Sensors **2014**, *14*, 19095-19114; doi:10.3390/s141019095

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Hybrid Optimal Design of the Eco-Hydrological Wireless Sensor Network in the Middle Reach of the Heihe River Basin, China

Jian Kang [1,4], Xin Li [1], Rui Jin [1,2,*], Yong Ge [3], Jinfeng Wang [3] and Jianghao Wang [3,4]

[1] Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou 730000, China; E-Mails: kangjian@lzb.ac.cn (J.K.); lixin@lzb.ac.cn (X.L.)

[2] Heihe Remote Sensing Experimental Research Station, Chinese Academy of Sciences, Lanzhou 730000, China

[3] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; E-Mails: gey@lreis.ac.cn (Y.G.); wangjf@lreis.ac.cn (J.W.); wangjh@lreis.ac.cn (J.W.)

[4] University of Chinese Academy of Sciences, Beijing 100049, China

* Author to whom correspondence should be addressed; E-Mail: jinrui@lzb.ac.cn; Tel.: +86-931-496-7965; Fax: +86-931-827-9161.

External Editor: Leonhard Reindl

Received: 19 July 2014; in revised form: 23 September 2014 / Accepted: 23 September 2014 / Published: 14 October 2014

Abstract: The eco-hydrological wireless sensor network (EHWSN) in the middle reaches of the Heihe River Basin in China is designed to capture the spatial and temporal variability and to estimate the ground truth for validating the remote sensing productions. However, there is no available prior information about a target variable. To meet both requirements, a hybrid model-based sampling method without any spatial autocorrelation assumptions is developed to optimize the distribution of EHWSN nodes based on geostatistics. This hybrid model incorporates two sub-criteria: one for the variogram modeling to represent the variability, another for improving the spatial prediction to evaluate remote sensing productions. The reasonability of the optimized EHWSN is validated from representativeness, the variogram modeling and the spatial accuracy through using 15 types of simulation fields generated with the unconditional geostatistical stochastic simulation. The sampling design shows good representativeness; variograms estimated by samples have less than 3% mean error relative to true variograms. Then, fields at multiple scales

are predicted. As the scale increases, estimated fields have higher similarities to simulation fields at block sizes exceeding 240 m. The validations prove that this hybrid sampling method is effective for both objectives when we do not know the characteristics of an optimized variables.

Keywords: eco-hydrological wireless sensor network; spatial sampling; hybrid optimization criterion; unconditional stochastic simulation

1. Introduction

Sensor networks, one of important components of Global Earth Observation System of Systems [1], promote the advancement of Earth system science and environmental science [2]. Sensor networks, as a revolutionary technique, are widely used in several huge observation projects, such as the Critical Zone Observatories [3], the National Ecological Observatory Network [4], the Terrestrial Environmental Observatories [5] and the Long-Term Ecological Research Network [6], to understand hydrological and ecological processes.

Compared with traditional observation methods, sensor networks have two significant advantages in catchment hydrology research: the real-time monitoring of hydrological, ecological and meteorological elements and the effective capture of the spatial and temporal variability of different parameters, as well as the local spatial mean.

With the development of international water science, the NSFC (National Nature Science Foundation of China) launched the Heihe Plan entitled "Integrated research on eco-hydrological process of the Heihe River Basin (HRB)" in 2010. The Heihe Water Allied Telemetry Experimental Research (HiWATER) program is the observation platform of the Heihe Plan, and the eco-hydrological wireless sensor network (EHWSN) is one of the fundamental experiments of HiWATER [7]. The EHWSN provides indispensable observations that allow HiWATER to address problems that include heterogeneity, scaling and uncertainty. To solve these problems, EHWSN is required to capture the spatial variability and temporal dynamics of soil moisture and temperature and to provide accurate ground-truth estimates at remote sensing pixel scales [8]. Therefore, an effective EHWSN must be optimally designed.

Because of the inherent stochastic character of natural processes, researchers are frequently faced with the problem of selecting a suitable sampling location [9,10]. Geostatistics are used to capture or represent spatial variations and have been widely applied in various fields, including geoscience, water resources, environmental science and soil science [11]. A large number of spatial sampling techniques have been discussed [12] and reviewed [13]. From the perspective of geostatistics, the optimal design of spatial sampling has three objectives.

First, it aims to accurately estimate variogram parameters. Warrick and Myers [14] proposed a method (the WM criterion) in which the distribution of paired points in the lag classes corresponds to a pre-specified distribution. Müller and Zimmerman [15] and Zhu and Stein [16] focused on increasing the estimation accuracy through comparing with assumed variogram parameters.

Second, an optimal design focuses on the precision of spatial statistical inference. Optimization strategies mainly include minimizing the maximum or average kriging variance using known variogram models [17,18] or evenly distributing the samples in the study region to indirectly reduce the kriging variance. The latter strategy uses the free model without assuming variogram parameters, employing methods, such as minimizing the mean of shortest distances (MMSD) [19], maximum entropy [20], fractal dimension [21] and mean squared distance to sides, vertices and boundaries [22]. These methods, which improve spatial predictions, are only applied in ideal fields that meet the second-order stationary assumption. However, land surface variables sometimes possess stratified characteristics, especially in larger research areas, and the second-order stationary assumption does not apply in these cases. A sampling method to address the spatial stratification based on the MSN (means of surfaces with non-homogeneity) theory was proposed by Wang *et al*. [23] and Hu and Wang [24]. It estimates a variogram for each stratum and requires a larger number of samples.

Of the above methods for spatial sampling, one type for estimating variogram leads to sample clustering, which provides poor field coverage for the spatial prediction. Though the other type of method can improve spatial prediction accuracy, it is inappropriate for variogram modeling, due to the limited availability of sampling for short lag classes. To eliminate the deficiencies in both methods, the third type of hybrid criteria is proposed. It is divided into two categories. The first category is the combined use of the WM and MMSD methods [25,26]. In this category, either WM or MMSD is performed first, followed by the other, each with a specified number of samples. This is a sequential optimization process that decreases the utilization rate of the samples. It is difficult to reasonably allocate the number of samples for each sub-criterion. The second category is the simultaneous minimization of kriging variance and improving estimation of variogram parameters by assuming a known variogram [27]. However, in practice, it is difficult to know the variogram of a target variable, and the rationality of the assumption parameters cannot be evaluated.

The ENWSN is required not only to capture the spatial variability of the observed variables, but also to infer true values at the pixel scale to validate remote sensing products. Therefore, a hybrid criterion needs to be established. Because we do not know the characteristic of the specific target variable, the criterion should be based on the free model. Furthermore, to use samples more efficiently, all samples should make a contribution to each sub-criterion of the hybrid criterion. Therefore, each sub-criterion should not be performed independently, but instead, both sub-criteria should be performed together. The existing hybrid methods, however, are not suitable. To achieve both goals of the EHWSN, we develop a combination criterion without any assumptions of the spatial autocorrelation structure of surface variables. It is expressed as an integrated objective function that makes the sample distribution as uniform as possible in both geography and feature (lag distance) space.

This paper is structured as follows. Section 2 introduces the requirements of the EHWSN and the study area. Section 3 describes the optimization criterion and assessment methods. Section 4 shows the tests of the developed hybrid criterion, the final results of the spatial EHWSN distribution and validates the results with a series of evaluation indexes, and Section 5 explains the merits and remaining questions associated with this hybrid criterion.

2. Requirements for Optimal Sampling by EHWSN

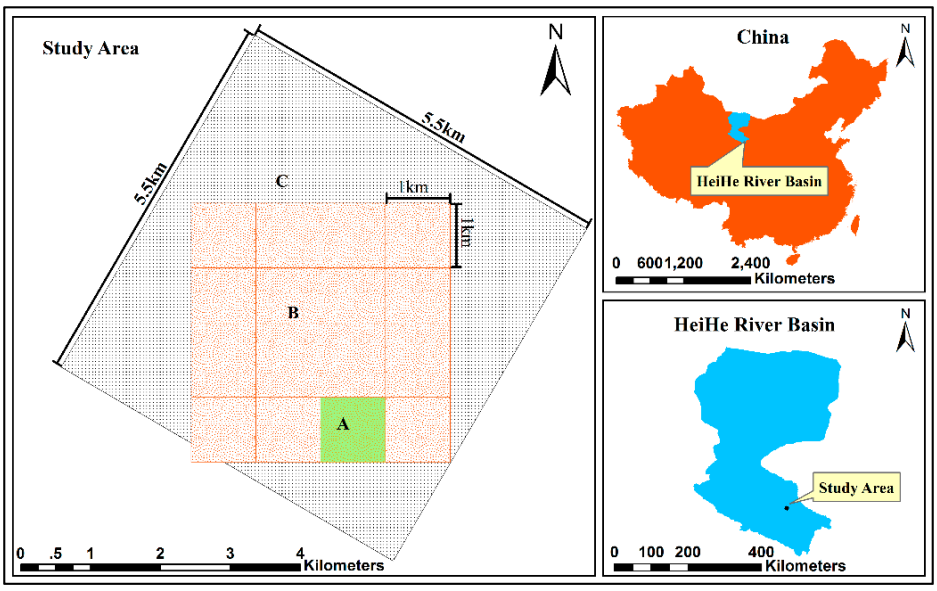
2.1. Objective of EHWSN

The EHWSN in the middle reaches of HRB aims to integrate a variety of distributed ecological and hydrological observations to capture the spatial-temporal variations of key eco-hydrological variables, including soil moisture, soil temperature and land surface temperature, and to obtain the ground truth for the validation of remote sensing products over a heterogeneous land surface. To better utilize multi-source satellite/airborne remote sensing data sets in studies of eco-hydrological processes, the EHWSN employs multi-scale validation for remote sensing products using kriging technology via nested and densely-distributed WSN nodes. Optimal spatial sampling of EHWSN nodes should be performed to help achieve the above objectives.

2.2. Experimental Area

The EHWSN is installed in a 5.5 km × 5.5 km observation matrix region located in the middle reaches of HRB, which covers both the Yingke and Daman irrigation districts of Zhangye oasis, in northwest China (Figure 1). The main crop type is seed corn, covering approximately 75% of the total area. Other plants, such as wheat, vegetables and fruits, are also represented. There is a dense canal network with five types of canals that forms the area's irrigation system. This irrigation management is the main source of land surface heterogeneity.

Figure 1. Map of the observation matrix region.



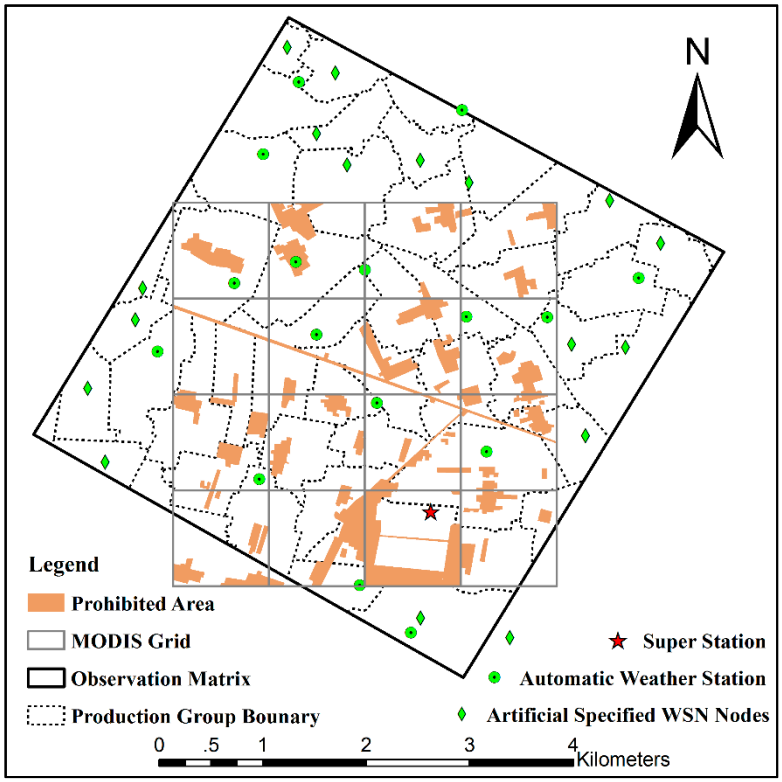
To effectively establish a nested WSN, the observation matrix is divided into three sub-regions: (A) the intensive region, with the same area as a MODIS pixel; (B) a 4 × 4 MODIS pixels region and (C) the surrounding region.

2.3. Arrangement of the EHWSN Nodes

A total of 180 EHWSN nodes will be installed, including 50 WATERNET nodes, 50 SoilNET nodes and 80 BNUNET nodes. The WATERNET nodes primarily observe the soil moisture, soil temperature and soil salinity at soil layer depths of 4 cm and 10 cm. The SoilNET nodes measure soil moisture and soil temperate at soil layer depths of 4 cm, 10 cm, 20 cm and 40 cm. The BNUNET nodes also observe soil moisture at 4 cm and soil temperature at 4 cm, 10 cm and 20 cm.

There are some artificial and natural condition limitations on the spatial distribution of EHWSN nodes. In total, 17 automatic meteorological stations (AMSs) have been installed in the observation matrix to measure the evapotranspiration over the heterogeneous land surface by observing the boundary layer conditions or the flux exchange between the atmosphere and land surface. Additionally, 17 BNUNET nodes have been artificially fixed in Region C, so that there is one node per production group. These pre-specified EHWSN nodes, together with the AMSs, are considered initial points during the optimization process (Figure 2). The remaining 163 nodes are optimized in Regions A and B. Of these, 56 nodes (50 SoilNET nodes and 6 WATERNET nodes) are designed to reveal the spatial variation at a scale of hundreds of meters in Region A. The remaining 63 BNUNET nodes and 44 WATERNET nodes in Region B are used to capture the spatial variation at an approximately kilometer scale. All of the EHWSN nodes are deployed on vegetation-covered land, because the instruments cannot be installed on other land surface types, such as roads, residential buildings, wind-defended forest or irrigation channels (Figure 2).

Figure 2. Initial eco-hydrological wireless sensor network (EHWSN) nodes in the optimization process.



3. **Methodology**

3.1. Hybrid Criterion

According to the requirements of EHWSN, we need to both effectively estimate the variogram parameters and ensure the spatial prediction accuracy when making inferences. Both requirements can be attributed to the optimal design of the sampling network; namely, how to simultaneously achieve the variogram estimation and minimizing the spatial estimation variance to the satisfied accuracy with the specified number of EHWSN nodes without assuming any variograms.

Based on the above goals, we establish an integrated hybrid model to simultaneously satisfy the two sub-criteria. This model is described by the following equation:

$$\Phi_{hybrid}(S) = w_1\Phi_{EP}^{norm} + w_2\Phi_{SP}^{norm}(S) \tag{1}$$

where Φ_{hybrid} is a weighted sum of two sub-criterions with weighted coefficients w_1 and w_2 and S is the optimized point set. *EP* represents a method that is effective for estimating variogram parameters, and *SP* is a method that is good for spatial prediction. Due to different dimensions, both Φ_{EP}^{norm} and Φ_{SP}^{norm} must be normalized in order to be added together.

3.1.1. Sub-Criterion to Estimate the Variogram Parameters

In geostatistics, the theoretical variogram $2\gamma(x,h)$ is a function that describes the degree of spatial dependence of a random spatial field V [28]. The estimator $2\hat{\gamma}(h)$ is the arithmetic mean of the squared differences between measurements Z at points x and $x + h$. The classical estimator of the variogram is defined as follows [29]:

$$2\hat{\gamma}(h) = \frac{1}{N(h)}\sum_{i=1}^{N(h)}[\mathrm{Z}(x+h) - \mathrm{Z}(x)]^2 \tag{2}$$

where $N(h)$ is the number of experimental pairs $[\mathrm{Z}(x), \mathrm{Z}(x + h)]$ with distance h and $\mathrm{Z}(x)$ is the value at location x.

We choose the WM criterion [14] for the *EP* method, which relies only on the distances between the points and does not depend on assumptions of the spatial autocorrelation structure of a variable. A predefined distribution of the number of coupled pairs in all lag classes should be optimized to improve the accuracy of variogram modeling. The desired distribution can be based on expert judgment, and Russo [30] suggests that a uniform distribution can reduce the estimation uncertainty. The objective function is a simple standard deviation between the expected number of point pairs $\overline{PP}$ and the realized value in the i-th lag class PP_i:

$$\Phi_{WM}(S) = \sqrt{\frac{1}{n_p}\sum_{i=1}^{n_p}(PP_i - \overline{PP})^2} \tag{3}$$

where S is a set of sampling points and n_p is the number of lag classes.

The function $\Phi_{WM}(S)$ is normalized as a coefficient of variation (CV) through division by $\overline{PP}$:

$$\Phi_{EP}^{norm}(S) = \frac{\Phi_{WM}(S)}{\overline{PP}} \tag{4}$$

where the $\overline{PP}$ value can approximated by the following:

$$\overline{PP} = \frac{N_p}{(\frac{1}{2} D_{max} / LS)} \tag{5}$$

where D_{max} is the maximum distance in the study area and LS the lag size. A rule of thumb is to multiply the number of lag classes by the lag size, which should equal approximately half of the maximum distance in the region. N_p denotes the total number of paired points and is given by $N_p = N(N-1)/2$, where N denotes the number of samples.

3.1.2. Sub-Criterion to Minimize the Estimation Variance

Normalization is difficult for the existing *SP* method based on a free model, such as MMSD, because we have little understanding of the statistical characteristics of objective function (e.g., mean, maximum and minimum). Thus, we need to develop a normalized *SP* criterion that can be compared with the WM criterion.

Yfantis *et al*. [31] confirmed that an equilateral triangle-shaped sampling network reduces estimation variance relative to a square or hexagonal network, and the best results are achieved when the regular point pattern forms equal-area Delaunay triangles. However, the measure of regularity is not sensitive to the area variance. With this in mind, a method to minimize the difference between the actual Delaunay triangle side length and the expected length (the MDS criterion) is proposed. The MDS is defined the same as in the WM criterion, *i.e*., as a standard deviation with the following form:

$$\Phi_{MDS}(S) = \sqrt{\frac{1}{n_s} \sum_{i=1}^{n_s} (SL_i - \overline{SL})^2} \tag{6}$$

where SL_i is the i-th side length of the Delaunay triangle network generated by the point set S, and the Delaunay triangle diagram is calculated by Fortune [32]. $\overline{SL}$ is the desired side length. The MDS criterion is chosen for the *SP* method, and Φ_{SP}^{norm} is defined as follows:

$$\Phi_{SP}^{norm}(S) = \frac{\Phi_{MDS}(S)}{\overline{SL}} \tag{7}$$

where Φ_{MDS} is normalized as *CV* though division by $\overline{SL}$. Because Thiessen polygons generated by equilateral Delaunay triangles are equal-area in an infinite region, we can approximate the $\overline{SL}$ value as follows:

$$\overline{SL} = \sqrt{\frac{2\sqrt{3}}{3} \times \frac{A}{N}} \tag{8}$$

where A denotes the area of study region and N is the total number of samples. A/N is the approximate area of the Thiessen polygons.

3.1.3. Determination of Weight Coefficients

Both Φ_{WM} and Φ_{MDS} are converted to *CV*, and the weight coefficients w in Equation (1) are calculated with the following equation:

$$w_i = \frac{CV_i}{\sum_{i=1}^{n} CV_i} \tag{9}$$

where w_i and CV_i are the weight coefficient and the coefficient of variation of the i-th indicator, respectively, and n is the number of indicators. Equation (9) implies that $\sum_{i=1}^{N} w_i = 1$. The two weight values in Equation (1) are not constant; their values change with CV_i during the optimization process, but their sum is equal to 1.

3.1.4. Optimization Algorithm

Our goal is to develop an optimal sampling scheme with a fixed number of sampling points via minimization of the Φ_{hybrid} value. It is necessary to find an effective way to optimize the objective function. In this paper, the simulated spatial annealing optimization algorithm (*SSA*) is employed to optimize a global sampling scheme [19,25,27,33–35]. *SSA* is a probabilistic method based on the Metropolis selection criterion [36], which can be written as follows:

$$\begin{aligned} P_T(S_i \to S_{i+1}) &= 1 && \text{if } \Phi_{hybrid}(S_{i+1}) \le \Phi_{hybrid}(S_i) \\ P_T(S_i \to S_{i+1}) &= exp\left(\frac{\Phi_{hybrid}(S_i) - \Phi_{hybrid}(S_{i+1})}{T}\right) && \text{if } \Phi_{hybrid}(S_{i+1}) > \Phi_{hybrid}(S_i) \end{aligned} \tag{10}$$

where T represents the annealing temperature, a positive control parameter that decreases with the optimization process, and i is the number of iterations. The parameter T is calculated by the follow equation:

$$T_{i+1} = \alpha T_i \tag{11}$$

where α is a parameter determined by users as a value slightly less than 1. In this paper, the α value is 0.95.

3.2. Generation of a Random Two-Dimensional Field for Validation

3.2.1. Unconditional Geostatistical Stochastic Simulation

To evaluate the ability of this proposed hybrid criterion to describe the spatial distribution characteristics of a regional variable, a two-dimensional stationary and isotropic field with values defined on the grid (x, y) with $x = 1, 2, \cdots, X$; $y = 1, 2, \cdots, Y$ is generated by stochastic simulation. The grid values are simulated using a sequential Gaussian simulation, in which ordinary kriging is used to estimate the local conditional probability distribution (LCPD) [37,38]. The simulation procedure requires the generation of spatial correlation values corresponding to a specified variogram or correlogram. Several models can be used for variogram modeling. In this study, we select the exponential model to express the spatial variation:

$$\gamma(h) = c_0 + c(1 - e^{-3h/a}) \tag{12}$$

where c_0 is the nugget, c is the sill and a is the range.

Assuming that a regional variable obeys the normal distribution with mean μ and standard deviation σ, the simulated value in each grid is given by the following equation [38]:

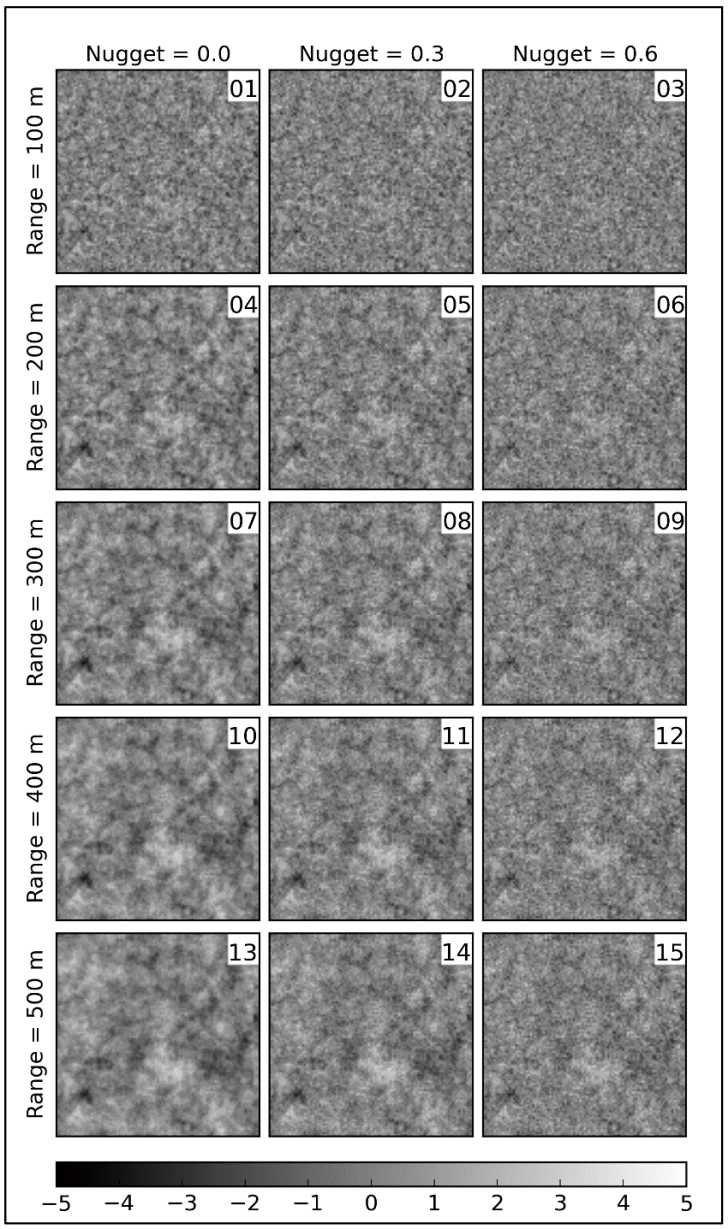
$$z(x,y) = xp \cdot std(x,y) + mean(x,y) \tag{13}$$

where *mean*(*x*, *y*) and *std* (*x*, *y*) are the kriging-based estimated mean and standard deviation for the grid (*x*, *y*), respectively. *xp* is drawn from the standard normal distribution and computed by the following equation:

$$xp = gauinv(p) \tag{14}$$

where *gauinv* is the numerically approximated inverse of the standard normal distribution function [39] and *p*, which represents a probability distribution function value, is a random number in the range 0–1.

Figure 3. Examples of the simulation fields with 95 × 95 grids produced by the variogram with different parameter combinations.



The LCPD on the grid (*x*, *y*) is estimated by searching all nearby grids with known value in the dependence distance. If there are less than 10 nearby grids, a value for the grid (*x*, *y*) is randomly chosen

from the normal distribution $N(\mu, \sigma)$. Each simulated datum becomes conditioning data for the next simulation step until all grids are simulated.

Setting $\mu = 0$ and $\sigma = 1$, a variogram model of $sill = 1$ is specified. The grid size is 40 m, and the number of grids is 95 × 95 for Regions A and B. In total, 50 realizations are generated for each of 15 parameter combinations of an exponential variogram model. Figure 3 shows one realization for each variogram.

3.2.2. Assessment Index

To validate the representativeness of samples, the *MAE* (mean absolute error) is defined to represent the degree of bias:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i^* - Y_i\right| \tag{15}$$

where Y_i^* is the prediction and Y_i is the true value.

After obtaining the experimental variogram, a curve is fitted for spatial interpolation. To compare the fitted curve with the true one, the *RE* (relative error) is defined as follows:

$$RE = \frac{\int_0^{range}\left|F(x \mid p_1, p_2, \cdots) - F^*(x \mid p_1^*, p_2^*, \cdots)\right|}{\int_0^{range}\left|F(x \mid p_1, p_2, \cdots)\right|} \times 100\% \tag{16}$$

where F and F^* represent the true and fitted variograms with different parameters p, respectively.

For the estimated block size, the prediction accuracy is usually assessed by the block kriging variance (*BKV*) as follows:

$$BKV = \sigma_A^2 - \omega^T D - \mu \tag{17}$$

where *BKV* is the estimated variance for the block A, σ_A^2 is the area-to-area covariance over Area A, μ is the Lagrange multiplier and D is a vector based on the estimated point-to-area covariance. To compare with the estimated fields from the different variograms and block sizes, the mean normalized *BKV*s ($MBKV_{norm}$) is calculated by the minimum-maximum normalization method:

$$MBKV_{norm} = \frac{1}{N}\sum_{i=1}^{N}\frac{BKV_i - BKV_{min}}{BKV_{max} - BKV_{min}} \tag{18}$$

where N is the number of estimation grids. BKV_{max} and BKV_{min} are the maximum and minimum *BKV*s, respectively. The BKV_i is the block kriging variance in the i-th grid.

Kriging variance only represents the estimation uncertainty, hence the similarity between the estimated field and the true field, and is defined as follows:

$$F(G,S) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|g_i - s_i|}{Max(g_i - s_i)}\right) \times 100\% \tag{19}$$

where G and S are histograms of two images. One image is simulated by stochastic simulation, and another is estimated by kriging. The image value is divided into N bins. The variables g_i and s_i are average values in the i-th bin.

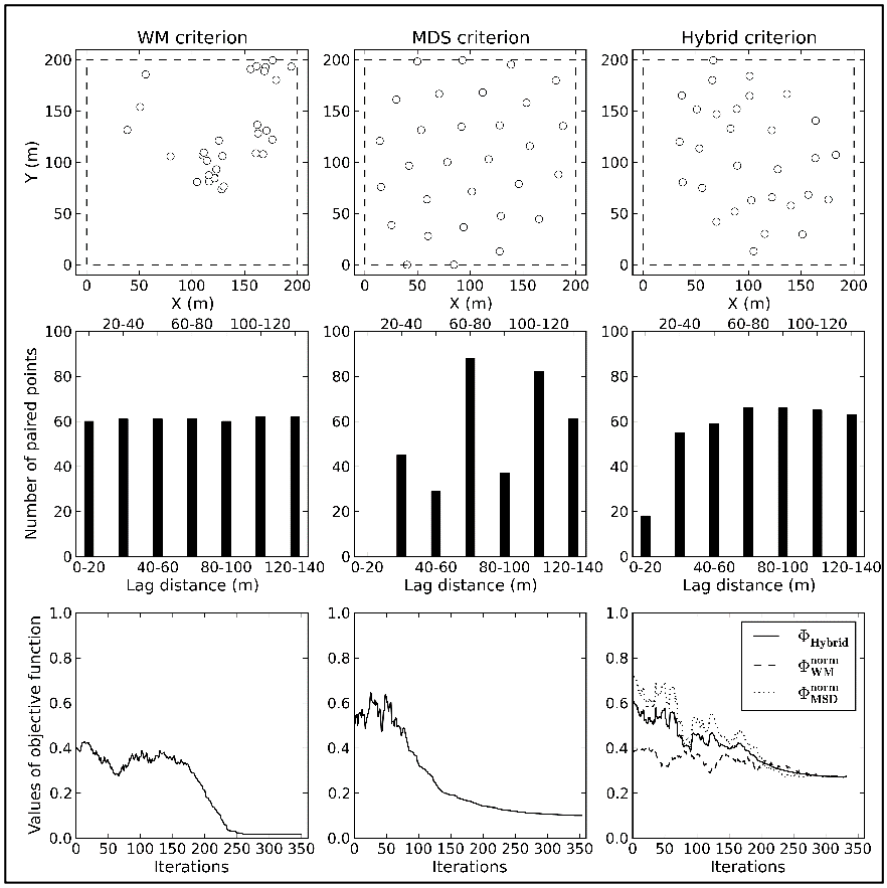
4. Results

4.1. Performance Test of Hybrid Criterion

As shown in Figure 4, there is a large contradiction between the spatial distributions of the WM and MDS criteria. WM is favorable for modeling spatial variation, but causes excessive aggregation, which negatively affects spatial estimation. MDS insufficiently captures the spatial variation due to the lack of information at short lag classes. Additionally, points near the boundaries do not have enough neighbors, leading to biases at the corners in MDS. This problem can be solved though initializing a few points in the corners.

The results of the hybrid optimal criterion are more or less similar to designs that are optimized by either WM or MDS. Though the hybrid model does not perfectly inherit all advantages of the sub-criteria, their defects are remedied. Compared to WM, the spatial distribution of samples generated by the hybrid criterion is superior for spatial statistical inferences, and compared to MDS, the distribution of samples in lag space is more complete.

Figure 4. Differences in spatial distribution, number of paired points and values of objective function between three optimization criteria.



Because the initial values of both sub-criteria are different, the spatial distribution of samples will tend toward the criterion with the smaller initial value if equal weights are assigned to each sub-criterion. In this test, the final results will tend toward the WM criterion. Hence, variable weight coefficients are reasonable during the entire optimization process.

4.2. EHWSN Optimization15

We have proposed a hybrid model-based optimization method with two sub-criteria for parameter estimation and spatial statistical inference. This method is applied to the EHWSN sampling design in the middle reaches of HRB. The relevant parameters are as follows: $\overline{PP}$ in Equation (5) is equal to 200 and $\overline{SL}$ in Equation (8) is equal to 130 m in Region A and 380 m in Region B. The size of each field block is approximately 40 m × 40 m. To avoid having more than one node in a field block (only the variability between fields is considered), a pair of points separated by less than 40 m is forbidden. The lag separation distance should coincide with the field spacing; thus, the lag size is set equal to 40 m. Lag classes less than the dependence distance should be given priority. However, it is difficult to select a suitable distance. Therefore, we attempt to obtain the dependence distance using the TVDI (Temperature-Vegetation Dryness Index) derived from a Thematic Mapper (TM) image. This image substitutes for soil moisture [40], because there is no prior high-resolution information of the soil moisture distribution. The TVDI can approximately represent the spatial distribution of soil moisture. Ultimately, the dependence distance is set equal to 680 m (Figure 5). The final result is shown in Figure 6. Optimization decreased the objective function Φ_{hybrid} from 0.84 to 0.28. The final values of Φ_{EP}^{norm} and Φ_{SP}^{norm} are 0.24 and 0.31, with weight coefficients of 0.44 and 0.56, respectively.

Figure 5. Experimental variogram (circles) and fitted curve (black line) of the Temperature-Vegetation Dryness Index (TVDI).

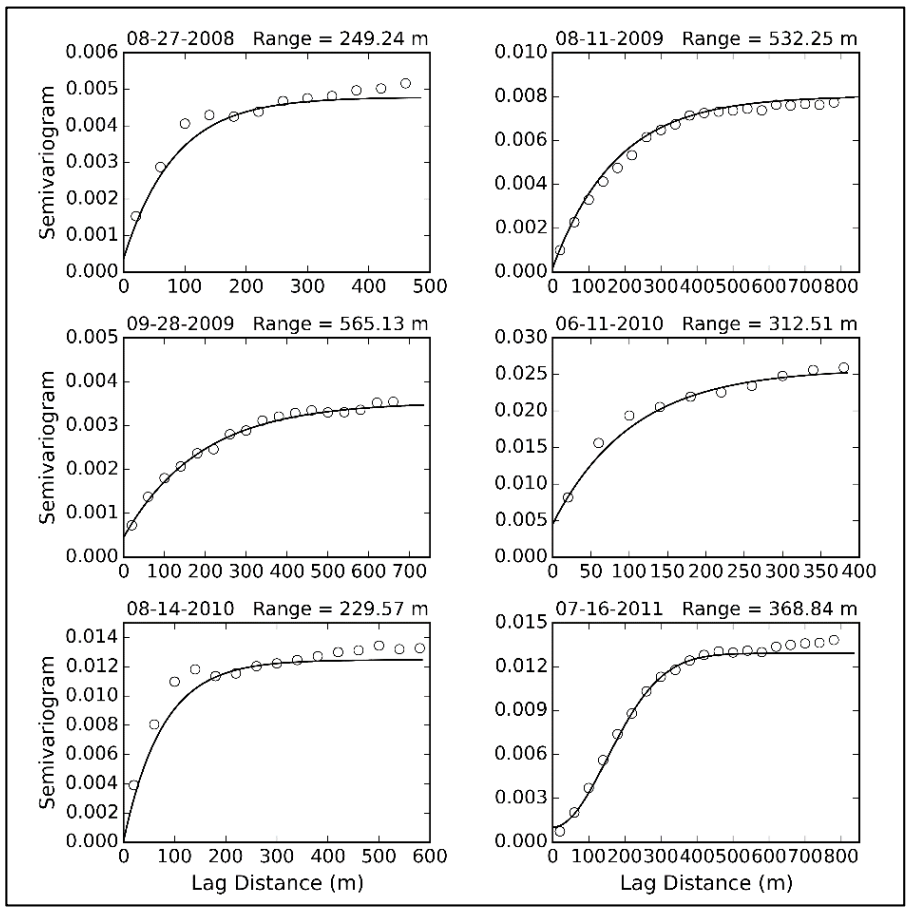
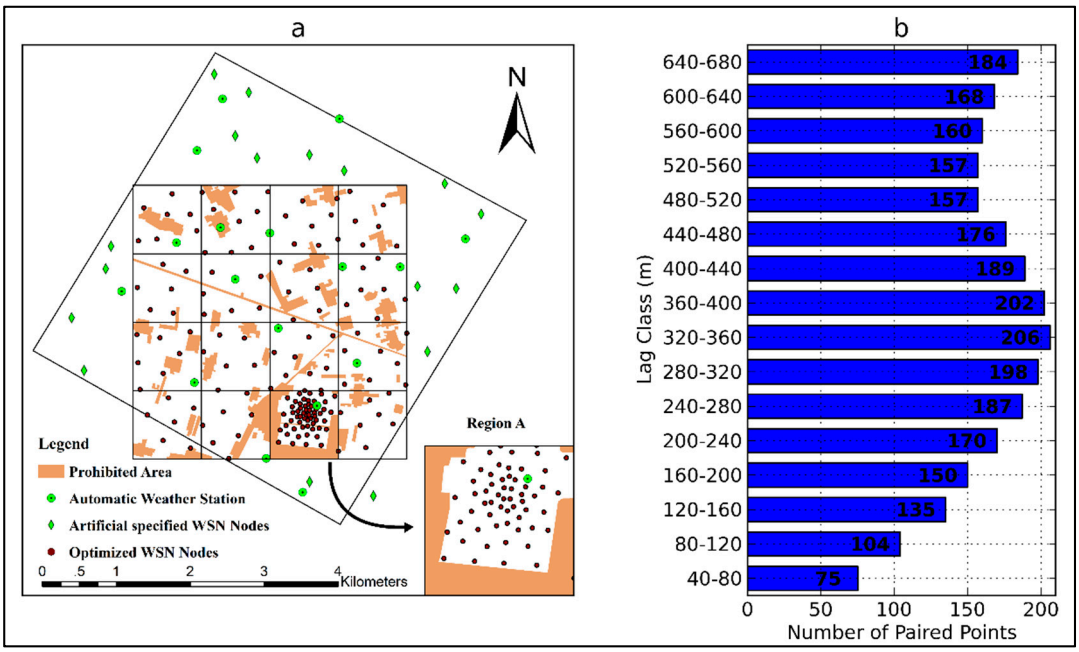


Figure 6. Optimized distribution of EHWSN nodes in the observation matrix (**a**) and the number of paired points in each lag class (**b**).



4.3. Validation

The optimization results of EHWSN need to be verified using simulated fields from several perspectives, including the representativeness, the accuracies of parameter estimation and spatial prediction.

4.3.1. Representativeness

The inferred values of target variables from spatial predictions at unsampled locations are based on the hypothesis that samples are representative. If data are sampled in an unrepresentative manner, the biased data cannot represent the overall properties in the area of interest, and the spatial predictions based on geostatistical techniques might be worse than those based on simple methods, such as inverse distance interpolation, surface interpolation and splines. To verify the representativeness, we compare the means and standard deviations (SD) between the samples and the populations from 50 stochastic realizations of each of 15 variogram parameter combinations. *MAE* values for mean and SD closing to zero indicate spatially representative samples. As shown in Figure 7, the scatterplot represents the degree of biases for 15 types of simulated fields. The final optimization results in the middle reaches of HRB exhibits good representativeness, and the maximum *MAE* is approximately 0.12. The representativeness of samples gradually increases with increasing the nugget and decreasing range. Intuitively, samples from fields with smaller nuggets and larger ranges should be more representative, but the results in Figure 7 are the opposite. This is because there are intensive observation nodes in Region A. Table 1 lists the *MAE*s of samples in Regions A and B. Large differences are observed between both regions. Regardless of the variability, samples in Region B with relatively uniform spatial distributions have low *MAE* values for different simulated fields. The *MAE* index is sensitive to the variability in regional variables when the spatial distribution of samples is excessively concentrated. Although cluster observations may lead to a

greater degree of deviation from the population, the local cluster points is necessary for estimating variogram parameters.

Figure 7. Scatter diagrams of the means (circles) and standard deviations (black dots). The *Y*- and *X*-axes represent the samples and the population, respectively.

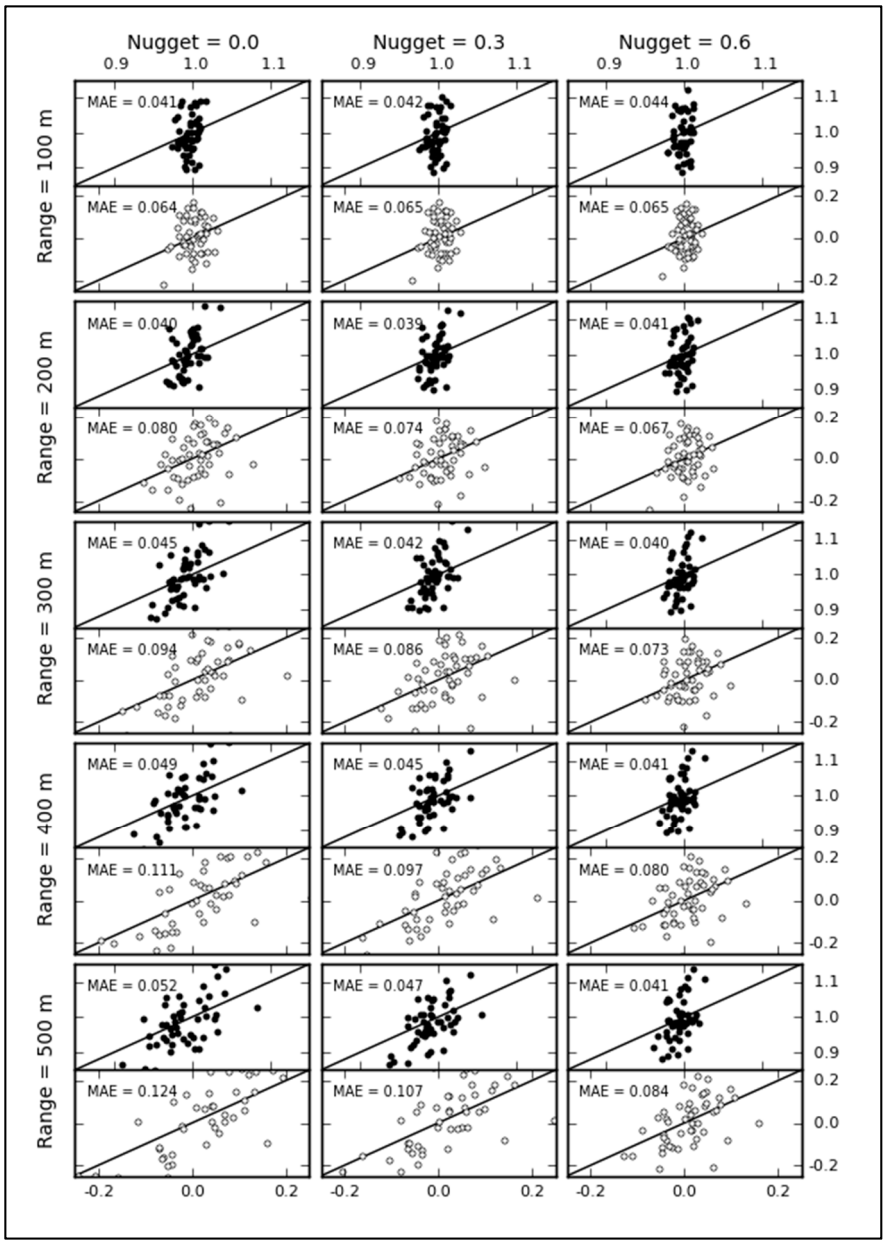


Table 1. Biases of the means and standard deviations between the samples and the population.

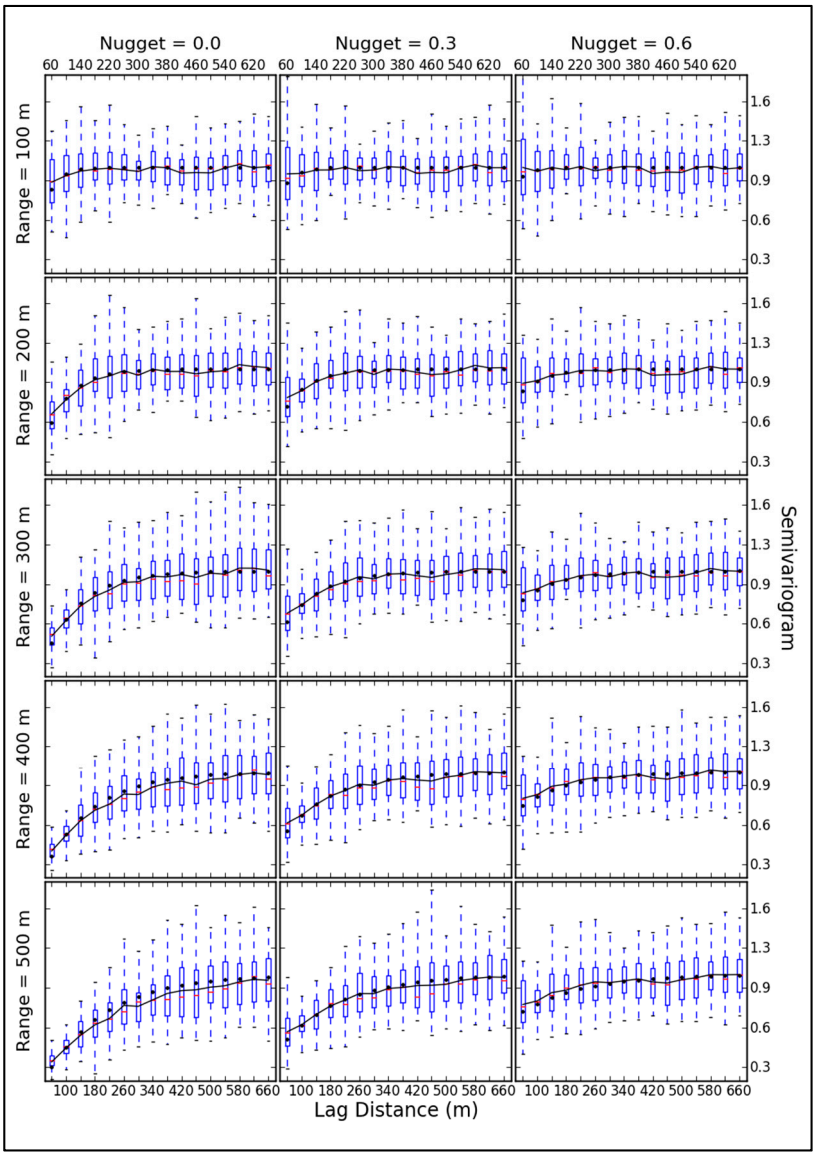
Range \ Nugget	0.0				0.3				0.6			
	Region A		Region B		Region A		Region B		Region A		Region B	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
100 m	0.138	0.059	0.071	0.055	0.130	0.063	0.074	0.055	0.121	0.068	0.076	0.055
200 m	0.212	0.079	0.063	0.054	0.185	0.066	0.067	0.053	0.151	0.062	0.072	0.053
300 m	0.271	0.105	0.060	0.053	0.231	0.083	0.065	0.051	0.179	0.065	0.071	0.051
400 m	0.325	0.133	0.058	0.052	0.272	0.101	0.064	0.050	0.201	0.072	0.070	0.050
500 m	0.368	0.160	0.056	0.050	0.312	0.116	0.063	0.049	0.218	0.076	0.069	0.050

4.3.2. Parameter Estimation

To capture the spatial variability, a specified number of paired points for 16 lag classes (Figure 6) are optimized to estimate the variogram. According to previous experience, at least 50 paired points are required [41]. Pairs of points in lag classes of 60 m, 100 m, 140 m, 180 m and 220 m are generated in the intensive Region A.

The dispersion and mean of the experimental variogram at different scales is described in Figure 8. With variations in the nugget and range, differences are generated in the dispersion at different scales. Small scales (60 m, 100 m and 140 m) especially exhibit large changes in dispersion. When the range increases, the dispersion of small scales decreases, because of the increased structure property, while at larger scales (>140 m), the dispersion increases slightly. At all estimated scales, the dispersion changes due to nugget variations are the opposite of the changes due to variations in range.

Figure 8. Experimental (black line) and theoretical (black dots) variograms. The boxplot represents the dispersion for the experimental variogram. The boxes with lines denote the median (thick line), lower quartile and upper quartile values (dotted lines).



The trend of the average line of the experimental variogram is close to the theoretical function, and it excellently expresses the spatial variability. When the structural characteristic is dominant (the proportion of the partial sill to sill is larger than that of the nugget to sill), the biases between theoretical and experimental variograms is low at small scales and high at large scales. However, with the growth of randomness, the biases at estimated scales are opposite.

To test the accuracy of estimated variogram parameters, the *RE* between the true variogram and the fitted curve is calculated (Table 2). For a dependence distance of 100 m, the estimated parameters are unavailable, because only one experimental variogram value in the distance of less than 100 m is used to fit the curve. With that exception, the relative error increases with increases in range when there is no non-spatial variance (Nugget = 0). When the major component of variance is non-spatial (Nugget = 0.6), the relative error of simulated fields decreases with increasing range. The maximum relative error is less than 6%, and the mean relative error is 3%.

Table 2. Accuracy of parameter estimations.

Nugget / Range	0.0				0.3				0.6			
	Nugget *	Sill *	Range *	*RE*	Nugget *	Sill *	Range *	*RE*	Nugget *	Sill *	Range *	*RE*
100 m	0.63	0.99	82.59	3.98%	0.88	0.99	181.61	3.59%	0.99	0.99	99.96	2.76%
200 m	0.31	1.00	262.33	2.54%	0.55	1.00	258.71	2.21%	0.75	0.99	220.66	1.86%
300 m	0.21	1.01	395.43	2.87%	0.46	1.00	368.75	1.95%	0.69	0.99	300.00	1.80%
400 m	0.16	1.00	516.82	4.58%	0.42	0.99	473.47	2.34%	0.69	0.99	422.18	1.69%
500 m	0.13	1.00	662.94	5.69%	0.40	0.99	589.32	2.71%	0.67	0.99	483.18	1.63%

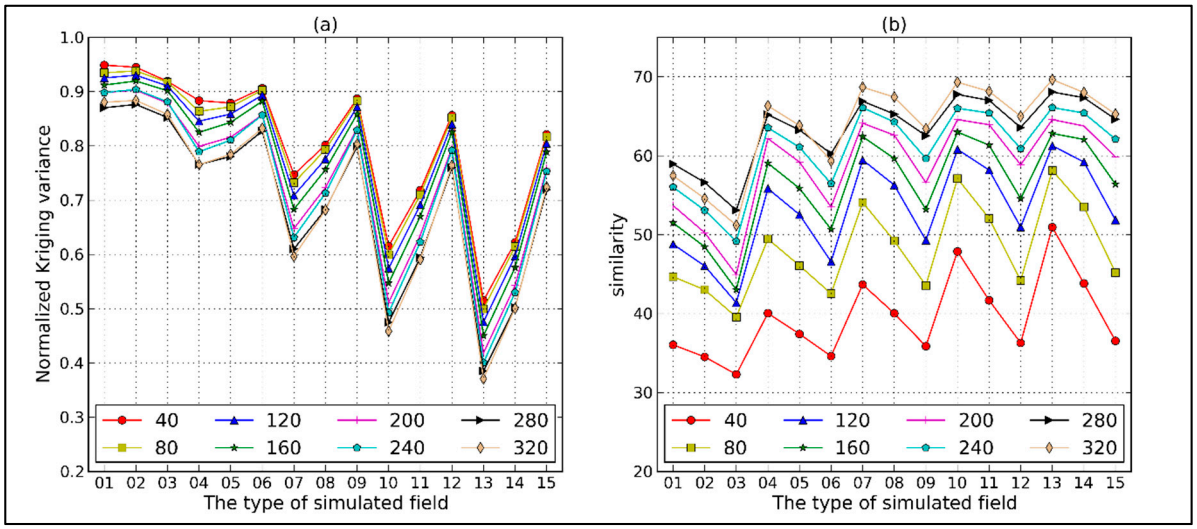
Note: * means the estimators of variogram parameters.

4.3.3. Spatial Prediction

After estimating the variogram, the interpolation accuracy also needs to be evaluated. Generally, the interpolation accuracy is evaluated by the kriging variance, which expresses the estimated uncertainty. We apply block kriging to obtain the estimated fields with eight grid sizes ranging from 40 m to 320 m. If the variogram parameters and estimated block sizes have been determined, the kriging variance is only related to the spatial distribution of samples and is independent of the sample values. As shown in Figure 9a, for each type of estimation field, the mean normalized kriging variances ($MBKV_{norm}$) in Equation (18) become small with increasing estimation scale. This is the spatial variance within a larger block that is cancelled out, leaving less uncertainty. For each estimation scale, the $MBKV_{norm}$ becomes large with decreasing dependence distance or increasing nugget value, which leads to the growth of uncertainty.

Kriging variance, however, cannot express the biases between true and estimated values. Therefore, we use another index to evaluate the estimation accuracy. The similarity between the simulated field and the estimated field with same block sizes is calculated by Equation (19). The similarity changes with estimation scales and parameters of variograms is opposite of the $MBKV_{norm}$ changes. As Figure 9b shows, the estimated field with the lowest randomness and maximum estimation scale (320 m) shows a maximum similarity of about 70% to the true field. In contrast, the estimated field with the strongest randomness and minimum estimation scale (40 m) shows a minimum similarity of about 32%.

Figure 9. Normalized kriging variance (**a**) and similarities (**b**) between the simulated and estimated fields with different block sizes.



5. Conclusions and Discussions

A hybrid sampling method is proposed to optimize samples for both spatial estimation and spatial interpolation when there is lack of prior information on the target variable. This hybrid method is used to optimally design the EHWSN in the HRB, and its effectiveness has been verified in terms of representativeness, parameter estimation and prediction accuracy, using various simulation fields.

The samples collected by the hybrid sampling method show the excellent representativeness. The relatively even spatial sampling in Region B enhances the representativeness of samples for different types of simulation fields. Though the nested samples in Region A introduce a slight sampling bias, it can improve the estimation accuracy of variogram parameters at small scales.

For variogram modeling, the stronger a regional variable shows spatial randomness, the more paired points are needed to capture the variability at small scales. When the structural features of a regional variable are obvious, more paired points are needed at larger scales. In this research, reliable prior information about the target variable is unavailable; therefore, using the equal treatment of paired points in each lag to estimate the variogram parameters is reasonable.

One of our objectives is to estimate ground truth at different remote sensing pixel scales. Both accurate parameter estimation and high sample representativeness are helpful for achieving this goal. The sampling design for estimation at block sizes exceeding 240 m has higher similarities. The differences of fluctuations between similarity curve lines with different block sizes indicates the influences of variogram parameters variation on the estimation accuracy. With the growth of the estimation scale, the amplitudes of fluctuations become small, which means that variogram parameters have less impact on the estimation accuracy. This information is meaningful, and if we estimate a quite large block, it is not necessary to be overly concerned with the estimation of variogram parameters. Instead, the sampling design may make more contributions to enhance the representativeness of samples.

For meeting multi-scale estimation requirements, the nested structure is designed. However, the cluster points may lead to some problems, such as decreasing the representativeness of the samples, enhancing the bias in the estimated variability at small scales and bringing a negative effect on the spatial

prediction. Ideally, the multi-cluster sampling, that points with a uniform spatial distribution are combined with multi-cluster points evenly distributed across the study area, can effectively eliminate these problems. However, such a sampling design needs to encompass a large number of samples. Due to budget limitations, only one point cluster is produced in our experiments. The validations prove that the nested sampling design is effective for both variogram modeling and spatial prediction based on limited samples.

In addition, unbiased sampling is important in the optimal design. In this work, the hybrid criterion considers both parameter estimation and spatial statistical inference. However, there is no quantitative expression to represent in the objective function. Therefore, future research should investigate how to quantify representation during the optimization process.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 91125001, the Chinese Academy of Sciences Action Plan for West Development Program Project under Grant KZCX2-XB3-15 and the National High-Tech Project under Grant 2012AA12A305.

Author Contributions

Jian Kang, Rui Jin and Xin Li conceived and designed the experiments; Jian Kang performed the experiments; Jian Kang and Rui Jin analyzed the data; Yong Ge, Jinfeng Wang and Jianghao Wang provided advice for improving the quality of this work; Jian Kang wrote the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zyl, T.V.; Simonis, I.; McFerren, G. The sensor web: Systems of sensor systems. *Int. J. Digit. Earth* **2009**, *2*, 16–30.
2. Hart, J.K.; Martinez, K. Environmental Sensor Networks: A revolution in the earth system science? *Earth Sci. Rev.* **2006**, *78*, 177–191.
3. Anderson, S.P.; Bales, R.C.; Duffy, C.J. Critical Zone Observatories: Building a network to advance interdisciplinary study of Earth surface processes. *Mineral. Mag.* **2008**, *72*, 7–10.
4. Kampe, T.U.; Johnson, B.R.; Kuester, M.; Keller, M. NEON: The first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. *J. Appl. Remote Sens.* **2010**, *4*, 043510–043524.
5. Zacharias, S.; Bogena, H.; Samaniego, L.; Mauder, M.; Fuss, R.; Putz, T.; Frenzel, M.; Schwank, M.; Baessler, C.; Butterbach-Bahl, K.; *et al*. A Network of Terrestrial Environmental Observatories in Germany. *Vadose Zone J.* **2011**, *10*, 955–973.
6. Waide, R.B.; Thomas, M.O. Long-Term Ecological Research Network. In *Earth System Monitoring*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 233–268.

7. Li, X.; Cheng, G.; Liu, S.; Xiao, Q.; Ma, M.; Jin, R.; Che, T.; Liu, Q.; Wang, W.; Qi, Y.; *et al*. Heihe Watershed Allied Telemetry Experimental Research (HiWATER): Scientific Objectives and Experimental Design. *Am. Meteorol. Soc.* **2013**, *94*, 1145–1160.
8. Jin, R.; Li, X.; Yan, B.; Li, X.; Luo, W.; Guo, J.; Ma, M.; Kang, J.; Zhu, Z.; Li, D. A Nested Eco-hydrological Wireless Sensor Network for Capturing the Surface Heterogeneity in the Midstream areas of the Heihe River Basin, China. *IEEE Geosci. Remote Sens.* **2014**, *11*, 2015–2019.
9. Haining, R. *Spatial Data Analysis: Theory and Practice*; Cambridge University Press: Cambridge, UK, 2003; pp. 93–96.
10. Hengl, T. *A Practical Guide to Geostatistical Mapping of Environmental Variables*; Office for Official Publications of the European Communities: Ispra, Italy, 2007; Volume 140, pp. 1–2.
11. Zhou, F.; Guo, H.-C.; Ho, Y.-S.; Wu, C.-Z. Scientometric analysis of geostatistics using multivariate methods. *Scientometrics* **2007**, *73*, 265–279.
12. Wang, J.-F.; Stein, A.; Gao, B.-B.; Ge, Y. A review of spatial sampling. *Spat. Stat.* **2012**, *2*, 1–14.
13. Stein, A.; Ettema, C. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agric. Ecosyst. Environ.* **2003**, *94*, 31–47.
14. Warrick, A.; Myers, D. Optimization of sampling locations for variogram calculations. *Water Resour. Res.* **1987**, *23*, 496–500.
15. Müller, W.; Zimmerman, D.L. Optimal Design for Variogram Estimation, 1997. Available online: http://epub.wu.ac.at/756/ (accessed on 13 October 2014).
16. Zhu, Z.; Stein, M.L. Spatial sampling design for parameter estimation of the covariance function. *J. Stat. Plan. Inference* **2005**, *134*, 583–603.
17. Van Groenigen, J.; Stein, A. Constrained optimization of spatial sampling using continuous simulated annealing. *J. Environ. Qual.* **1998**, *27*, 1078–1086.
18. Zimmerman, D.L.; Holland, D.M. Complementary co-kriging: Spatial prediction using data combined from several environmental monitoring networks. *Environmetrics* **2005**, *16*, 219–234.
19. Van Groenigen, J.W. Spatial Simulated Annealing for Optimizing Sampling. In *GeoENV I: Geostatistics for Environmental Applications. Lisbon*; Kluwer Academic Publishers: Lisbon, Portugal, 1997; pp. 351–361.
20. Banjevic, M.; Switzer, P. Optimal Network Designs in Spatial Statistics, 2004. Available online: http://searchworks.stanford.edu/view/5707036 (accessed on 13 October 2014).
21. Di Zio, S.; Fontanella, L.; Ippoliti, L. Optimal spatial sampling schemes for environmental surveys. *Environ. Ecol. Stat.* **2004**, *11*, 397–414.
22. Stevens, D.L., Jr. Spatial properties of design-based *versus* model-based approaches to environmental sampling. In Proceedings of Accuracy 2006: The 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon, Portugal, 5 July 2006; pp. 119–125.
23. Wang, J.F.; Christakos, G.; Hu, M.G. Modeling Spatial Means of Surfaces with Stratified Nonhomogeneity. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 4167–4174.
24. Hu, M.-G.; Wang, J.-F. A spatial sampling optimization package using MSN theory. *Environ. Modell. Softw.* **2011**, *26*, 546–548.

25. Simbahan, G.C.; Dobermann, A. Sampling optimization based on secondary information and its utilization in soil carbon mapping. *Geoderma* **2006**, *133*, 345–362.
26. Van Groenigen, J.W.; Siderius, W.; Stein, A. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* **1999**, *87*, 239–259.
27. Zimmerman, D.L. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **2006**, *17*, 635–652.
28. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266.
29. Matheron, G. *The Theory of Regionalized Variables and Its Applications*; École National Supérieure des Mines: Fontainbleau, France, 1971; Volume 5, p. 210.
30. Russo, D. Design of an optimal sampling network for estimating the variogram. *Soil Sci. Soc. Am. J.* **1984**, *48*, 708–716.
31. Yfantis, E.A.; Flatman, G.T.; Behar, J.V. Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Math. Geol.* **1987**, *19*, 183–205.
32. Fortune, S. Voronoi diagrams and Delaunay triangulations. *Comput. Euclidean Geom.* **1992**, *1*, 193–233.
33. Melles, S.; Heuvelink, G.B.; Twenhöfel, C.J.; van Dijk, A.; Hiemstra, P.H.; Baume, O.; Stöhlker, U. Optimizing the spatial pattern of networks for monitoring radioactive releases. *Comput. Geosci.* **2011**, *37*, 280–288.
34. Debba, P. Field Sampling Scheme Optimization Using Simulated Annealing. In *Simulated Annealing, Theory with Applications*; Sciyo: Rijeka, Croatia, 2010; pp. 113–129.
35. Brus, D.J.; de Gruijter, J.J. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* **1997**, *80*, 1–44.
36. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, doi.org/10.1063/1.1699114.
37. Borgman, L.; Taheri, M.; Hagan, R. Three-dimensional frequency-domain simulations of geological variables. *Geostat. Nat. Resour. Charact. Part* **1984**, *1*, 517–541.
38. Deutsch, C.; Journel, A. *Gslib: Geostatistical Software Library And Use's Guide*; Oxford University Press: New York, NY, USA, 1992; pp. 169–174.
39. Kennedy, W.J.; Gentle, J.E. *Statistical Computing*; Marcel Dekker: New York, NY, USA, 1980; pp. 90–95.
40. Sandholt, I.; Rasmussen, K.; Andersen, J. A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sens. Environ.* **2002**, *79*, 213–224.
41. Journel, A.G.; Huijbregts, C.J. *Mining Geostatistics*; Academic Press: London, UK, 1978; pp. 193–194.