Design of Environmental Sensor Networks Using Evolutionary Algorithms

Ferry Susanto, Setia Budi, Paulo de Souza Jr., Ulrich Engelke, and Jing He

***Abstract*—An evolutionary algorithm (EA)-assisted spatial sampling methodology is proposed to assist decision makers in sensor network (SN) deployments. We incorporated an interpolation technique with leave-one-out cross-validation (LOOCV) to assess the representativeness of a particular SN design. For the validation of our method, we utilized Tasmania's South Esk Hydrological Model developed by the Commonwealth Scientific and Industrial Research Organisation, which includes a range of environmental variables describing the landscape. We demonstrated that our proposed methodology is capable of assisting in the initial design of SN deployment. Ordinary Kriging is shown to be the best suited spatial interpolation algorithm for the EA's LOOCV under the current empirical study.**

***Index Terms*—Evolutionary algorithm (EA), inverse distance weighting (IDW), leave-one-out cross-validation (LOOCV), multiobjective, optimization, ordinary Kriging (OK), sensor network (SN) deployment, spatial data interpolation, spatial sampling, thin plate spline (TPS).**

I. Introduction

DEPLOYING sensor networks (SNs) from ground–up has never been a simple task without knowledge of historical environmental information within the landscape. The random distribution of nodes does not necessarily establish a fit-to-purpose network. While adding more sensor nodes within the region of interest (ROI) is likely to enhance the data usage and robustness of the SN, it would also introduce undesirable increase in both deployment and maintenance costs. Careful design is therefore a critical process prior to the deployment of SN. It is one of the most significant factors in ensuring that the network delivers fit-for-purpose data in a cost-effective way. Two fundamental questions arise while planning the deployment of SN: How many sensor nodes have to be deployed to meet certain application purposes and how should the nodes be deployed within the ROI [1].

The optimization of SN deployment is a process of determining the best possible locations of sensor nodes within the area under study (spatial sampling method). Mamun provided detailed descriptions of existing topologies in wireless sensor networks, including a comparative discussion of performance of different topologies [2]. As an overview, heuristic-based approaches (mathematical programming) have been extensively utilized to address such NP-hard problem [3], for instance, evolutionary algorithms (EAs) [4], [5], swarm algorithms [6], linear programming [7], spatial simulated annealing [8]–[11], and signal processing technique (e.g., wavelet [12]). However, the aforementioned literatures are mainly focusing on the following aspects: coverage of sensing area, network connectivity, and energy consumption. Apart from that, other factors such as cost, spatial analysis, and data reconstruction have been described and addressed in [13] and [14].

In this letter, we present a method to deploy an SN that is able to represent a certain environmental variable of the entire ROI, given a certain number of nodes. Section II describes in detail the problem to be addressed in this letter, followed by a methodology of the deployment strategy in Section III. Experimental simulation results and discussion are demonstrated in Section IV, and conclusions are drawn in Section V.

II. Problem Formulation

A. Assumption

The SN design in this letter is specifically tailored for weather stations acting as sensor nodes, which are stationarily deployed in the ROI. Each node is connected to a telemetry device that enables it to send data to a base station, and is also equipped with a solar panel as an energy source. Therefore, network connectivity and energy consumption are not of concern in this letter and are not considered for optimization.

B. Experimental Data Set

This study is conducted using Tasmania's South Esk Hydrological Model, developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) [15]. The model covers a set of environmental parameters in the North East of Tasmania (-41.0° to -42.0° latitude and 147.0° to 148.5° longitude). The ROI is mapped as a 2-D data grid of size of 151×101 and is in netCDF [16] format.

A total of one year of averaged daily data are utilized in our experiments. We focused on a number of parameters for

Manuscript received November 27, 2015; revised January 21, 2016; accepted February 1, 2016. Date of publication February 26, 2016; date of current version March 23, 2016.

F. Susanto is with the College of Engineering and Science, Victoria University, Footscray, Vic. 3011, Australia, and also with the Data61, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, Tas. 7005, Australia (e-mail: ferry.susanto@vu.edu.au; ferry.susanto@csiro.au).

S. Budi is with Data61, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, Tas. 7005, Australia, and also with the School of Engineering and Information and Communication Technology, University of Tasmania, Sandy Bay, Tas. 7005, Australia (e-mail: setia.budi@utas.edu.au; budi.budi@csiro.au).

P. de Souza Jr. and U. Engelke are with the Data61, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, Tas. 7005, Australia (e-mail: paulo.desouza@csiro.au; ulrich.engelke@csiro.au).

J. He is with the College of Engineering and Science, Victoria University, Footscray, Vic. 3011, Australia (e-mail: jing.he@vu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2016.2525980

different purposes in our study: 1) *elevation* data are mainly used for the sensor placement optimization purpose; such selection is justified based on the fact that such data can be accessed from a designated source, and Tang et al. also suggested that a high-elevation sample is crucial within the data set for meteorological studies [17]; and 2) environmental parameters such as *temperature*, *relative humidity*, and *solar radiation* are employed for the evaluation.

C. Problem Statement

In this letter, the quality of an SN is measured by the "degree of representativeness" toward the ROI in which the SN is deployed. We formulate the representativeness as how a certain placement of sensor nodes allows an interpolator to best estimate particular environmental variables across the ROI. To achieve a high representativeness of an SN design, we take the following considerations into account.

Spatial interpolation is criticized because of the incapability to estimate extreme values [17]. For example, within a mountainous landscape, if we merely deploy the sensors on the lower ground (i.e., foothills in this case), we will not be able to estimate the measurement of interest at the peak of the mountain. Considering such scenario, the focus of our sampling method is to discover the distribution of sensors (encompass those extreme values) to allow the best estimation/interpolation of particular environmental phenomena on the later stage. Furthermore, a good interpolator can also be evaluated by the extrapolation capability of distinct methods. Therefore, this issue will also be incorporated in the optimization criteria.

Toward this end, this letter aims to use EA to optimize the locations of N nodes that minimizes the error of the estimated (interpolated and extrapolated) surface under study.

III. Deployment Strategy

The aforementioned problem statement leads us to a multiobjective optimization problem, where it is typically impossible to have a single solution that satisfies all the objectives. Therefore, the focus is looking for a tradeoff among the objectives instead of looking for a single solution [18].

A. EA

We employ an EA [19] to address the multiobjective optimization problem in this letter. The EA mimics the process of natural selection principles to solve complex searching and optimization problems. The algorithm starts with a randomly generated population (a set of possible solutions), and it executes the reproduction process in each generation, including parent selection, crossover, and mutation. In this letter, each possible solution (individual) represents a single deployment of SN, which incorporates the position of sensor nodes within the ROI. Elitism is performed at the end of each generation to ensure that the best solution seen so far is not lost. Gradually, the most successful individuals evolve to discover the near-optimal solutions (Pareto front) [20].

We utilized EA from a Python library—Distributed Evolutionary Algorithms in Python [21]. Table I presents the parameters that we used in our experiment. These values were chosen from the literature [22]. The process is terminated when the Pareto front remains the same over the last 50 generations (the stopping criterion).

TABLE I
EA Parameters

Parameter	Value
Population size	50
Crossover probability	0.7
Mutation probability	0.05
Crossover operation	cxOnePoint
Mutation operation	mutUniformInt
Selection operation	NSGA2 [23]

B. Fitness Function

The EA discovers near-optimal solutions according to the so-called *fitness* functions that define the quality (SN representativeness) of a particular individual (SN design). Let $X = \{x_1, x_2, \ldots, x_N\}$ be a set of N sensor nodes deployed within the ROI. Based on the objectives of this letter (see Section II-C), we formulate the functions as follows.

1) Fitness Function 1: This function aims to identify the extreme values within the landscape. We leverage the main pitfall of the spatial interpolation technique in conjunction with the leave-one-out cross-validation (LOOCV) to assist us in identifying those nodes

$$\text{LOOCV}(\hat{f}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{f}^{(-n)}(x_n)\right)^2} \tag{1}$$

where $\hat{f}$ is a particular interpolation technique (see Section III-C), y_n is the observed value at the nth location, and $\hat{f}^{(-n)}(x_n)$ is the estimated value using $\hat{f}$ with the absence of the nth node (such that $X \setminus \{x_n\}$). Then, the fitness function is calculated by maximizing such equation, so that an interpolator is able to estimate a good representation of the ROI in the later stage. In other words, this method is trying to find a set of node locations in a way that each node is important and must be deployed within the network. The absence of any node will greatly degrade the representativeness of the area under study.

2) Fitness Function 2: We also want to consider the extrapolation capability of $\hat{f}$ by minimizing the root-mean-squared error (RMSE) of the estimated map's corners

$$\text{RMSE}(C) = \sqrt{\frac{1}{4}\sum_{n=1}^{4}\left(y_n - \hat{f}(x_n)\right)^2} \tag{2}$$

where C is the locations at the map's corners that are located at the top left, top right, bottom left, and bottom right of the surface.

C. Spatial Interpolation Techniques

This letter adopted three of the most frequentlyused spatial data interpolation techniques [24]: inverse distance weighting (IDW), ordinary Kriging (OK), and thin plate spline (TPS).

TABLE II
NUMBER OF REPLICATIONS (n) FOR EACH RESPECTIVE METHOD AND NUMBER OF NODES, CALCULATED USING (3)

No. of Nodes	5	10	15	20	25	30	35	40	45
OK	15	4	6	3	4	1	5	3	3
IDW	7	2	2	3	4	2	4	2	1
TPS	21	43	27	31	41	56	28	56	67

We are interested in observing how the application of different interpolation techniques on the proposed methodology will affect the performance. Brief descriptions of these methods are as follows.

1) IDW: Utilizes the spatial distance between the point to be interpolated and sample points as the main weighting mechanism [25]. This method has been extensively used because it is computationally efficient and produces acceptable results.

2) OK: A geostatistical method that incorporates the local spatial variances of its neighboring data points within the interpolation process [26]. Kriging-based techniques have been suggested as the optimal method overall from the literature [24], with the downside of being computationally heavy.

3) TPS: A spline-based technique for spatial data interpolation introduced by Duchon in 1976 that passes through each sample point [27]. It is based on the physical analogy involving the bending of a thin metal to create an interpolated surface.

IV. RESULTS AND DISCUSSIONS

In order to evaluate our work, we replicate the simulation using several different runs. While there is no empirical method to determine how many replications are needed, we utilized the method proposed in [28] as follows:

$$n = \left(\frac{z \times \sigma}{\mu \times acc} \right)^2 \tag{3}$$

where n is the required number of replications, z refers to the z-score of 1.96 which leads to 95% of confidence interval (CI), μ and σ are the mean and standard deviation obtained from preliminary simulations of 10 runs, and acc is the percentage of μ that we want to get as deviation (5% of accuracy in our case). The following sections are generated using replications based on Table II. We adopt 10 as the minimum number of replications in the case where $n < 10$.

A. Optimal Interpolator for EA's LOOCV

The first experiment is aimed to determine which interpolation method is best suited for the EA's LOOCV. Since a single run will generate a set of Pareto front (PF) that consists of a number of SN designs, we choose the solution that favors the first objective function (Fig. 1). In this experiment, we evaluate the techniques based on two criteria.

1) *Ability to estimate close-to-reality measurements.* RMSEs between the observed and the estimated values are calculated throughout the map, and the result is shown in Fig. 2(a).
2 *"Extrapolation" capability of each method.* We calculated the RMSE at the map's corner. The result is shown in Fig. 2(b).

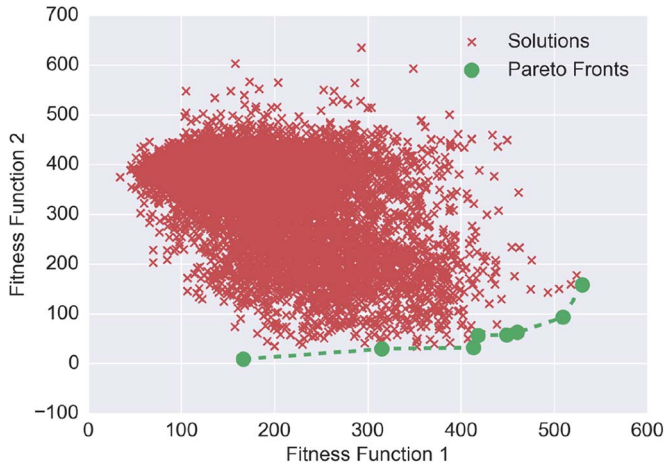

Fig. 1. Typical plotting of multiobjective EA with two objective functions: (a) x-axis, maximization of (1); and (b) y-axis, minimization of (2). The red markers are the explored solutions throughout the EA process, and the green dots are the Pareto front (a list of nondominated solutions).

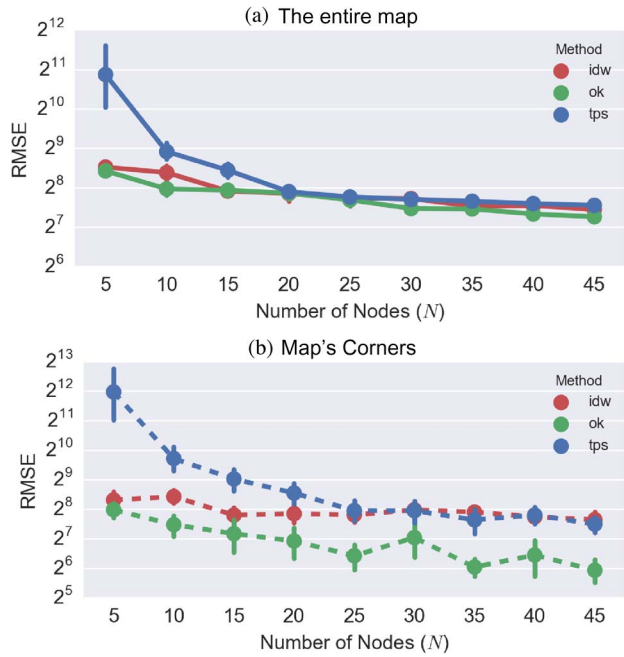

Fig. 2. Interpolation and extrapolation performance of different interpolation techniques. The x-axis represents the number of nodes (N), and the y-axis is the RMSE value (the average and 95% CI). The RMSE is the error calculation between the observed and the estimated surface height data. (a) Throughout the landscape and (b) the map's corner.

Fig. 2(a) presents the performance among three compared spatial interpolation techniques: OK, IDW, and TPS. The result shows that OK performs the best (lowest RMSE), and it is followed by IDW and TPS. The CIs of OK and IDW are relatively small and are barely noticeable, which indicates that both methods produce relatively stable results. TPS, on the other hand, has a very large CI with the number of nodes (N) being 5, which becomes less significant as N increases. The performances among these methods progress to converge as $N = 20$; further increase in N does not substantially reduce the RMSE.

According to Fig. 2(b), the extrapolating capabilities of the compared techniques have the similar performance behavior as in Fig. 2(a) (OK, followed by IDW and TPS). Interestingly,

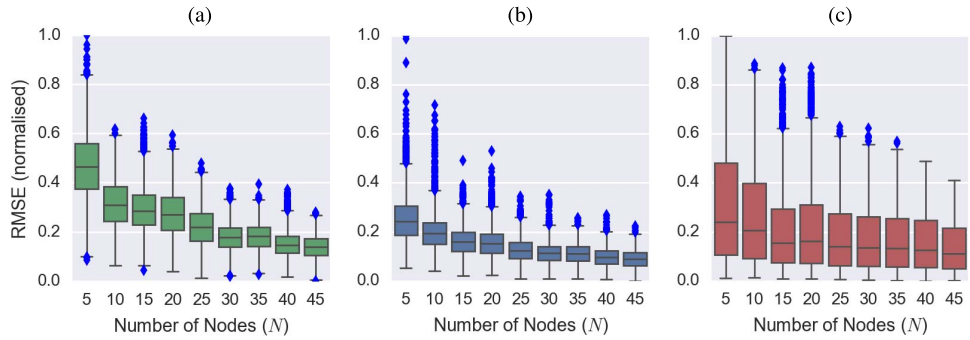

Fig. 3. Performance comparisons among different parameters. (a) Air temperature, (b) relative humidity, and (c) solar radiation. The x-axis is the number of nodes, and the y-axis is the normalized RMSE.

the extrapolating performance between IDW and TPS becomes comparable starting from $N = 25$, where IDW performs relatively stable regardless of N. The fact that OK outperforms other techniques is not surprising, considering that OK incorporates geostatistical analysis within its calculation.

These results have shown that OK suits the best for our proposed method. Thus, the subsequent simulation will utilize OK as the interpolation technique within EA's LOOCV.

B. Performance Assessment

This simulation is used to validate the proposed method by comparing the interpolation RMSE generated from the sampling design of different environmental parameters: temperature, relative humidity, and solar radiation. The main objective of this experiment is to determine whether using surface height data within the optimization (spatial sampling design) is able to generate a close-to-reality measurement of other parameters within the ROI.

Fig. 3 demonstrates the simulation results. In general, it shows that the RMSE decreases for each parameter as the number of nodes N increases. However, it is noticeable that air temperature has the most significant quality improvement as N increases [see Fig. 3(a)], as reflected in median, interquartile range, whiskers, and outliers. It is then followed by relative humidity [see Fig. 3(b)] that behaves similarly with temperature (median), while the only discrepancy is the less significant improvement in terms of whiskers and outliers. Finally, for the case of solar radiation [see Fig. 3(c)], a slight improvement can be observed for the median, but not the variability, particularly whiskers and outliers (e.g., $N = 20$).

As a result, under the current empirical study, we suggest that using elevation in the sampling design suffices to obtain confident temperature and relative humidity data. In a less restricted budget situation, the best representativeness of all the variables could be achieved by employing 35 nodes (e.g., Fig. 4).

V. Conclusion

The main objective of this letter is to obtain near-optimal sensor node placements that allow any *interpolator* to best estimate a particular environmental phenomenon of interest. A novel spatial sampling design approach is proposed by using multiobjective EA to minimize the prediction error (see Section III-B).

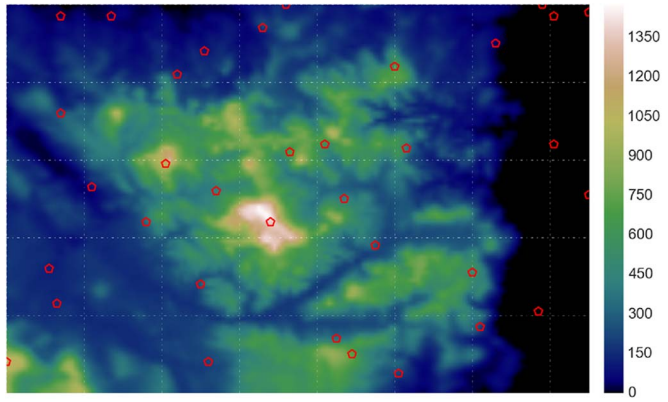

Fig. 4. Example of an SN deployment (35 sensor nodes) within the ROI, generated by the proposed methodology. This figure demonstrates the elevation map from the South Esk model, and the red dots are the node locations to be deployed within the landscape.

In order to define which interpolation technique is suitable for our work, three different spatial interpolation techniques are compared. Our results indicate that using OK produces the best results, which yields the lowest interpolation error and best extrapolating capability (see Section IV-A).

This letter provides a simulation which would help decision makers while designing an SN that could deliver fit-for-purpose data without introducing undesirable costs resulting from the excessive placement of nodes. Based on the outputs (Pareto optimality solutions) generated by the proposed method (see Fig. 1), the decision makers are required to select a single SN design that incorporates additional consideration of their domain knowledge (i.e., sparsity and the feasibility of the deployment) that is suitable for their purpose.

Acknowledgment

The first author would like to thank Victoria University and the Vale Institute of Technology for the Ph.D. scholarship and CSIRO for the Office of the Chief Executive (OCE) top-up scholarship. The second author would like to thank Sense-T for the Ph.D. scholarship and CSIRO for the top-up scholarship. The authors would also like to thank the anonymous reviewers who have given excellent comments that greatly improved the contents of this letter.

REFERENCES

[1] E. Onur, C. Ersoy, and H. Deliç, "How many sensors for an acceptable breach detection probability?" *Comput. Commun.*, vol. 29, no. 2, pp. 173–182, Jan. 2006.

[2] Q. Mamun, "A qualitative comparison of different logical topologies for wireless sensor networks," *Sensors*, vol. 12, no. 11, pp. 14 887–14 913, Nov. 2012.

[3] M. Younis and K. Akkaya, "Strategies and techniques for node placement in wireless sensor networks: A survey," *Ad Hoc Netw.*, vol. 6, no. 4, pp. 621–655, Jun. 2008.

[4] M. Mansouri, H. Nounou, and M. Nounou, "Genetic algorithm-based adaptive optimization for target tracking in wireless sensor networks," *J. Signal Process. Syst.*, vol. 74, no. 2, pp. 189–202, Jun. 2013.

[5] S. Budi, P. de Souza, G. Timms, V. Malhotra, and P. Turner, "Optimisation in the design of environmental sensor networks with robustness consideration," *Sensors*, vol. 15, no. 12, p. 29 765, 2015.

[6] R. Kulkarni and G. Venayagamoorthy, "Particle swarm optimization in wireless-sensor networks: A brief survey," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 41, no. 2, pp. 262–267, Mar. 2011.

[7] M. Rebai, M. Le berre, H. Snoussi, F. Hnaien, and L. Khoukhi, "Sensor deployment optimization methods to achieve both coverage and connectivity in wireless sensor networks," *Comput. Oper. Res.*, vol. 59, pp. 11–21, Jul. 2015.

[8] W. Jianghao, G. Yong, G. B. M. Heuvelink, and Z. Chenghu, "Spatial sampling design for estimating regional GPP with spatial heterogeneities," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 539–543, Feb. 2014.

[9] G. B. M. Heuvelink, Z. Jiang, S. De Bruin, and C. J. W. Twenhöfel, "Optimization of mobile radioactivity monitoring networks," *Int. J. Geograph. Inf. Sci.*, vol. 24, no. 3, pp. 365–382, 2010.

[10] Y. Ge *et al.*, "Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the Babao River basin, China," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 1, pp. 92–110, 2015.

[11] J. Kang *et al.*, "Hybrid optimal design of the eco-hydrological wireless sensor network in the middle reach of the Heihe River basin, China," *Sensors (Basel)*, vol. 14, no. 10, pp. 19 095–19 114, 2014.

[12] A. L. L. Aquino, R. A. R. Oliveira, and E. F. Wanner, "A wavelet-based sampling algorithm for wireless sensor networks applications," in *Proc. Symp. Appl. Comput.*, 2010, pp. 1604–1608.

[13] A. Frery, H. S. Ramos, J. Alencar-Neto, E. Nakamura, and A. A. F. Loureiro, "Data driven performance evaluation of wireless sensor networks," *Sensors*, vol. 10, no. 3, p. 2150, 2010.

[14] C. D'Este, P. d. Souza, C. Sharman, and S. Allen, "Relocatable, automated cost-benefit analysis for marine sensor network design," *Sensors*, vol. 12, no. 3, pp. 2874–2898, Mar. 2012.

[15] J. Katzfey and M. Thatcher, "Ensemble one-kilometre forecasts for the South Esk hydrological sensor web," in *Proc. 19th Int. Congr. modsim*, 2011, pp. 3511–3517.

[16] R. Rew and G. Davis, "NetCDF: An interface for scientific data access," *IEEE Comput. Graph. Appl.*, vol. 10, no. 4, pp. 76–82, Jul. 1990.

[17] L. Tang, X. Su, G. Shao, H. Zhang, and J. Zhao, "A clustering-assisted regression (CAR) approach for developing spatial climate data sets in China," *Environmental Modelling Softw.*, vol. 38, pp. 122–128, 2012.

[18] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, 1st ed. New York, NY, USA: Wiley, Mar. 2009.

[19] C. C. Coello, G. B. Lamont, and D. A. V. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed. New York, NY, USA: Springer-Verlag, Sep. 2007.

[20] Y. Censor, "Pareto optimality in multiobjective problems," *Appl. Math. Optim.*, vol. 4, no. 1, pp. 41–59, Mar. 1977.

[21] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2171–2175, Jul. 2012.

[22] M. Srinivas and L. Patnaik, "Genetic algorithms: A survey," *Computer*, vol. 27, no. 6, pp. 17–26, Jun. 1994.

[23] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[24] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecol. Informat.*, vol. 6, no. 3/4, pp. 228–241, 2011.

[25] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proc. 23rd ACM Nat. Conf.*, 1968, pp. 517–524.

[26] P. A. Burrough and R. A. McDonnell, Principles of geographical information systems. Oxford, U.K.: Oxford Univ. Press, 1998.

[27] J. Duchon, *Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces*, ser. Lecture Notes in Mathematics, vol. 571. Berlin, Germany: Springer-Verlag, 1977, Sect. 7, pp. 85–100.

[28] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. New York, NY, USA: Wiley, 1991.