deal with the differential probabilities of the potential observation points. The model is built upon information on prior knowledge and assumptions (Geuna, 2000). It contains the prescription for the statistical inference. A detailed description of sampling theory and the contrast between the two strategies can be found in Särndal (1978), Hansen et al. (1983), Brus and de Gruijter (1997) and Brus (2010). Figure 2 shows an overview of the observation network designs compared in this study.

2.3.1. Design-Based Approach: Spatial Statistical Sampling and Probability Sampling

Four standard types of statistical sampling methods are generally used in "classical" statistical surveys (Kish, 1995). These methods are *simple random sampling* (*SRS*), *random grid sampling* (random systematic sampling), *stratified sampling*, and *cluster sampling*. For this study, *SRS* and random grid sampling were chosen. Stratified sampling and cluster sampling are usually used, when a heterogeneous distribution of values can be broken down in parts that are internally homogenous or when values tend to cluster together (Gilbert, 1987). Neither is the case for groundwater elevation.

2.3.1.1. Simple Random Sampling

SRS is the simplest and most frequently used sampling design approach in survey sampling, since it is assumed that n wells are located randomly from N potential sites with an "equal probability of selection" (EPS) throughout a domain. The advantages are that it is simple to use and free from bias and prejudice. Disadvantages are poor spatial coverage and the possible occurrence of clustering, redundancy, and regions that contain no observation points. The additional observation points with *SRS* were selected in this study using the ESRI ArcGIS tool *GenerateRandomPoints*. This tool creates a specified number of random points in a defined extent.

2.3.1.2. Random Grid Sampling

In *random grid sampling* (or random systematic sampling), samples are taken at regularly spaced intervals over space with the first point m chosen randomly. The approach is classified as probability sampling since, with the first point selected randomly, each location within the study area initially has the same probability of being selected. Examples of a systematic grid are *rectangular*, *triangular*, *hexagonal*, or *radial grids*. If the

starting point is not chosen randomly, the approach is classified as model based (see section 2.3.2), as, for example, in *regular systematic sampling* and *centric systematic sampling* (Delmelle, 2014). The main advantage of a systematic sampling over SRS is that it can be more conducive to covering more extensive areas through a maximized sampling coverage while clustering and redundancy are prevented. Moreover, it allows to add a degree of system into the process of random selection (Fischer & Nijkamp, 2014). Gridded sampling designs are particularly suitable for large investigation areas, which should be covered with a limited number of sampling points (Gruijter et al., 2006). The disadvantage of this method is that the smallest separation distance is fixed. Since the kriging variance is described as a function of the separation distance, this can lead to unnecessarily large nugget effects for the model semivariogram (Baxter, 2016). Furthermore, the approach is not progressive, which means that the number of overall observation points has to be known beforehand, and it is not possible to progressively add more points without breaking the order. Though therefore systematic sampling methods are not suitable to construct extensible observation networks, we added two methods for comparison, as they are often referred to as the most efficient design for survey sampling (Birch et al., 2007; Olea, 1984).

The systematic random approach used in this study is a *triangular grid* based on regular hexagonal polygons. The points are placed in the center as well as at the corners of a hexagon (shared with three other hexagons). The hexagons have a width $w = \sqrt{3}{\cdot}sidelength$ and a height $h = 2 \cdot sidelength$. Therefore, the lateral distance d_{hz} between adjacent hexagon center points is w while the longitudinal distance is $h{\cdot}3/4$. In contrast to other methods, here no additional points can be added progressively without breaking the grid symmetry. Instead, the existing points were replaced in the next iteration step by as many points as were necessary to maintain a regular grid with the next-larger number of points.

2.3.2. Model-Based Approach

2.3.2.1. Centered Grid Sampling

Though very similar to random grid sampling, the centered grid sampling is referred to as a model-based approach, since in contrast to random grid sampling, the starting point is not chosen randomly, but purposefully, so that the area of investigation is well covered, especially when the overall number of points is low. For the *rectangular grid* used in this study, the investigated area was divided into $n{\cdot}n$ *square* intervals, and the observation points were set in their center (*centric systematic sampling*). The shortest distance d between the observation points equals the side length of the square divided by the square root of the sample size, $d = \sqrt{A/n}$, where A is the area of the square. As with all regular grids, this method is not progressive. An existing network cannot be extended under the applicable laws.

2.3.2.2. Spatial Coverage Sampling

Spatial coverage sampling is a technique that optimizes an objective function of the distance between the observation points (Brus et al., 2006). We have chosen two low-discrepancy sequences as objective functions. In *quasi-random* or *low-discrepancy sampling*, the position of sampling points is based on low-discrepancy sequences (also called quasi-random or subrandom sequences). These sequences represent numbers that are better equidistributed than pseudo-random numbers (Dalal et al., 2008). To construct higher-dimension low discrepancy, as in the case of two-dimensional sampling design, several one-dimensional sequences are combined in a component-wise manner, that is, that the x and y coordinates of a two-dimensional area are constructed by pairing consecutive numbers of two different low discrepancy series in an $[0,1] \times [0,1]$ space and then adjusted to the actual spatial extent of the area to be sampled. In a two-dimensional context, discrepancy refers to the density of points on an area or sampling space. A high discrepancy means that there are large areas of empty space or regions with a disproportionally high point density (as it may be the case in a random distribution). Therefore, *SRS* can lead to a high discrepancy while *systematic sampling* has the lowest discrepancy. Fully deterministic low-discrepancy sequences were developed to optimize Monte Carlo simulations because they fulfill requirements as if they were genuinely random numbers. At the same time, higher accuracy and faster convergence can be achieved with fewer samples compared to pseudo-random numbers, which reduces the computational costs. Low-discrepancy sampling methods thus constitute a good compromise of being progressive (like SRS) and having a low discrepancy (like systematic sampling). Therefore, they are frequently used in sampling problems. Furthermore, they allow for a better distribution of sample separation distances than gridded sampling schemes, while minimizing sampling bias and clustering. To our knowledge, low-discrepancy sequences have not yet been applied to the development of groundwater level monitoring design before.

A categorization of the different types of low-discrepancy sequences is mostly done by the method of constructing their bias (hyper)parameter. These are either prime numbers (*Van der Corput*, *Halton* and *Faure sequence*), polynomials (*Sobol* and *Niederreiter sequence*), or irrational fractions (*Kronecker* and *R-sequence*). In this study, the *Halton sequence* and the *R-sequence* were investigated, as they are more suitable for low-discrepancy in two dimensions than other low-discrepancy series (Roberts, 2018).

The *Halton sequence* (Halton, 1960) is a generalization of the van der Corput sequence (van der Corput, 1935) to higher dimensions. It uses arbitrary coprime numbers as a base for each dimension. The most frequent selection for two dimensions, due to its apparent simplicity and sensibility, is to select the first primes, which is referred to as the (2,3)-*Halton sequence*. To generate the sequence for 2, the interval [0,1] is divided in half, then in fourths, eighths, and so forth, which generates ½, ¼, ¾, 1/8, 5/8, 3/8, and so forth. Equivalently, the *Halton* sequence for 3 is generated by dividing the [0,1] interval in thirds, ninths, twenty seventh, and so forth, giving 1/3, 2/3, 1/9, 4/9, 7/9, 2/9, and so forth. The coordinates for the sampling points are constructed by placing the *x* coordinates according to the *2-Halton sequence* and *y* coordinates according to the *3-Halton sequence*, adjusting the numbers of the sequence to the actual spatial coordinate extent of the area.

The *Halton* sequence constitutes a good source for a low discrepancy in two dimensions since the selection of small coprime bases ensures a minimal correlation between dimensions (Worley, 2016) and is therefore regularly used in ecological sampling (Brown et al., 2015; Kermorvant et al., 2019).

Recently, a new low discrepancy quasi-random sequence that offers a substantial improvement over current state-of-the-art sequences has been proposed by Roberts (2018). The new additive recurrence sequence (R-sequence) is a recurrence method based on irrational numbers (generally called Kronecker sequences), which uses the golden ratio as a basis. For the two-dimensional case (R_2-*sequence*), it produces more evenly spaced points than any of the other known methods. The generalized version of the golden ratio Φd is defined as the unique positive root $xd + 1 = x + 1$. That is, for $d = 2$, $\Phi 2 = 1.3247\ldots$. This value was conjectured to most likely be the optimal value for a related two-dimensional problem (Hensley & Su, 2004). In two dimensions the x and i coordinates of the nth term ($n = 1, 2, 3, \ldots$) are defined as

$$x_n = \left(0.5 + \frac{1}{\Phi_2}\cdot n\right) \text{ and } y_n = \left(0.5 + \frac{1}{\Phi_2{}^2}\cdot n\right). \tag{4}$$

2.3.2.3. Geostatistical Sampling

Model-based sampling techniques, more specifically geostatistical sampling when the postulated model is a geostatistical model, have been applied in several studies in recent years for the assessment and location of additional monitoring wells in a GLMN (see section 1).

In geostatistical sampling, a geostatistical model, the model variogram, is used to identify locations for additional sampling points (Gruijter et al., 2006). These are selected based on the predefined selection criterion, usually by minimizing the mean or the maximum kriging variance (Olea, 1984).

$$\sigma^2_{OK} = Var\,(Z) - \sum_{\alpha=1}^{n}\lambda_\alpha C\,(h_{0\alpha}) \tag{5}$$

The kriging variance depends on the covariance model and the data configuration but is independent of the data values. For a given covariance model two identical sample location distributions would yield the same kriging variance independently no matter what the data were (Goovaerts, 1997). Heuvelink and Pebesma (2002) examined the validity of kriging variance by numerical analysis of two-dimensional Gaussian distributed realizations and by mathematical-statistical description. They show that the prediction error variance is independent of the data values and therefore the kriging variance is still a correct assessment of the local uncertainty even if in parts of the area the variations are larger or smaller than elsewhere. However, these findings do not apply to the non-Gaussian case.

We used the maximum prediction standard error (square root of the kriging variance) as the selection criterion for placing additional sampling points. This approach is implemented in the ArcGIS Tool "Densify Sampling Network" and is therefore referred to as *DSN* (Johnston, 2004).

2.3.3. Reference

We introduce a reference method (*REF*) to evaluate the quality of the tested approaches. After each interpolation, the predicted surface is compared with the "real," surface, and an additional well is set at the point

with the largest difference between the two surfaces. The idea behind this approach is that the resulting arrangement is the best possible design for each respective surface. Hence, we assume that it theoretically represents the lower limit for the smallest prediction error achievable with a given number of observation wells.

2.3.4. Progressive Versus Nonprogressive Designs

Progressive (sequential) sampling design creates an observation network by optimally adding one or several new points step by step, whereas in nonprogressive (simultaneous) sampling design, all points have to be added at once.

While some methods are either progressive or nonprogressive, some others can be used in both ways (though they are usually referred to as progressive only). The latter is the case for *SRS* and the spatial coverage sampling strategies (R_2 and *Halton*), where the results are the same whether one, several, or all points are added at a time. The *DSN* method constitutes an exception. It can be used progressive or nonprogressive, but the results may differ. *DSN* places additional points at the highest prediction standard error, which depends on the locations of the existing observations points as well as the semivariogram model. The semivariogram is recomputed in each densifying process and therefore leads to different results, depending if a number of points is added one by one (sequential) or all at one time (simultaneously). In practice, the differences in the results are often small, at least when a certain number of observations is reached, since the addition of new points only leads to minor changes in the modeled semivariogram.

The grid sampling methods are nonprogressive, in the sense that only a certain number of points can be added at a time without breaking the grid symmetry (and therefore are usually referred to as nonprogressive only).

Whether a progressive or nonprogressive method is preferred may depend from case to case, though in GLMN design, progressive methods have the advantage to allow for an extension of the network at a later stage.

2.4. Validation Methods

The *absolute prediction error* (*APE*) is the absolute mean difference between the real GWS and the predicted (interpolated) GWS, divided by the area of the surface. For the methods with a random component (*REF*, *SRS*, *triangular random grid*, and *DSN*), a mean value was calculated from 10 runs.

$$APE = \frac{\sum_{i=1}^{n} \left| \hat{Z}_{(x)} - Z_{(x)} \right|}{\text{Area}} \tag{6}$$

The *standardized absolute prediction error* (*SAPE*) is the APE divided by the maximum value of the APEs of all methods per surface (except *REF*). The SAPE makes it easier to compare the error propagations of the individual surfaces with each other.

$$SAPE = \frac{APE}{\max(APE)_{all\ methods}} \tag{7}$$

The *mean standardized absolute prediction error* (*MSAPE*) is the mean SAPE of the individual design approaches for all nine surfaces.

Since the real value for an actual GWS at the time of the prediction is generally unknown, the real prediction error is also unknown. Validation of the interpolation can therefore only be performed at the observation points, using CV and error statistics to assess the accuracy of the interpolation. Based on the results of the CV, the following error measures were used to compare the accuracy of the different interpolation methods, where n is the number of observations, m_i is the measured value, and p_i is the predicted CV value at the position i:

The *mean absolute error* (*MAE*) is the arithmetic mean of the absolute error values. It indicates the magnitude of the error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - m_i|. \quad (8)$$

The *root mean square error* (*RMSE*) represents the root of the averaged square error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - m_i)^2}. \quad (9)$$

The *average standard error* (*ASE*) is the average of the prediction standard errors:

$$ASE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_i - \left(\sum_{i=1}^{n}p_i\right)/n\right)^2}. \quad (10)$$

The *root mean square standardized error* (*RMSSE*) with p_{si} as the standardized predicted value and m_{si}:

$$RMSSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{si} - m_{si})^2}. \quad (11)$$

The Nash-Sutcliffe model efficiency coefficient (NSE; Nash & Sutcliffe, 1970) with $\overline{m_i}$ as the mean of the measured values is used to quantify how well a model simulation can predict the outcome variable. The NSE ranges from minus infinity to 1 (perfect fit):

$$NSE = \frac{\sum_{i=1}^{n}(p_i - m_i)^2}{\sum_{i=1}^{n}(m_i - \overline{m_i})^2}. \quad (12)$$

In a detailed review, Li and Heap (2008) have compiled a comprehensive assessment of error statistics. They conclude that MAE and RMSE are similar measures that give an estimate of the average error but do not provide information about the relative size of the average difference nor the nature of the difference. In contrast to MAE, RMSE is very sensitive to outliers (Hernandez-Stefanoni & Ponce-Hernandez, 2006; Ikechukwu et al., 2017; Vicente-Serrano et al., 2003; Willmott, 1982). Nonetheless, both are among the best measures of model performance (Willmott, 1982). The following criteria for using error measurements are proposed to assess the performance of spatial interpolation. MAE, RMSE, and ASE should be as small as possible. ASE and RMSE should be nearly identical, and RMSSE should be close to 1, indicating the estimated prediction uncertainty is consistent. ASE > RMSE or RMSSE > 1 implies an overestimation of the variability of the predicted values; ASE < RMSE or RMSSE < 1 implies an underestimation (Hu et al., 2004).

3. Results and Discussion

We divided the presentation of the results into three sections. In section 3.1, the spatial distributions of the added monitoring wells from the tested approaches are shown and compared on example (ACF A). Since a priori known groundwater surfaces are used, not only the CV errors but also the "real" prediction errors based on the GLMN can be computed. Therefore, in section 3.2, a comparison of the tested design approaches based on these real prediction errors (APE) is drawn. Section 3.3 finally contains a comparison based on the CV results.

3.1. Resulting Spatial Distribution

The GLMN designs resulting from each approach are shown for the example of ACF A in Figure 3. The Reference method (*REF*) results in a network design with concentrated observation density and clustering in the more variable southern part of the area, while in the less variable northern part with lower gradients, there are large regions with a low density even with 500 monitoring points. This clustering is because an additional point is set at the location of the largest deviation between the interpolated and the actual surface (APE). Thus, the first 37 extra wells are placed exclusively in the south (light blue points). After that, points

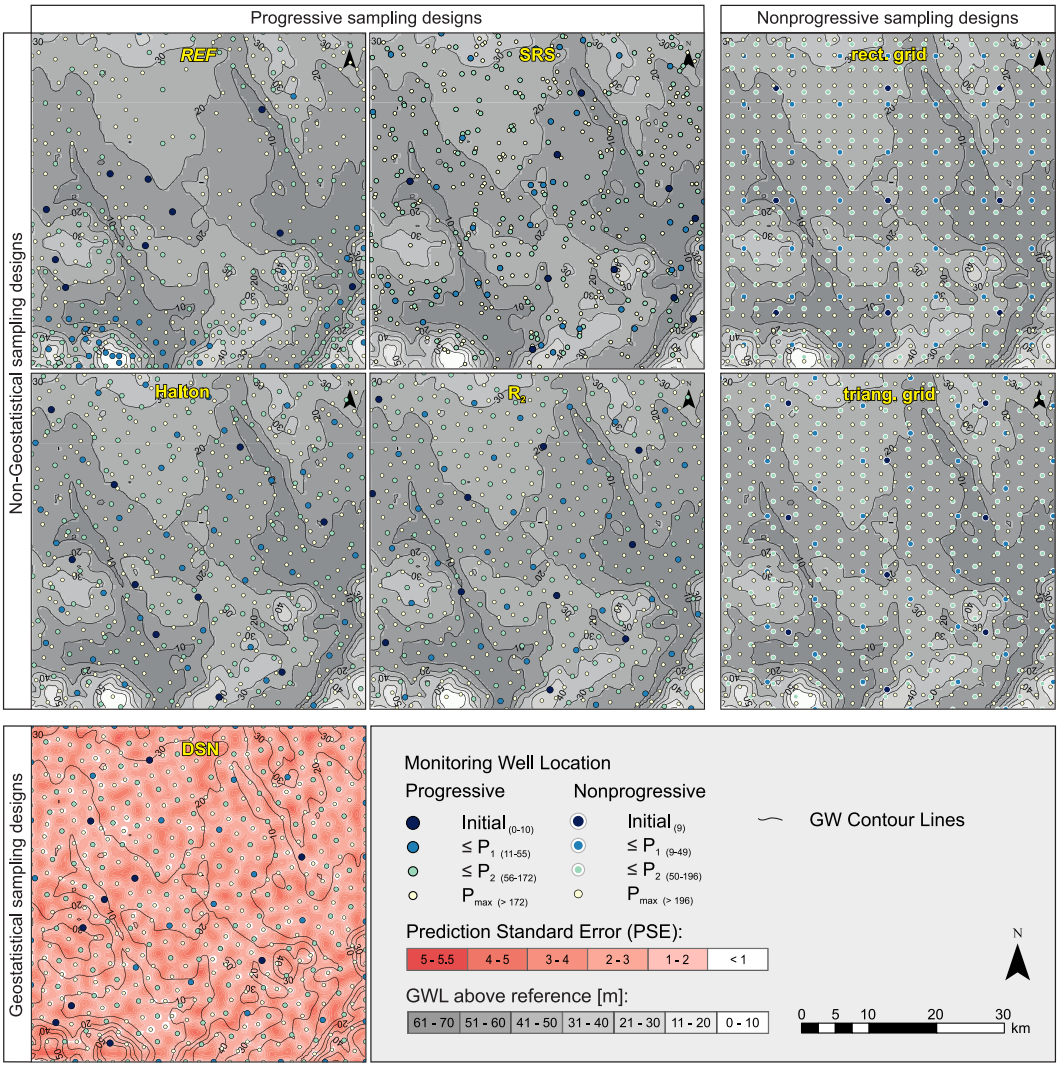


Figure 3. Resulting spatial distribution of the observation wells for surface ACF A based on the investigated network designs at the start (dark blue), less than 55 ($\leq P_1$, light blue), less than 172 ($\leq P_2$, cyan), and for 500 ($\leq P_{max}$, off white) observation wells (the numbers for the nonprogressive methods are adapted according grid symmetry). Progressive: design that can be extended with n additional observation wells. Nonprogressive: design that is not extensible without breaking the symmetry of the grid (points marked by white outlines). The Geostatistical sampling design (DSN) uses the prediction standard error as selection criterion for the placement of additional sample points.

will also be set further north although a concentration of additional points will remain in areas of high variability (cyan points).

SRS exhibits point clustering and regions without points. Since the points are placed randomly, this effect can be stronger or weaker from case to case. To take this random component into account, the placement was repeated ten times for *SRS*. The figure shows one of the ten examples.

To varying degrees, the low discrepancy methods (*Halton* and R_2 method) and the geospatial method *DSN* show uniformly distributed arrangements of the observation points. The *Halton* method shows a very even global point distribution. Locally, however, the method tends to place individual points closer together as the number of monitoring points increases (cyan and light yellow points). *DSN* shows a uniform local point distribution with relatively consistent neighboring distances. Since the kriging variance depends on the distance to the closest n-observation points, a high prediction standard error results at the boundary of the research area, and additional points are placed along that edge. The R_2 method shows both an even global distribution and uniform distances to the neighboring n-observation points across all observation densities.

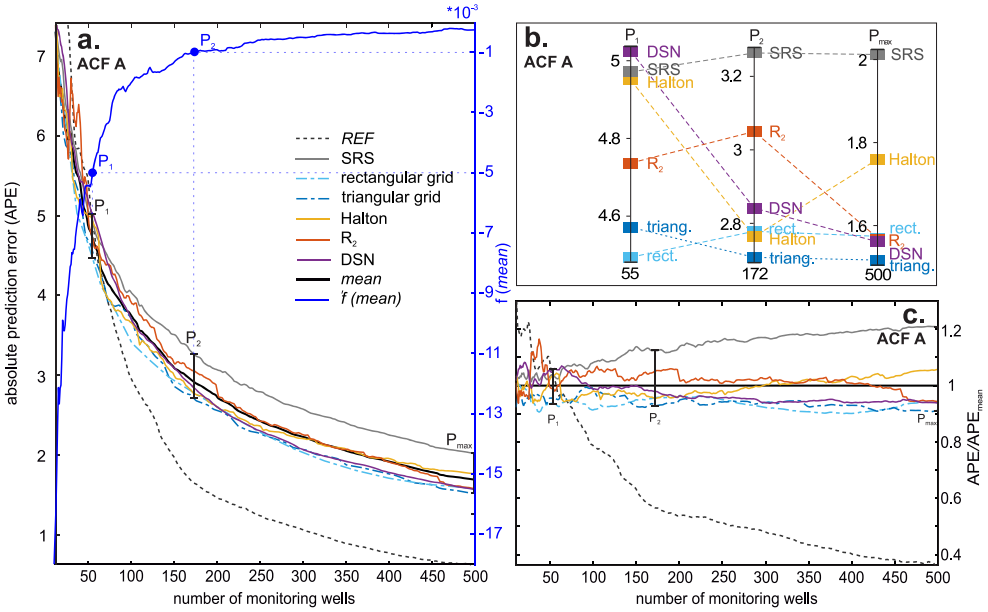


Figure 4. (a) Absolute prediction error (APE) of the tested methods for ACF A (left *y* axes) as a function of the number of observation wells and the derivative of the mean SAPE of the methods (excluded REF; right *y* axes). (b) Detailed view of P_1, P_2, and P_{max}. P_1 and P_2 are points at which the numerical derivative of the mean SAPE of the methods (excluded REF) has the value 0.005 and 0.001. (c) Ratio of the APE of the individual design approaches to the mean APE of all methods (excluding REF). A value <1 means that the individual design approach is better than average, >1 that it is poorer than average.

3.2. Absolute Prediction Errors

Figure 4a shows the *APE* for all tested design approaches for surface ACF A as an example (diagrams for all surfaces are provided as supporting information, Figure S1) as a function of the number of observation wells. The points P_1 and P_2 were defined such as the numerical derivatives of the mean SAPE of all methods (excluding *REF*) are 0.005 and 0.001, respectively. Consequently, an additional monitoring point results in a reduction of the initial mean SAPE by 0.5% for P_1 and 0.1% for P_2 and hence can be used for the definition of a well density for the respective information/cost ratio. The idea behind the thresholds P_1 and P_2 is that below P_1, the observation density seems insufficient, and the placement of additional wells leads to a substantial improvement of the results, while above P_2 further observation wells lead only to minor and thus possibly inefficient improvements. The decision at which information/cost ratio these points are defined can only serve as a rough guideline and must, of course, be adapted to the respective requirements of each site.

The theoretical maximum information content is reached at a density of 1 point per pixel, that is, the value is known for each grid cell. The grid size of the output maps needs to match the sampling density and scale at which the processes of interest occur (Hengl, 2009), since small-scale variabilities can only be displayed up to pixel size and the coarser the resolution of a GWS, the more small-scale variations disappear. P_1 was achieved for ACF A with 55 observation points and P_2 with 172 observation points. Figures 4b and 5b show a detailed view of P_1, P_2, and P_{max} (for the maximum tested number of 500 observation points). Figures 4c and 5c show the deviations of the MSAPE of the individual design approaches from the mean MSAPE of all methods (except *REF*). A value <1 means that the individual design approach performs better than average, >1 that it performs worse than average.

To make more general statements about the different methods, the individual APE of each surface and method was standardized (SAPE) by division through the maximum APE of all methods per surface to make the absolute error values, which have different magnitudes for each surface, comparable. Afterward, the mean over all surfaces of the SAPE for each method was computed (MSAPE; Figure 5a) to determine the overall performance of the method.

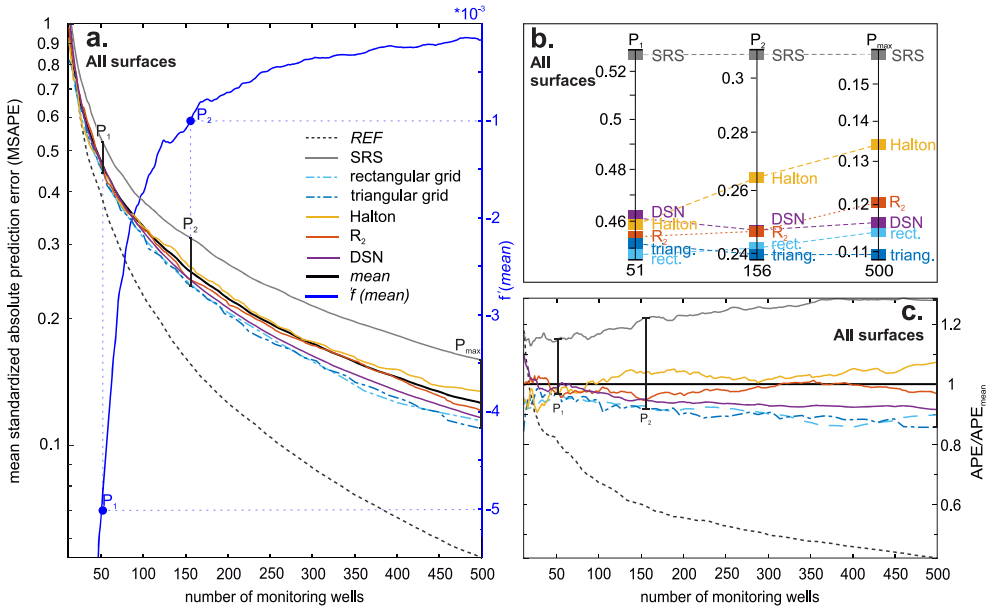


Figure 5. (a) Mean standardized absolute prediction error (MSAPE) of the individual methods for all nine surfaces (left *y* axes) and the derivative of the averaged MSAPE of the methods (excluded REF; right *y* axes). P_1 and P_2 are points at which the numerical derivative of the averaged MSAPE of the methods (excluded REF) has the value 0.005 and 0.001. (b) Detailed view of P_1, P_2, and P_{max}. (c) Ratio of the MSAPE of the individual design approaches to the averaged MSAPE of all methods (excluding REF). A value <1 means that the individual design approach is better than average, >1 that it is poorer than average.

In Both Figures 4a and 5a as well as Figure S1, three facts stand out:

1. Visible at first glance is that *SRS* underperforms, and it produces the largest error consistently for all surfaces as well as all point densities.
2. For the low end of number of monitoring wells (less than about 50, respectively <P_1), it seems arbitrary which design method performs best. This can be explained by the fact that below a certain well density, it is not possible to record and map small-scale variabilities of the tested surface and, thus, it is more or less a matter of luck whether the observations points are placed in locations that allow for an interpolation that is similar to the actual surface. For the ACF A, for example, P_1 would correspond to an observation density of one observation well per 50 km^2 only. Some of the designs even seem to outperform *REF*. This can be explained by the fact that with *REF* an additional point is set on the location of the largest difference between actual and interpolated surface. With a few observation points, however, the elimination of the largest but possibly very small-scale (regarding area) error does not always lead to the smallest global error.
3. For higher point densities, a ranking of the design strategies becomes identifiable, but the resulting errors (except *SRS*) are in a narrow range, and none of the methods is clearly superior to the other methods for every surface, regardless of the number of monitoring wells.

It can, therefore, be concluded that unless there is a sufficiently dense observation point coverage, there is more benefit in adding an additional observation point than in the use of the supposedly best design (as long as the wells are reasonably distributed). Which method performs best in detail depends on the characteristics of the surface and the density of the existing monitoring wells (Figure S1). This is mainly due to large-magnitude but small-scale variabilities in the surfaces. If there is no monitoring well in such an area, this can increase the APE of a method excessively although most of the surface is well represented. In addition, none of the tested method comes close to the (known by a priori) theoretically best possible design *REF*.

In detail, for the majority of point densities and tested surfaces, the regular grid arrangements (*triangular grid* and *rectangular grid*) achieved slightly lower errors than the other methods (Figures 5a and S1). However, their advantage of the excellent spatial sampling coverage is in contrast to the nonprogressive characteristic of these design approaches, which makes them unsuitable for many applications, unless the

final number of available wells is known in advance. On average, the *triangular grid* shows better results than the *rectangular grid*.

In the case of the tested spatial coverage sampling methods, the R_2 method, in particular, shows good results across all observation point densities and for most of the surfaces. The resulting error is only slightly higher for the majority of surfaces than for the grid methods. Since the locations are only based on mathematical sequences, the spatial coverage sampling methods have the advantage, which they can be used without prior measurements to construct an effective and evenly spaced monitoring network from scratch. Furthermore, the placement of the observations points is reproducible and does not change over time (with possibly different measurements, as it is the case for *DSN*). The *Halton* method also delivers acceptable results, especially at low observation densities. At higher densities, it produces only mediocre results. As a conclusion, the R_2 method is preferable.

The geostatistical *DSN* method led to the lowest errors among the tested progressive network designs, only slightly higher compared to the grid methods. As an advantage over the grid methods, it can be sequentially extended and is, therefore, more appropriate for most the observation of groundwater levels networks. However, since *DSN* selects the location of new observation points based on the maximum prediction standard error, it requires an existing kriging interpolation and can therefore only be used for adding new sampling locations to an existing monitoring network.

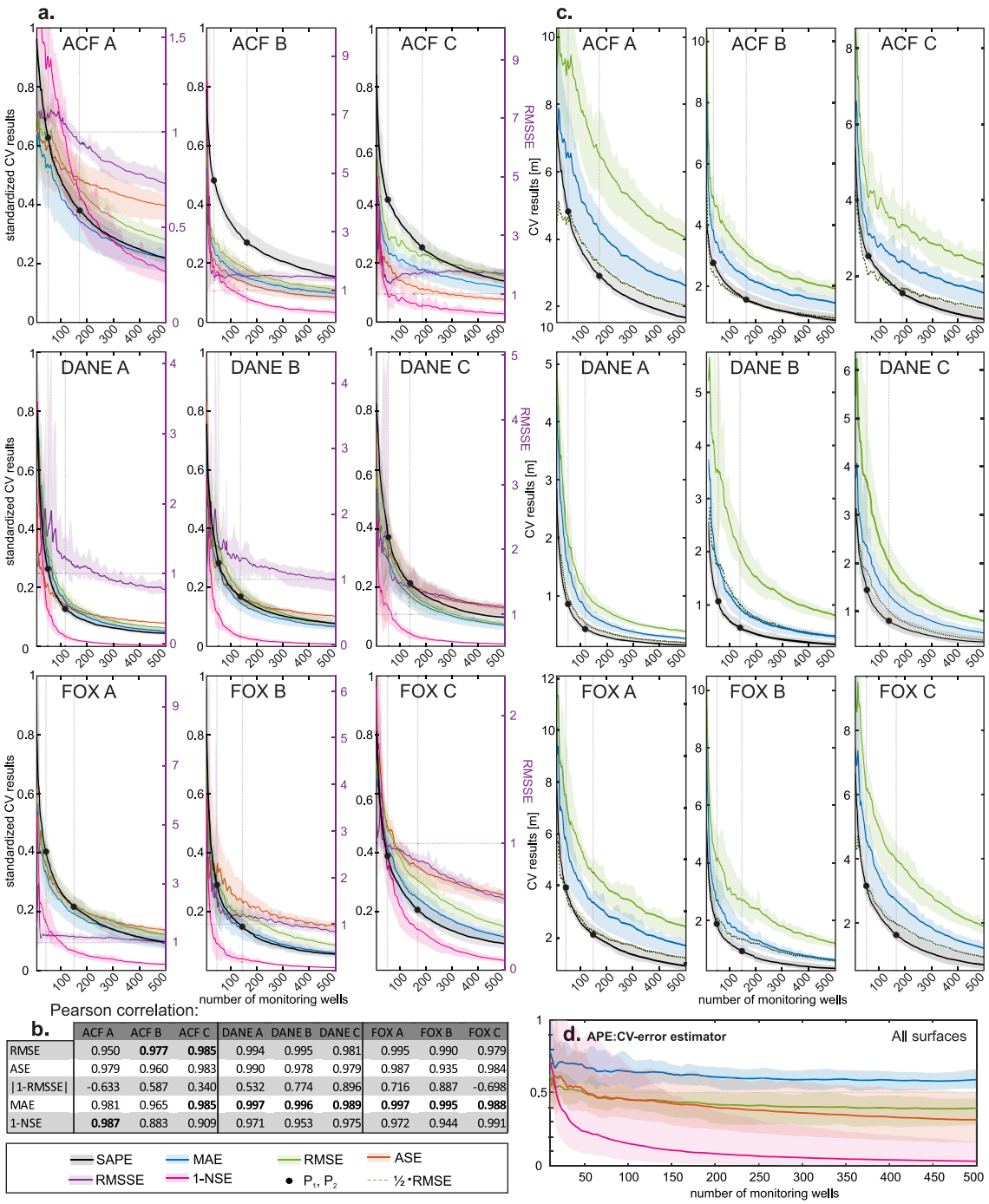
P_1 was achieved for the nine surfaces with 37–64 observation points (mean 51) and P_2 with 118–186 observation points (mean 156). Since the surfaces have different spatial resolutions, with pixel sizes between 100 and 500 m (and therefore varying degrees of detail), an observation density in observation wells per km^2 at which P_1 or P_2 is reached could be misleading. However, since all surfaces are 100×100 pixels in size, a comparison can still be made regarding an observation density per pixel. In the case of P_1, that means that on average for all surfaces and methods, for a density of $5.1 \cdot 10^{-3}$/pixel, a reduction of the initial mean SAPE by 0.5% is achieved, and for P_2, $1.56 \cdot 10^{-2}$/pixel, respectively. That means that on average, about 0.51% (with a range of 0.37–0.64%) of all possible sampling options (imaginary pixels of a grid reflecting the assumed variability of the groundwater) should be sampled, since until then, there is a significant error reduction of 0.5% with each additional well, while on average, with above 1.56% of all possible sampling options being sampled (with a range of 1.18–1.86%), the error reduction becomes considerably less (0.1% with each additional well), and the resulting information/cost ratio might become too low. To compare these values (in density per pixel) with a real observational density in wells per km^2, it must be assumed that the resolution of an imaginary grid is adapted to the actual variability of the GWS to be expected. As a rule of thumb, the observation density at P_1 and P_2 can be used as an indicator for evaluating existing observation networks and planning new ones.

3.3. CV results

Although an assessment of the error based on the deviation from the real groundwater surface (APE) seems necessary to determine both the number of wells required and the best method for their placement, this deviation will not be available in practice. Only the error at the existing observation points can be determined by CV. Thus, we compare the APE with different CV error statistics to examine whether one of them is more appropriate as a proxy for the averaged APE as well as suitable point densities for the assessment of an information/cost ratio.

Figure 6a shows a comparison of the averaged standardized APE (SAPE) of all design approaches and the averaged, standardized global CV error estimates (MAE, RMSE, ASE, and NSE), as well as the averaged RMSSE for all surfaces. Along with the averages, the ranges between the minimum and maximum curves are given as shaded areas in the same color. For a better comparison, NSE is shown as 1-NSE. The standardization aims to present the error developments of the different design approaches so that they can be compared with each other. Furthermore, it should be shown whether and to what extent conclusions can be drawn from the shape of the individual CV curves compared to the SAPE curve regarding the assessment of an information/cost ratio.

Figure 7b shows the Pearson correlation coefficient r between the CV-error estimates and the APE. Except for RMSSE, all methods consistently have a very strong correlation with APE. RMSE (between 0.950 and 0.995) and MAE (0.965 and 0.997) show the strongest correlation to the APE, which makes them to



	ACF A	ACF B	ACF C	DANE A	DANE B	DANE C	FOX A	FOX B	FOX C
RMSE	0.950	**0.977**	**0.985**	0.994	0.995	0.981	0.995	0.990	0.979
ASE	0.979	0.960	0.983	0.990	0.978	0.979	0.987	0.935	0.984
|1-RMSSE|	-0.633	0.587	0.340	0.532	0.774	0.896	0.716	0.887	-0.698
MAE	0.981	0.965	**0.985**	**0.997**	**0.996**	**0.989**	**0.997**	**0.995**	**0.988**
1-NSE	**0.987**	0.883	0.909	0.971	0.953	0.975	0.972	0.944	0.991

Figure 6. (a) Comparison of the standardized average prediction error (SAPE), RMSSE, and the standardized averaged MAE, RMSE, and ASE along with the ranges (shaded colors). (b) Pearson correlation between the SAPE and the respective CV-results (best method in each case is in bold type). (c) Comparison of the averaged prediction error (APE), the MAE, and RMSE. (d) Development of the ratio of APE and the individual cross-validation results (mean values of all tested methods and areas) for APE:MAE, APE:RMSE, APE:ASE, and APE:(1-NSE).

reliable qualitative error estimators. This is followed by 1-NSE (0.883–0.987). Since the RMSSE should go to 1, the Pearson's *r* of|1-RMSSE|instead of RMSSE was calculated. This variable shows the worst correlation with the APE (−0.633 to 0.896) of the tested estimators.

Since both MAE and RMSE have the same units as the estimated quantity, Figure 6c compares them with the APE to examine to what extent their value is suitable for quantifying the actual absolute error. Since

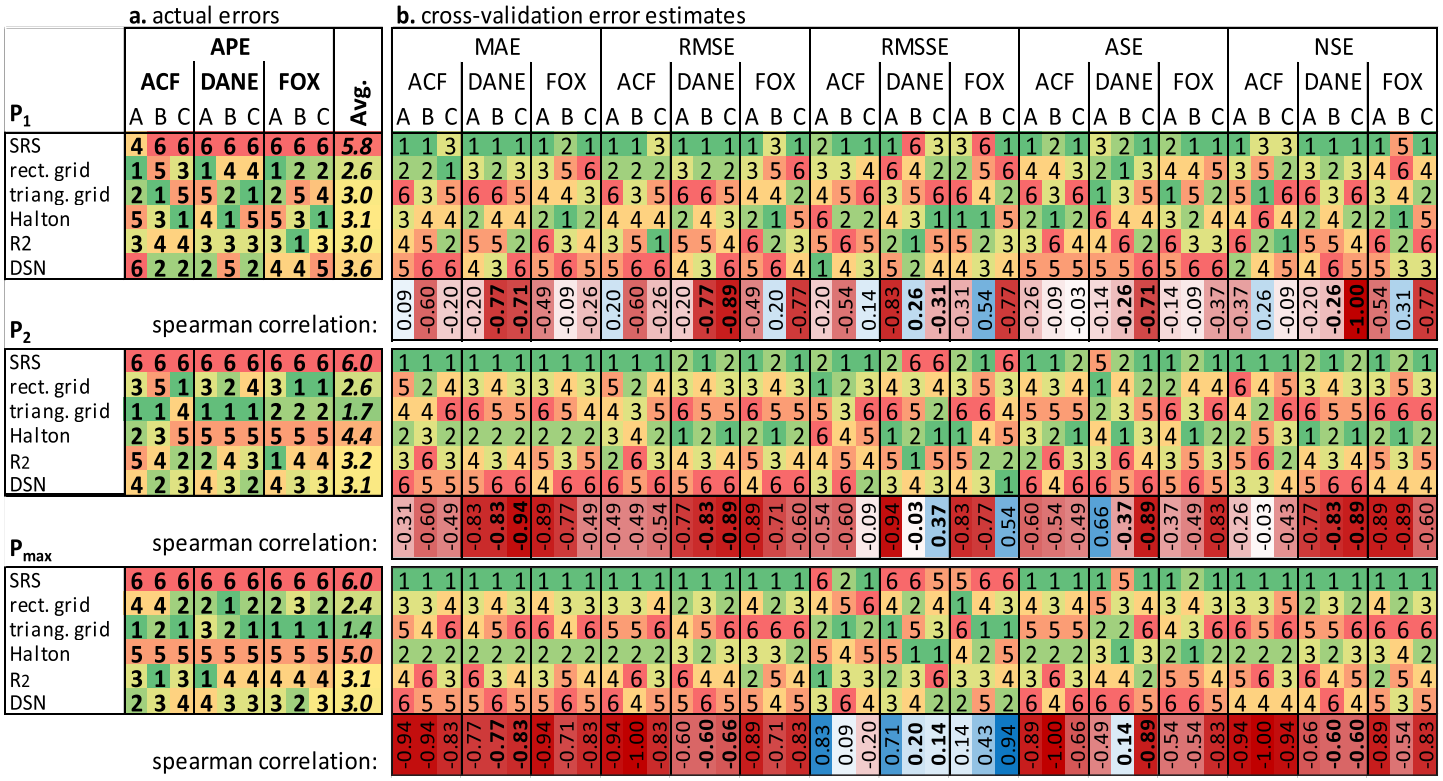


Figure 7. Comparison of the (a) ranking of the six investigated methods on all surfaces based on their absolute prediction errors (APE) and (b) the CV error estimates mean absolute error (MAE), root mean square error (RMSE), root mean square standardized error (RMSSE) and average standard error (ASE), Nash-Sutcliffe Efficiency (NSE), from best (1, green) to worst (7, red, RMSSE and NSE ranked by proximity to 1), along with the Spearman rank correlation between the APE and the CV error estimates from 1 (blue) to −1 (red).

RMSE squares the errors before averaging them, it gives a relatively high weight to large errors. Thus, the RMSE is clearly above the APE and the MAE for every surface. The APE corresponds approximately to ½ RMSE and to 2/3 MAE, which can, according to our data, therefore be used as a rough quantitative estimate for the actual error.

This can also be seen in Figure 6d, which illustrates the development of the ratios of APE and the CV error estimators (APE/MAE, APE/RMSE, APE/ASE, and APE/(1-NSE)). The standard deviation caused by the different design approaches is essentially lower for RMSE (0.09) and MAE (0.08) than for APE (0.22), RMSSE (1.29), and 1-NSE (0.24). 1-NSE shows the most irregular course. The ratio changes from 0.7 at the beginning to 0.03 at 500 observation points.

Figure 7 shows a comparison of the *APE* for all design approaches (Figure 7a) and the individual results of the CV errors (Figure 7b) as a ranking at the points P_1, P_2 and with 500 observation wells (P_{max}) for all nine surfaces. For better clarity, the results are ranked according to the error size (*RMSSE* and *NSE* ranked by proximity to 1). A table with the actual errors can be found in the supporting information (Table S2). The results show that it is not possible to evaluate the different design approaches based on their CV errors. With increasing measurement density, there is a moderate to strong negative Spearman rank correlation between the APE-based error ranking and the CV error estimates (except RMSSE). Thus, the *SRS* design, which actually led to the largest errors on any surface, is mostly classified as the supposedly best design approach, while the *triangular grid* producing on average the smallest APE is predominantly classified as the least suitable design. We therefore advise not to evaluate different designs based on their CV-error statistics.

4. Conclusions

The focus of this study was to compare different design approaches for groundwater level monitoring networks (GLMN) to find out if there is an extensible design that allows reliable spatial estimates of

groundwater level with an optimal information/cost ratio. Additionally, we examined what quality differences result from the use of different design approaches and which are the most suitable error statistics to evaluate the quality of the interpolated groundwater surface. For this purpose, we used nine potentiometric groundwater surfaces, extracted from three regional MODFLOW groundwater flow models to compare the interpolation results to an a priori reference. The sampling designs examined were random sampling, grid sampling (*triangular* and *rectangular grid*), spatial coverage sampling (low-discrepancy methods), and geostatistical sampling (densify sampling network).

The results show that the number of monitoring wells has more beneficial influence on the interpolation result than their spatial distribution (design), as long as a reasonably even spatial distribution is given. All tested sampling designs led to significantly better results than *SRS*, but none of these designs proved to be clearly superior to the others. Which method performs best is mostly dependent on the density of the GLMN and the characteristics of each individual groundwater surface.

Interpolated groundwater surfaces based on systematic sampling approaches (rectangular and triangular grid) showed on average the smallest actual APE at all observation densities. Due to their nonprogressive nature, they are only suitable for the construction of a new GLMN with a defined number of wells, which will not be extended in the future. The *triangular grid* design showed on average better results than the *rectangular grid* design.

The geostatistical *DSN* method, in which the location of an additional observation point is selected based on the maximum prediction standard error, resulted in the lowest APE among the tested progressive network designs. Despite its insignificantly higher APE compared to the grid designs, *DSN* has the advantage that the resulting design can be sequentially extended and is therefore more appropriate for most groundwater level observation networks. However, since *DSN* selects the location of new observation points on the basis of the maximum prediction standard error, it requires an existing kriging interpolation and can therefore only be used for adding new sampling locations to an existing monitoring network.

Consistently good results have been achieved with the low-discrepancy methods (*Halton* and R_2 *method*), which, to our knowledge, have not yet been used for GLMN designs before. Moreover, their locations are only based on mathematical sequences and can therefore be determined without prior measurements. Furthermore, the placement of the observations points is reproducible and does not change over time (with possibly different measurements, as it is the case for *DSN*). Among the low-discrepancy methods, the R_2 *method* delivers better results than the *Halton method* and should be preferred.

Based on the SAPE, we defined the points P_1 and P_2 where an additional well leads to a reduction of the initial mean SAPE by 0.5% for P_1 and 0.1% for P_2. Below P_1, the observation density seems insufficient, and the placement of additional wells leads to a substantial improvement of the results, while above P_2 further observation wells only lead to minor and thus inefficient improvements. On average for all surfaces and methods, the observation density for P_1 is $5.1 \cdot 10^{-3}$/pixel and $1.56 \cdot 10^{-2}$/pixel for P_2, respectively, that is, about 0.51% of all possible sampling options (imaginary pixels of a grid reflecting the assumed variability of the groundwater) should be sampled definitely, while on average when over 1.56% of all possible sampling options are sampled, the error reduction becomes considerably less, and the resulting information/cost ratio might become too low. To compare the density per pixel with a real observation density in wells per km^2, it must be assumed that the resolution of the grid is adapted to the actual variability of the GWS to be expected. Therefore, P_1 and P_2 can be used as rough guideline values for the required number of observation wells in the planning of a GLMN as well as the evaluation of an existing one.

From the results of global CV, we conclude that one should avoid comparing different designs based on the global average CV error estimation since there are strong negative correlations between APE and the CV-error estimates, especially at higher observation point densities. Thus, the CV error statistics are not appropriate to evaluate the results of different methods and to compare different design approaches. Which methods performs best can differ significantly from the actual error depending on the surface and on the CV statistics used. The actual benefit of CV comes not from using it in a global sense but rather in looking at the spatial distribution of the individual CV errors (correlated residuals, bias, normal distribution, etc.). According to our results, the CV error statistics, especially MAE and RMSE, can be helpful as a rough quantitative estimate for the actual error, though.

Acknowledgments

No conflicts of interest are declared. All used data sets can be obtained from USGS Water Resources NSDI Node. The authors like to thank Ty Ferré and two anonymous reviewers for their valuable comments.

References

Aguilar, F. J., Agüera, F., Aguilar, M. A., & Carvajal, F. (2005). Effects of terrain morphology, sampling density, and interpolation methods on grid DEM accuracy. *Photogrammetric Engineering and Remote Sensing*, *71*(7), 805–816. https://doi.org/10.14358/PERS.71.7.805

Ahmadi, S. H., & Sedghamiz, A. (2007). Geostatistical analysis of spatial and temporal variations of groundwater level. *Environmental Monitoring and Assessment*, *129*(1-3), 277–294. https://doi.org/10.1007/s10661-006-9361-z

Alfonso, L., Ridolfi, E., Gaytan-Aguilar, S., Napolitano, F., & Russo, F. (2014). Ensemble entropy for monitoring network design. *Entropy*, *16*(3), 1365–1375. https://doi.org/10.3390/e16031365

Babbar-Sebens, M., & Minsker, B. (2010). A Case-Based Micro Interactive Genetic Algorithm (CBMIGA) for interactive learning and search: Methodology and application to groundwater monitoring design. *Environmental Modelling & Software*, *25*(10), 1176–1187. https://doi.org/10.1016/j.envsoft.2010.03.027

Baxter, E. (2016). *Distributed hydrologic modeling using GIS* (3rd ed., Vol. 48). Dordrecht, Netherlands: Springer.

Birch, C. P. D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, *206*(3-4), 347–359. https://doi.org/10.1016/j.ecolmodel.2007.03.041

Bohling, G. (2005). Introduction to geostatistics and variogram analysis (pp. 1–20). http://people.ku.edu/~gbohling/

Brown, J. A., Robertson, B. L., & McDonald, T. (2015). Spatially balanced sampling: Application to environmental surveys. *Procedia Environmental Sciences*, *27*, 6–9. https://doi.org/10.1016/j.proenv.2015.07.108

Brus, D. (Ed.) (2010). *Design-based and model-based sampling strategies for soil monitoring*. The Netherlands: Soil Science Centre, Wageningen University and Research Centre.

Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. *Geoderma*, *80*(1-2), 1–44. https://doi.org/10.1016/S0016-7061(97)00072-4

Brus, D. J., de Gruijter, J. J., & van Groenigen, J. W. (2006). Chapter 14 Designing spatial coverage samples using the k-means clustering algorithm. In *Digital Soil Mapping - An Introductory Perspective, Developments in Soil Science* (pp. 183–192). Amsterdam: Elsevier.

Cameron, K., & Hunter, P. (2002). Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: A case study. *Environmetrics*, *13*(5-6), 629–656. https://doi.org/10.1002/env.582

Chandan, K. S., & Yashwant, B. K. (2017). Optimization of groundwater level monitoring network using GIS-based geostatistical method and multi-parameter analysis: A case study in Wainganga Sub-basin, India. *Chinese Geographical Science*, *27*(2), 201–215. https://doi.org/10.1007/s11769-017-0859-9

Chunmei, W., Qinke, Y., Hongyan, L., Weiling, G., Jupp, D. L., & Rui, L. (2013). Influence of resolution on elevation and slope at watershed scale in Loess Plateau. *Springerplus*, *2*(Suppl 1), S13. https://doi.org/10.1186/2193-1801-2-S1-S13

Dalal, I. L., Harwayne-Gidansky, J., & Stefan, D. (2008). On the fast generation of long-period pseudorandom number sequences. In *2008 IEEE Long Island Systems, Applications and Technology Conference* (pp. 1–9). New York: IEEE.

Delmelle, E. M. (2014). Spatial sampling. In M. M. Fischer & P. Nijkamp (Eds.), *Handbook of Regional Science* (pp. 1385–1399). Berlin Heidelberg, Berlin, Heidelberg: Springer.

DeSimone, L. A., Walter, D. A., Eggleston, J. R., & Nimroski, M. T. (2002). *Simulation of ground-water flow and evaluation of water-management alternatives in the Upper Charles River Basin* (p. 94). Eastern Massachusetts: United States Geological Survey.

Dhar, A., & Patil, R. S. (2012). Multiobjective design of groundwater monitoring network under epistemic uncertainty. *Water Resources Management*, *26*(7), 1809–1825. https://doi.org/10.1007/s11269-012-9988-1

Esquivel, J. M., Morales, G. P., & Esteller, M. V. (2015). Groundwater monitoring network design using GIS and multicriteria analysis. *Water Resources Management*, *29*(9), 3175–3194. https://doi.org/10.1007/s11269-015-0989-8

Fischer, M. M., & Nijkamp, P. (2014). *Handbook of regional science* (5071103). Heidelberg: Springer Reference.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. New York: John Wiley and Sons.

Geuna, S. (2000). Appreciating the difference between design-based and model-based sampling strategies in quantitative morphology of the nervous system. *The Journal of Comparative Neurology*, *427*(3), 333–339. https://doi.org/10.1002/1096-9861(20001120)427:3<333::AID-CNE1>3.0.CO;2-T

Giustolisi, O., & Simeone, V. (2010). Optimal design of artificial neural networks by a multi-objective strategy: Groundwater level predictions. *Hydrological Sciences Journal*, *51*(3), 502–523. https://doi.org/10.1623/hysj.51.3.502

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*, *Applied geostatistics series* (p. 483). New York: Oxford University Press.

Gruijter, J. J., Bierkens, M. F. P., Brus, D. J., & Knotters, M. (2006). *Sampling for natural resource monitoring*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, *2*(1), 84–90. https://doi.org/10.1007/BF01386213

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*(384), 776–793. https://doi.org/10.2307/2288182

Hengl, T. (2009). *A practical guide to geostatistical mapping* (2nd ed., p. 270). Amsterdam: Hengl.

Hensley, D., & Su, F. E. (2004). Random walks with badly approximable numbers: Unusual applications of number theory. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, *64*, 95–101. https://doi.org/10.1090/dimacs/064/10

Hernandez-Stefanoni, J. L., & Ponce-Hernandez, R. (2006). Mapping the spatial variability of plant diversity in a tropical forest: Comparison of spatial interpolation methods. *Environmental Monitoring and Assessment*, *117*(1), 307–334. https://doi.org/10.1007/s10661-006-0885-z

Heuvelink, G. B. M., & Pebesma, E. J. (Eds.) (2002). *Is the ordinary kriging variance a proper measure of interpolation error*. Melbourne: RMIT University, Melbourne.

Hu, K., Li, B., Lu, Y., & Zhang, F. (2004). Comparison of various spatial interpolation methods for non-stationary regional soil mercury content. *Environmental Sciences*, *25*(3), 132–137.

Ikechukwu, M. N., Ebinne, E., Idorenyin, U., & Raphael, N. I. (2017). Accuracy assessment and comparative analysis of IDW, spline and kriging in spatial interpolation of landform (topography): An experimental study. *Journal of Geographic Information System*, *09*(03), 354–371. https://doi.org/10.4236/jgis.2017.93022

Johnston, K. (2004). *ArcGis 9: Using ArcGis geostatistical analyst* (p. 300). Redlands, CA: *GIS by ESRI*, Esri Press.

Júnez-Ferreira, H. E., & Herrera, G. S. (2013). A geostatistical methodology for the optimal design of space-time hydraulic head monitoring networks and its application to the Valle de Querétaro aquifer. *Environmental Monitoring and Assessment*, *185*(4), 3527–3549. https://doi.org/10.1007/s10661-012-2808-5

Kambhamettu, B. V. M. P., Allena, P., & King, J. P. (2011). Application and evaluation of universal kriging for optimal contouring of groundwater levels. *Journal of Earth System Science*, *120*(3), 413–422. https://doi.org/10.1007/s12040-011-0075-4

Kermorvant, C., Caill-Milly, N., Bru, N., & D'Amico, F. (2019). Optimizing cost-efficiency of long term monitoring programs by using spatially balanced sampling designs: The case of manila clams in Arcachon bay. *Ecological Informatics*, *49*, 32–39. https://doi.org/10.1016/j.ecoinf.2018.11.005

Kish, L. (1995). *Survey sampling* (p. 643). New York: Wiley classics library, Wiley.

Kollat, J. B., & Reed, P. M. (2006). Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. *Advances in Water Resources*, *29*(6), 792–807. https://doi.org/10.1016/j.advwatres.2005.07.010

Krige, D. G. (1951). A statistical approaches to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, *52*(6), 119–139.

Krivoruchko, K. (2011). *Spatial statistical data analysis for GIS users* (p. 1). Redlands, CA: Esri Press.

Kumar, S., Sondhi, S. K., & Phogat, V. (2005). Network design for groundwater level monitoring in Upper Bari Doab Canal tract, Punjab, India. *Irrigation and Drainage*, *54*(4), 431–442. https://doi.org/10.1002/ird.194

Leach, J. M., Coulibaly, P., & Guo, Y. (2016). Entropy based groundwater monitoring network design considering spatial distribution of annual recharge. *Advances in Water Resources*, *96*, 108–119. https://doi.org/10.1016/j.advwatres.2016.07.006

Li, J. (2008). *A review of spatial interpolation methods for environmental scientists*, *Record/Geoscience Australia, 2008/23* (Vol. xvi, p. 137). Canberra: Geoscience Australia.

Li, J., & Heap, A. D. (2008). *A review of spatial interpolation methods for environmental scientists*. Canberra: Geoscience Australia.

Loaiciga, H. A., Charbeneau, R. J., Everett, L. G., Fogg, G. E., Hobbs, B. F., & Rouhani, S. (1992). Review of ground-water quality monitoring network design. *Journal of Hydraulic Engineering*, *118*(1), 11–37. https://doi.org/10.1061/(ASCE)0733-9429(1992)118:1(11)

Luo, Q., Wu, J., Yang, Y., Qian, J., & Wu, J. (2016). Multi-objective optimization of long-term groundwater monitoring network design using a probabilistic Pareto genetic algorithm under uncertainty. *Journal of Hydrology*, *534*, 352–363. https://doi.org/10.1016/j.jhydrol.2016.01.009

Ma, T.-S., Sophocleous, M., & Yu, Y.-S. (1999). Geostatistical applications in ground-water modeling in South-Central Kansas. *Journal of Hydrologic Engineering*, *4*(1), 57–64. https://doi.org/10.1061/(ASCE)1084-0699(1999)4%3A1(57)

Masoumi, F., & Kerachian, R. (2008). Assessment of the groundwater salinity monitoring network of the Tehran region: Application of the discrete entropy theory. *Water Science and Technology: a Journal of the International Association on Water Pollution Research*, *58*(4), 765–771. https://doi.org/10.2166/wst.2008.674

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

Mogheir, Y., de Lima, J. L. M. P., & Singh, V. P. (2009). Entropy and multi-objective based approach for groundwater quality monitoring network assessment and redesign. *Water Resources Management*, *23*(8), 1603–1620. https://doi.org/10.1007/s11269-008-9343-8

Mogheir, Y., Singh, V. P., & de Lima, J. L. M. P. (2006). Spatial assessment and redesign of a groundwater quality monitoring network using entropy theory, Gaza Strip, Palestine. *Hydrogeology Journal*, *14*(5), 700–712. https://doi.org/10.1007/s10040-005-0464-3

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nunes, L. M., Cunha, M. C., & Ribeiro, L. (2004). Groundwater monitoring network optimization with redundancy reduction. *Journal of Water Resources Planning and Management*, *130*(1), 33–43. https://doi.org/10.1061/(ASCE)0733-9496(2004)130:1(33)

Ohmer, M., Liesch, T., Goeppert, N., & Goldscheider, N. (2017). On the optimal selection of interpolation methods for groundwater contouring: An example of propagation of uncertainty regarding inter-aquifer exchange. *Advances in Water Resources*, *109*, 121–132. https://doi.org/10.1016/j.advwatres.2017.08.016

Olea, R. A. (1984). Sampling design optimization for spatial functions. *Journal of the International Association for Mathematical Geology*, *16*(4), 369–392. https://doi.org/10.1007/BF01029887

Parsen, M. J., Bradbury, K. R., Hunt, R. J., & Feinstein, D. T. (2016). The 2016 groundwater flow model for Dane County, Wisconsin. In *Bulletin/Wisconsin Geological and Natural History Survey* (Vol. 110, p. 56). Madison, Wisconsin: Wisconsin Geological and Natural History Survey. https://water.usgs.gov/GIS/dsdl/gwmodels/WGNHS2016-Dane_County/WGNHS_B110-report.pdf

Prakash, M. R., & Singh, V. S. (2000). Network design for groundwater monitoring—A case study. *Environmental Geology*, *39*(6), 628–632. https://doi.org/10.1007/s002540050474

Reed, P., Kollat, J. B., & Devireddy, V. K. (2007). Using interactive archives in evolutionary multiobjective optimization: A case study for long-term groundwater monitoring design. *Environmental Modelling & Software*, *22*(5), 683–692. https://doi.org/10.1016/j.envsoft.2005.12.021

Roberts, M. (2018). The unreasonable effectiveness of quasirandom sequences. http://extremelearning.com.au/unreasonable-effectiveness-of-quasirandom-sequences/#GeneralizingGoldenRatio

Romero, V., Slepoy, R., Swiler, L., Giunta, A., & Krishnamurthy, T. (2005). Error estimation approaches for progressive response surfaces. In *46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Structures, Structural Dynamics, and Materials and Co-located Conferences* (pp. 269–289). Austin, TX: American Institute of Aeronautics and Astronautics.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics (SJS), Theory and Applications*, *5*(1), 27–52.

Sepulveda N., and Painter J.A. (2017), MODFLOW-2005 simulation of groundwater-flow budget for the lower Apalachicola-Chattahoochee-Flint River Basin in southwestern Georgia and parts of Florida and Alabama, 2008 12.

Siska, P., Goovaerts, P., Hung, I.-K., & Bryant, V. (2005). Predicting ordinary kriging errors caused by surface roughness and dissectivity. *Earth Surface Processes and Landforms*, *30*, 601–612. https://doi.org/10.1002/esp.1164

Theodossiou, N., & Latinopoulos, P. (2006). Evaluation and optimisation of groundwater observation networks using the Kriging methodology. *Environmental Modelling & Software*, *21*(7), 991–1000. https://doi.org/10.1016/j.envsoft.2005.05.001

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Proceedings//International Geographical Union, Commission on Quantitative Methods*, *46*(2), 234–240.

Uddameri, V., & Andruss, T. (2014). A GIS-based multi-criteria decision-making approach for establishing a regional-scale groundwater monitoring. *Environment and Earth Science*, *71*(6), 2617–2628. https://doi.org/10.1007/s12665-013-2899-5

van der Corput, J. G. (1935). *Verteilungsfunktionen: Koninklijke Akademie van Wetenschappen te Amsterdam*, *N. V. Noord-Hollandsche Uitgevers Maatschappij* (Vol. 68-70). Amsterdam: N. Z. Voorburgwal.

Vicente-Serrano, S. M., Saz-Sánchez, M. A., & Cuadrat, J. M. (2003). Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): Application to annual precipitation and temperature. *Climate Research*, *24*, 161–180. https://doi.org/10.3354/cr024161

Wackernagel, H. (1995). Ordinary kriging. In H. Wackernagel (Ed.), *Multivariate Geostatistics* (pp. 74–81). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, W. (Ed.) (2011). *International Symposium on Water Resource and Environmental Protection (ISWREP), 2011: 20-22 May 2011, Xi'an, China; proceedings.* Piscataway, NJ: IEEE.

Wild, B. (2009), Minimizing error variance in estimates by optimum placement of samples: A comparison of optimization techniques, *CCG Annual Report 11*, (Paper 406,).

Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, *63*(11), 1309–1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2

Worley, B. (2016). Subrandom methods for multidimensional nonuniform sampling. *Journal of magnetic resonance (San Diego, California)*, *269*, 128–137. https://doi.org/10.1016/j.jmr.2016.06.007

Wu, Y. (2004). Optimal design of a groundwater monitoring network in Daqing, China. *Environmental Geology*, *45*(4), 527–535. https://doi.org/10.1007/s00254-003-0907-x

Yang, F.-G., Cao, S.-Y., Liu, X.-N., & Yang, K.-J. (2008). Design of groundwater level monitoring network with ordinary kriging. *Journal of Hydrodynamics*, *20*(3), 339–346. https://doi.org/10.1016/S1001-6058(08)60066-9

Yarus, J. M., & Chambers, R. L. (1994). Stochastic modeling and geostatistics: Principles, methods, and case studies. In *AAPG computer applications in geology* (Vol. 3, p. 379). Tulsa, Okla: American Association of Petroleum Geologists.

Yeh, M.-S., Lin, Y.-P., & Chang, L.-C. (2006). Designing an optimal multivariate geostatistical groundwater quality monitoring network using factorial kriging and genetic algorithms. *Environmental Geology*, *50*(1), 101–121. https://doi.org/10.1007/s00254-006-0190-8

Zhang, Y., Pinder, G. F., & Herrera, G. S. (2005). Least cost design of groundwater quality monitoring networks. *Water Resources Research*, *41*, W08412. https://doi.org/10.1029/2005WR003936

Zhou, Y., Dong, D., Liu, J., & Li, W. (2013). Upgrading a regional groundwater level monitoring network for Beijing Plain, China. *Geoscience Frontiers*, *4*(1), 127–138. https://doi.org/10.1016/j.gsf.2012.03.008