scientific reports

Check for updates

OPEN

Regulation-based probabilistic substance quality index and automated geo-spatial modeling for water quality assessment

Artyom Nikitin[1]✉, Polina Tregubova[2], Dmitrii Shadrin[2], Sergey Matveev[3,4], Ivan Oseledets[1,4] & Maria Pukalchik[2]

Natural environments are recognized as complex heterogeneous structures thus requiring numerous multi-scale observations to yield a comprehensive description. To monitor the current state and identify negative impacts of human activity, fast and precise instruments are in urgent need. This work provides an automated approach to the assessment of spatial variability of water quality using guideline values on the example of 1526 water samples comprising 21 parameters at 448 unique locations across the New Moscow region (Russia). We apply multi-task Gaussian process regression (GPR) to model the measured water properties across the territory, considering not only the spatial but inter-parameter correlations. GPR is enhanced with a Spectral Mixture Kernel to facilitate a hyper-parameter selection and optimization. We use a 5-fold cross-validation scheme along with R^2-score to validate the results and select the best model for simultaneous prediction of water properties across the area. Finally, we develop a novel Probabilistic Substance Quality Index (PSQI) that combines probabilistic model predictions with the regulatory standards on the example of the epidemiological rules and hygienic regulations established in Russia. Moreover, we provide an interactive map of experimental results at 100 m^2 resolution. The proposed approach contributes significantly to the development of flexible tools in environment quality monitoring, being scalable to different standard systems, number of observation points, and region of interest. It has a strong potential for adaption to environmental and policy changes and non-unified assessment conditions, and may be integrated into support-decision systems for the rapid estimation of water quality spatial distribution.

Freshwater—probably the most precious resource on the planet—plays a crucial role for humans since it is exploited in farming, industry, domestic consumption, and power supply[1–4]. Yet, in the light of drastically changing environmental conditions freshwater resources are highly vulnerable. They are affected both by natural climatic shifts as well as by anthropogenic impact manifested in pollution and catchment disturbance. To enhance freshwater storage protection active monitoring and quality assessment are required.

A freshwater quality assessment is complicated at both spatial and temporal scales and in terms of data collection. In other words, numerous points of observation are needed, some flows are partly hidden or even unavailable for the observers without specific equipment[5]. Another bottleneck for assessment is the high complexity and heterogeneity of water composition. It consists of a number of parameters of distinctive nature-physicochemical (such as acidity, alkalinity, turbidity, the content of cations and anions, including toxic chemicals, such as pesticides and toxic trace metals) and biological (presence and structure of living organisms' community). These characteristics are highly interconnected with each other and sensitive to the external stressors and processes at

[1]Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow, Russian Federation 121205. [2]Digital Agriculture Laboratory, Skolkovo Institute of Science and Technology, Moscow, Russian Federation 121205. [3]Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russian Federation 119991. [4]Marchuk Institute of Numerical Mathematics of Russian Academy of Science, Moscow, Russian Federation 119333. ✉email: artem.nikitin@skolkovotech.ru

the same time. Multiple factors, ranging from natural to anthropogenic ones, determine the significant spatial variability of water characteristics and the overall quality on large territories. The former includes aquifer characteristics heterogeneity, substance migration patterns, while the latter—diverse sources of potential pollution from different land-use types in urbanised and developed lands[6–8].

In order to monitor complex natural systems, such as freshwater reservoirs, an investigator has to answer two key questions: (1) how to contemplate as much information as possible in a most conscientious way and (2) how to cover maximal territory using the available data, which tends to be quite limited. A wide-spread approach to tackle the first problem is to evaluate the water system state by reducing the overall complexity, e.g. by calculating one integrative parameter, such as Water Quality Index (WQI). The idea of introducing a single aggregated parameter, such as WQI, was firstly proposed by Horton, 1965[9]. It has been significantly elaborated since then[10–13], being used even by some governmental agencies, such as National Sanitation Foundation Water Quality Index (NSFWQI), Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI), British Columbia Water Quality Index (BCWQI)[14,15]. The quality index approach is widespread in assessing other complex natural environments, e.g. soil[16–18]. The main objective of classic WQI is the aggregation of multiscale data, based on the relative importance of parameters, and further categorisation according to the obtained results. However, the applicability of WQI raised a number of concerns as it lacks unity and coherence in estimation workflow and evaluating the parameters.

As a consequence, these factors led to a high divergence in interpreting the obtained results[15,19–22]. The existing aggregation outlooks rarely reflect the normative thresholds directly[23], and often miss other than "less is better" possible motivations for parameters' consideration. Thus, the cases of the optimum range, when the permissible parameter content is defined by some lower and upper bounds, are underrepresented. Significant part of the recent developments focuses on to the approaches of picking up the most important features to construct the index from them via assigning different weights[24–26] in the contrast to subjective recommendations[27], or systematizing them, using such tools as Multi-Criteria Decision Analysis, Analytic Hierarchy Process, Fuzzy Logic[10,14,28]. In case of implementing the expert opinion systematization techniques the authors identified several significant uncertainties accompanying a non-stable data aggregation process and a high risk of misinterpretation. Some of the new approaches are based on implementing numerical tools to consider the overall variability of characteristics across the territory of study, e.g. use of the Principal Component Analysis (PCA). However, if the high variance reflects noticeable parameter changes from an excellent to an appalling state, low variance does not allow to distinguish whether conditions are very poor or not. Finally, one may doubt whether it is expedient to aggregate the information into one index value at all after measuring tens of parameters. Specifically, considering that monitoring observations and private assessments are already based on the plethora of different characteristics[29]. Thus, the development of the new unified (i.e. non-specific to study sites) approaches to quality assessment are needed[30,31].

In terms of observation and monitoring water quality spatial dynamics, data imputation, and prediction of possible system shifts, modeling approaches are in common use. Among them, two modeling approaches can be distinguished: process-based (PB) and data-driven solutions. Classic PB solutions are widely used for the tasks such as description of transport and fate of contaminants in water flows[32], recharge-depletion and consumption dynamics[33]. These approaches, built on structural equations, are connected with the description of stochastic processes underlying visible outputs with specified initial and boundary conditions[34]. Despite being comprehensive and fundamental, i.e. based on the observed dependencies in exploratory researches, the PB solutions are often considered as too complicated. Such models are usually limited by the demand of complex explanatory infrastructure related to various natural environments to describe the principles behind the processes; up-scaling challenges, slow and clumsy calculations, as well as biases caused by the established assumptions and conditions[34–36].

A suitable solution in the environmental modeling and assessment is using the data-driven modeling solutions, to be more specific, machine learning (ML) to supply and improve PB techniques and as a self-contained approach. In the last few decades the popularity of ML-based approaches used for the modeling the water characteristics' distribution, including over-all water quality, has been increasing. The ML techniques have been successfully introduced in the evaluation of the most important aspects of freshwater reservoirs, e.g. surface water quality and its mapping[37], determining the key parameters for accurate quality estimation[38], predicting groundwater contamination[39,40] and level dynamics[41]. Although ML approaches require relatively large training sets and leave behind the physical mechanisms of processes, combined with geostatistical techniques, they allow to establish the distribution of characteristics more precisely and in higher resolution on both spatial and temporal scales. As compared to the PB modeling techniques, the ML approaches allow to model complex nonlinear relationships between independent and target parameters using black-box approaches[42,43]. Thus, there is no need to rely on any empirical models that are not always able to embrace all aspects of the considered system. This in turn opens an avenue for geo-spatial modeling automatization.

One of the most popular tools for successful predicton of the spatial distribution of parameters related to the natural environment (e.g. water quality, groundwater level, soil organic matter, air pollutant) is the implementation of Gaussian Process (GP)[44–47]. GP is a kernel-based model able to handle different types of input data without any limitations of the particular parametric form of relation to the output. The flexibility of GP allows to use it for the most frequent tasks in environmental studies such as regression (also called kriging) or classification[48] with the ability to perform simultaneous multi-parameter predictions giving confidence for the predicted values. Apart from the advantages of GP, a list of weaknesses is normally mentioned: limited computational efficiency with the growing number of samples, poor scaling with increasing data dimensionality. Additionally, it requires to choose variogram and mean (trend) functions structure and make some assumptions about the data distribution type (e.g., normality)[49]. To handle the efficiency problems several approaches may be exploited, including batch-learning[50], kernel approximation[51] and dimensionality reduction techniques[52]. Therefore, considering a high popularity of GP in the environmental science community, the studies applying GP to environmental issues

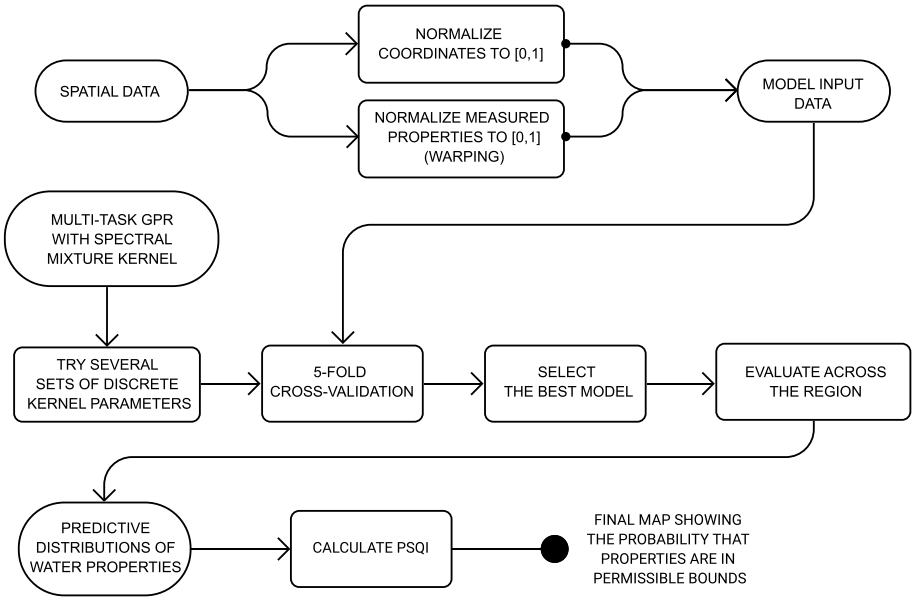


Figure 1. Main steps of the workflow: from data pre-processing to final Probabilistic Substance Quality index mapping.

and showing the ways to decrease handcrafting (i.e., through the automated kernel structure selection) and increase computational efficiency are of paramount interest and practical importance. It should be noted, that there are other ML tools potentially capable of obtaining similar results, e.g. Support Vector Regression[53], neural networks[54–56].These models may give reasonable results, however black-box approaches are usually reported to be difficult in interpretation. At the same time, GP benefits over the above mentioned approaches due to the results of GP applications are usually easier interpreted. Still, the comparison between different modeling frameworks to solve the multi-task problems might be the promising direction in advancing assessment approaches.

This paper presents a part of the project aimed to implement the ML techniques to the environmental monitoring issues. Considering the above-mentioned developments of the community, the objective of this research is to show an automated Geographical Information Systems (GIS) approach for the freshwater assessment and spatial modeling applied to the existing sample network based on the data including 1526 samples obtained from 448 unique points across the New Moscow region described by 21 parameters and spatial coordinates[57].

The detected concentrations of parameters used in the modeling vary significantly across the territory. Some of them, e.g. Cl, NO_3, PO_4 ions, as well as metal ions, Fe, Mn, Ni, may exceed established permissible limits 2-8 times while their content in other locations may be equal to 0. Our proposed modeling workflow is based on the multi-task Gaussian process regression (GPR) featured by the automatic kernel structure selection and hyper-parameter optimization. An important advantage of the developed approach is that it enables predicting the spatial distribution of all of the measured properties in one consistent procedure. Particularly, it considers both spatial and inter-parameter dependencies and allows to assess not only the precise values of water properties but also their probabilistic ranges as well as enables the accuracy control with the minimal user efforts. This modeling framework is supplied with a limit-driven assessment system: one can easily check whether the selected parameter is in the permissible range (considering the diapason, not only the upper limit) in the selected location. Considering the convenience of the joint concise characteristic to describe the overall system state, we propose a probabilistic substance quality index (PSQI). It incorporates both observations and the established regulations, denoting the probability that all of the characteristics lie in permissible diapasons. The presented approach is devised to meet the requirements to enhance reproducibility and fairness of the assessment. It is open to scaling to different standard systems, set of points of observation, region of interest and has a strong potential for adaption to environmental and policy changes and non-unified conditions of assessment. Therefore, it ensures direct integration into support-decision systems.

Modeling tools and water quality index

In the following section the key concepts and stages of the proposed assessment approach are described in detail. Firstly, the theory behind the Gaussian process regression is explained and the modeling objectives are formulated. In order to deal with one of the key difficulties of the natural systems quality assessment—representative consideration of the overall complexity in spatial modeling—the concept of multi-task Gaussian process regression is given. It is supplied with an explanation of the automatization procedure (hyper-parameter selection) and details for reproducibility: data pre-processing, model validation, and technical requirements to handle calculations. Finally, the definition of the proposed Probabilistic Substance Quality Index is given. The general scheme of the workflow is presented in Fig. 1.

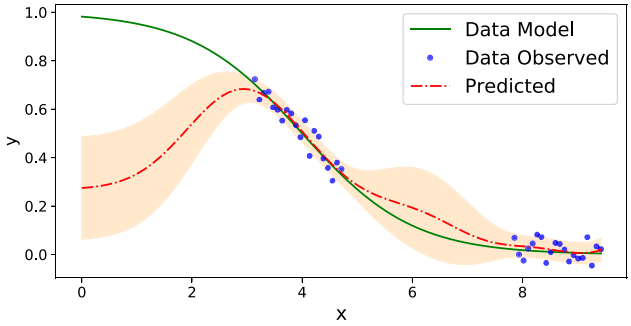


Figure 2. Example of Gaussian process regression (red dashed line stands for the predictive mean and orange fill stands for the standard deviation intervals) with noisy measurements (blue dots) of a sigmoid function (solid green line) using Gaussian kernel and constant mean function.

Gaussian process regression. In order to perform geo-spatial modeling of multiple water properties from the collected dataset, we refer to the *Gaussian process regression* (GPR) framework[49], known as *kriging* in geostatistics. *Mean* $\mu(\cdot)$ and *covariance* (or *kernel*) $k(\cdot,\cdot)$ functions completely determine a Gaussian process:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}\left(\mu(\mathbf{x}), k(\mathbf{x},\mathbf{x}')\right), \\ \mu(\mathbf{x}) &= \mathbb{E} f(\mathbf{x}), \\ k^x(\mathbf{x},\mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x})-\mu(\mathbf{x}))(f(\mathbf{x}')-\mu(\mathbf{x}'))\right], \end{aligned} \tag{1}$$

where $\mathbb{E}$ is a mathematical expectation and $\mathbf{x} \in \mathbb{R}^d$ is a vector of d input parameters, which are 2D coordinates in our case (for instance, represented in the Mercator projection). As an example, consider a simple GP model:

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0,\sigma^2)$ accounts for noise in measurements, hence, helping to avoid model over-fitting. Given the training samples $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^\intercal \in \mathbb{R}^{N\times d}$, $\mathbf{Y} = \left(y_1,\ldots,y_N\right)^\intercal \in \mathbb{R}^N$, where N denotes the number of available measurements and $(\cdot)^\intercal$ denotes a transpose, the predictive distribution at arbitrary point $\mathbf{x}_*$ can be found as

$$\begin{aligned} \hat{f}(\mathbf{x}_*) &\sim \mathcal{N}(\hat{\mu},\hat{\sigma}^2), \\ \hat{\mu}(\mathbf{x}_*) &= \mu(\mathbf{x}_*) + k^x_* \Sigma(\mathbf{y}-\mu(\mathbf{X})), \\ \hat{\sigma}^2(\mathbf{x}_*) &= k(\mathbf{x}_*,\mathbf{x}_*) - (k^x_*)^T \Sigma^{-1} k^x_*, \\ \Sigma &= K^x + \sigma^2 I, \end{aligned} \tag{3}$$

where $K^x = k^x(\mathbf{X},\mathbf{X}) = k(\mathbf{x}_i,\mathbf{x}_j), i,j = 1,\ldots,N$ and $k^x_* = k^x(\mathbf{X},\mathbf{x}_*)$ are spatial covariance matrices between all of the training points and between training points and the single prediction point, respectively; $\mu(\mathbf{X}) = \mu(\mathbf{x}_i), i = 1,\ldots,N$ is the mean vector-function evaluated at the training points; and I is an identity matrix. A choice of the mean and kernel functions depends on the assumptions about the model and the particular application. An example of a kernel function is a widely used *Gaussian* kernel, which corresponds to Gaussian variogram in kriging. The kernel hyper-parameters are usually optimized using *Maximum Likelihood Estimation* (MLE)[58] or its variations.

Figure 2 illustrates an example of GPR using the Gaussian kernel and the constant mean over the sigmoid function with noisy measurements. Predictive variance increases notably at the points with missing measurements. Moreover, outside of the interpolation region, a predictive mean fails to capture the behavior of the underlying model due to the structure of its mean and kernel functions.

There are several issues that should be addressed to perform efficient modeling:

- Basic approach allows modeling only a single output or multiple independent outputs, whereas our aim is to capture both geo-spatial and inter-feature dependencies at once.
- Naive GPR computational requirements increase cubically with the dataset size, as it requires matrix inversion.
- GPR model requires selection of multiple hyper-parameters, e.g., kernel and mean functions.

Multi-task Gaussian process regression. Let's consider a more complex model than in the Eq. (2):

$$\mathbf{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \epsilon, \tag{4}$$

where $\mathbf{y}$ is a vector of M measured properties, $\epsilon \sim \mathcal{N}(0, D)$ with D being an $M \times M$ diagonal noise matrix. In order to capture both inter-feature and geo-spatial dependencies in covariance function construction, we refer to multi-task approach[59]:

$$k_{kl}(\mathbf{x},\mathbf{x}') = \langle f_k(\mathbf{x}), f_l(\mathbf{x}')\rangle = K^f_{kl}k^x(\mathbf{x},\mathbf{x}'),\ k,l = 1,\ldots,M \tag{5}$$

where K^f is an $M \times M$ inter-feature covariance matrix and $\langle\cdot,\cdot\rangle$ denotes a scalar product. Then, given the training data $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)^\intercal \in \mathbb{R}^{N\times d}$, $\mathbf{Y} = \left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)^\intercal \in \mathbb{R}^{N\times M}$, the predictive distribution at the unobserved point $\mathbf{x}_*$ is found as

$$\begin{aligned}
\hat{f}(\mathbf{x}_*) &\sim \mathcal{N}(\hat{\mu},\hat{\Sigma}),\\
\hat{\mu}(\mathbf{x}_*) &= \mu(\mathbf{x}_*) + (K^f \otimes k^x_*)^T\Sigma^{-1}(\overline{\mathbf{y}} - \overline{\mu}(\mathbf{X})),\\
\hat{\Sigma}(\mathbf{x}_*) &= K^f k^x(\mathbf{x}_*,\mathbf{x}_*) - (K^f \otimes k^x_*)^T\Sigma^{-1}(K^f \otimes k^x_*),\\
\Sigma &= K^f \otimes K^x + D \otimes I,
\end{aligned} \tag{6}$$

where $\otimes$ denotes a Kronecker product, $\overline{\mathbf{y}}$ and $\overline{\mu}$ are flattened $N \cdot M$-dimensional vectors obtained from $N \times M$ matrices $\mathbf{y}$ and $\mu(\mathbf{X})$, respectively. As a "side-effect" of this approach, after the model is built and hyper-parameters are optimized, we can analyze the dependencies between modeled properties using matrix K^f.

Decreasing computational complexity. One of the main disadvantages of the proposed multi-task GPR approach is that it induces a lot of additional calculations. In the naive case it becomes $O(N^3M^3)$ instead of $O(N^3)$, and to alleviate it, we turn to GPU computations. We perform modeling using Python programming language, GPyTorch library[60] based on the PyTorch framework[61] and NVIDIA Tesla K80 GPU. However, to further improve performance and to easily account for the possibly correlated components, we parametrize covariance matrix K^f using low-rank approximation as follows:

$$K^f = B^TB + \mathrm{diag}(\mathbf{v}), \tag{7}$$

where B is an $M \times r$ matrix, $\mathbf{v}$ is an M-dimensional positive vector and r is the supposed rank of K^f.

Hyper-parameter selection. In order to simplify the hyper-parameter selection, we refer to *Spectral Mixture Kernel*[62], which can be represented as:

$$k^x(\mathbf{x},\mathbf{x}') = k^x(\mathbf{x}-\mathbf{x}') = k^x(\tau) = \sum_{q=1}^{Q} w_q \prod_{p=1}^{d} \exp\{-2\pi^2\tau_p^2 v_p^{(q)}\}\cos\left(2\pi\tau_p\mu_p^{(q)}\right), \tag{8}$$

where Q is the number of components in the mixture, w_q is a weight of the q^{th} component, $v_p^{(q)}$ and $\mu_p^{(q)}$ are p^{th} variance and mean of the q^{th} mixture component, respectively. The weights influence the importance of each separate component in the mixture, whereas variance and the mean allow to model effects of different scales. Hence, we are able to model arbitrary stationary kernels and control the complexity with a number of components Q in the mixture. Depending on the size and structure of the input data, the number of model parameters may require certain tuning. The main advantage of this approach is that it does not require any handcrafting of the potentially effective kernels, but instead, enables automatic hyper-parameter selection and optimization. Since the kernel assumes stationarity, we use a quadratic polynomial in two variables as the mean function to eliminate potential trend in data:

$$\mu(\mathbf{x}) = \mathbf{c}_0 + \mathbf{c}_1 \cdot x_1 + \mathbf{c}_2 \cdot x_2 + \mathbf{c}_{12} \cdot x_1x_2 + \mathbf{c}_{11} \cdot x_1^2 + \mathbf{c}_{22} \cdot x_2^2 \tag{9}$$

where $\mathbf{c}_0,\ \mathbf{c}_1,\ \mathbf{c}_2, \mathbf{c}_{12}, \mathbf{c}_{11}, \mathbf{c}_{22}$ are vectors of size M also being optimized during the model training. Optimization of the hyper-parameters is performed with MLE approach by solving the following maximization problem numerically:

$$\max_{\theta}\ \log p(\mathbf{Y}|\mathbf{X},\theta) = -\frac{1}{2}(\overline{\mathbf{y}} - \overline{\mu}(\mathbf{X}))^T\Sigma^{-1}(\overline{\mathbf{y}} - \overline{\mu}(\mathbf{X})) - \frac{1}{2}\log|\Sigma| - \frac{NM}{2}\log 2\pi, \tag{10}$$

where θ denotes all hyper-parameters of the model.

Data normalization. Spatial coordinates first were converted from EPSG:4326 (latitude, longitude) format to EPSG:32637 (UTM zone 37N) and, then, scaled down to [0, 1] range using the min-max normalization.

The scaling of the measured properties required a more complex approach. All of the properties are limited from below by zero and from above with some reasonable values, e.g., concentration can not be more than 100%, and, as another example, pH values can lie only in [0, 14] range. Unfortunately, a direct GPR does not consider such limits on outputs as the predictive distribution is normal and has infinite support. To incorporate such bounds we follow a *warping* approach[63,64], which allows to map measurements from bounded space to unbounded and apply GPR directly. First, let M-size vectors $\mathbf{b}^L$ and $\mathbf{b}^U$ denote bounds of parameters dictated by regulatory documents from the Table 1. For every modeled parameter without the explicit value domain (all, except for pH in our case) we calculate their maximum values across all of the measurements $\mathbf{y}^{max} = \max\limits_{1\le i\le N}\mathbf{y}_i$. Then, we choose the maximum between the obtained values and $\mathbf{b}^U$ and multiply it by 10 (to certainly avoid out-of-bounds problem), thus, defining upper limits as $10 \cdot \max\{\mathbf{y}^{max}, \mathbf{b}^U\}$. The boundaries for pH are set as [0, 14] and the obtained limits are used for the min-max scaling, mapping all of the parameters to [0, 1] region. Furthermore, to map the scaled parameters to an unbounded space we use the inverse cumulative-distribution function (ICDF)

of the standard normal distribution. Unfortunately, it can not be applied straightforwardly, because our dataset contains strict zero values (which coincide with lower bounds), thus, yielding $-\infty$ values after the transformation. To tackle this issue, we simply replace zeros with sufficiently small values of 10^{-10} before applying ICDF. Finally, we use another min-max scaling to end up with [0, 1] range of values. Thus, GPR predictive mean ends up in the required bounds after the appropriate inverse steps. Noteworthily, predictive distribution is Gaussian in the transformed measurement space, however, it is different in the original space and heavily depends on the warping function.

Validation. To validate and compare the trained models we apply a standard cross-validation scheme with 5 random splits, with 80% and 20% of a train and test data, respectively. For each split we (i) perform the model fitting on training data, (ii) obtain predictions for each modeled property for the test data points and (iii) calculate R^2-score (or the coefficient of determination). This quality metric shows the proportion of the observed variation explained by the variation in the input data using the model. It is equal to one (1) if a predictive error is zero (0), zero (0) if a predictive error equals the test data variance, and negative if it is larger. The particular choice of the metric based on the Mean Square Error (MSE) is justified by our model selection. The predictive mean of GPR is tightly connected with a solution of Kernel Ridge Regression[65], which involves the minimization of the exact MSE of the training data with additional regularization. As the model is intrinsically trained to minimize the Euclidean error, it is natural to use the MSE-based metrics for its evaluation. To select the best model we have to disregard the poorly modeled properties. Thus, we average R^2-scores over all splits and remove all of the properties for which the maximum, calculated over every model, yielded a negative value. Then, we average the scores over all kept properties and splits and select the model with the highest value.

Probabilistic substance quality index. To evaluate the water quality we propose a new technique that takes advantage of GPR and allows to incorporate regulatory standards into assessment procedure directly. First, we note that the output of the GPR model is not just a vector of predicted values of properties, but a probabilistic distribution. Namely, at any location it gives a multi-dimensional normal distribution $\mathbf{z} \sim \hat{p}(\mathbf{z} \mid \mathbf{x}_*) = \mathcal{N}\left(\hat{\mu}(\mathbf{x}_*), \hat{\Sigma}(\mathbf{x}_*)\right)$ described by the mean vector and covariance matrix from the group of Eqs. (6). Second, we appeal to the fact that the water quality is considered high if concentrations of different elements are located in admissible safe bounds, defined by governmental standards. Taking into account the above mentioned, we propose a measure coined Probabilistic Substance Quality Index (PSQI), which depicts the probability that all the measured properties will be within the admissible bounds. Therefore, it seems natural to integrate the probability density function $\hat{p}(\mathbf{z} \mid \mathbf{x}_*)$ over these bounds. Unfortunately, due to the "curse of dimensionality", an increase in the number of properties leads to a drastic decrease of the integral value and overall interpretability. Thus, we utilize the marginalization approach instead and define PSQI as follows:

$$\begin{aligned} \mathrm{PSQI}(\mathbf{x}_*) &= \sum_{i=1}^{M} w_i \cdot \hat{p}_i(\mathbf{x}_*), \ \sum_{i=1}^{M} w_i = 1, \\ \hat{p}_i(\mathbf{x}_*) &= \int_{-\infty}^{+\infty} \dots \int_{b_i^L}^{b_i^U} \dots \int_{-\infty}^{+\infty} \hat{p}(\mathbf{z} \mid \mathbf{x}_*) \mathrm{d}z_1 \dots \mathrm{d}z_M, \end{aligned} \tag{11}$$

where w_i denotes the importance of each individual property in water quality and $\mathbf{b}^L$, $\mathbf{b}^U$ are admissible bounds for parameters from Table 1. By its construction, it is normalized to [0,1] interval, where zero value corresponds to zero possibility that properties are within the admissible bounds (bad quality) and one corresponds to the opposite (good quality). Since the integral in Eq. (11) does not have an analytical solution for the arbitrary bounds, we use *SciPy* library[66] to perform numerical integration of multivariate Gaussian probability density.

Unfortunately, PSQI alone does not allow us to distinguish between the following cases of the index values being small: (a) predictive mean lies outside of the bounds and predictive variance is small (i.e., water is bad and we are certain about it); and (b) predictive mean lies inside of the bounds and predictive variance is large (i.e., water is good, but we are uncertain about it). To tackle this issue, we propose an additional confidence metric:

$$\begin{aligned} \mathrm{conf}(\mathbf{x}_*) &= \sum_{i=1}^{M} w_i \cdot \hat{q}_i(\mathbf{x}_*), \ \sum_{i=1}^{M} w_i = 1, \\ \hat{q}_i(\mathbf{x}_*) &= \int_{-\infty}^{+\infty} \dots \int_{b_i^L}^{b_i^U} \dots \int_{-\infty}^{+\infty} \hat{q}(\mathbf{z} \mid \mathbf{x}_*) \mathrm{d}z_1 \dots \mathrm{d}z_M, \\ \hat{q}(\mathbf{z} \mid \mathbf{x}_*) &= \mathcal{N}\left((\mathbf{b}^L + \mathbf{b}^U)/2, \hat{\Sigma}(\mathbf{x}_*)\right). \end{aligned} \tag{12}$$

Confidence is calculated similarly to PSQI, although with an important difference—predictive distribution is centered within the bounds. This way, the confidence value is high, if the standard deviation is much smaller than the bounds and low, otherwise. One can note that PSQI and confidence values are interconnected, e.g., if the predictive distribution is already centered within admissible bounds, then, they match. Figure 3 shows PSQI and confidence calculations for a single modeled parameter ($M = 1$).

The selection of weights w_i can be governed by the hazardousness of deficiency or excess of individual properties. Unfortunately, it may be not very straightforward to choose particular values of the weights with such an approach. Instead, we use the model training results to determine the "importance" of model properties for PSQI

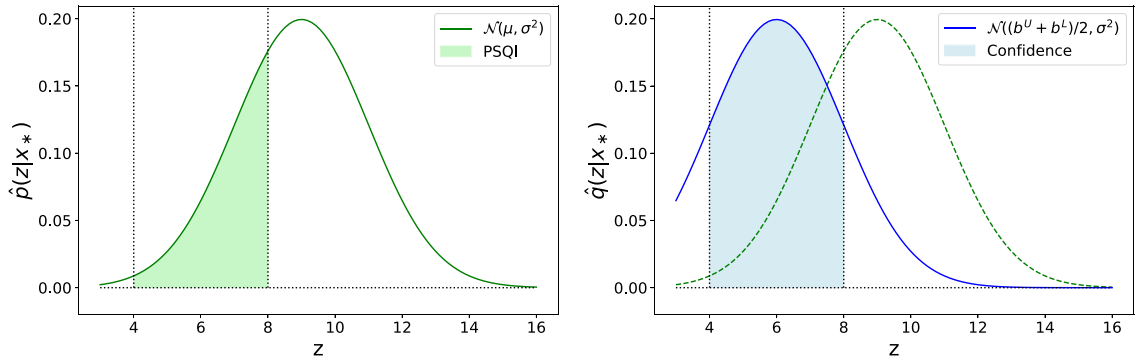


Figure 3. Example of PSQI (left) and confidence (right) calculation procedures for a single parameter. Dotted vertical lines denote admissible bounds b^L and b^U, green and blue solid lines denote obtained predictive distribution and its centered variant, respectively. Both PSQI and confidence values are calculated as the area under the respective distribution curves between the bounds.

computation. During training, we apply five-fold cross-validation and select the best model based on average R^2-scores for each modeled property. It is worth noting that R^2-score computed in transformed and original measurement space is different due to the non-linear nature of transformation (see in "Data normalization" section). Thus, we perform the model selection based on R^2 computed in the original state space. For some of the properties the model can perform poorly and yield negative R^2-score values, which implies that the simple mean has a better predicting capacity than the model. In this case, the properties with non-positive R^2-scores of their predictive means and variances are replaced with respective dataset means and variances, and all their inter-parameter correlations are considered zero. Further, to deal with the discrepancy of prediction accuracy among different properties, we propose to compute weights from Eqs. (11) and (12) using R^2-scores and *softmax* function:

$$w_i = \frac{\exp\left(\max(R_i^2, 0)\right)}{\sum_{i=1}^{M} \exp\left(\max(R_i^2, 0)\right)}, \tag{13}$$

where R_i^2 denotes R^2-score obtained for i^{th} property during training. This way, we incorporate modeling accuracy into PSQI computations directly and reduce potential over- or under-estimations of the index. Moreover, it is still possible to incorporate ad-hoc importance of particular properties with an additional re-weighting.

Results

The source code, data, and results can be found in our repository[67] with available interactive visualization via kepler.gl platform.

Experiment. To apply and analyze the proposed approach we used a dataset obtained from a large environmental investigation in the New Moscow area, Russia. Firstly, we modeled spatial distribution of multiple available parameters, Alkalinity, Hardness, Mineralization, pH, ions of sodium (Na), potassium (K), calcium (Ca), magnesium (Mg), manganese (Mn), iron (Fe), copper (Cu), nickel (Ni), chrome (Cr), zinc (Zn), bicarbonate (HCO_3), ammonium (NH_4), nitrate (NO_2), nitrite (NO_3), chloride (Cl), orthophosphate (PO_4), and sulfate (SO_4), implementing multi-task GPR framework (see in "Gaussian process regression" section).

To avoid serious over-fitting during the training phase, we bounded the corresponding length-scales of the mixture components within [0.1, 100] interval. We trained several models with different number of mixtures Q (from 1 to 5) in the spatial kernel (see Eq. 8) and different ranks r (3, 5, 7, 10, 15) of the inter-feature covariance matrix K^f (see Eq. 7). To select the best model, we considered prediction accuracy only for 20 components, except for Cr as it yielded very poor results for every model. Figure 4 shows the average R^2-score over every kept component and split for each model for different values of Q and r. It can be seen, that increasing model complexity does not necessarily lead to model improvements, as it can bring about over-fitting. Moreover, it typically causes time-performance degradation. The best model corresponds to a single mixture component and rank-10 covariance matrix. Training of the best model took 6.8 s, whereas evaluation over test data took 0.6 s.

Modeling. Figure 5a illustrates per-parameter performance with average, min, and max values of R^2-score over splits for different properties obtained using the best selected model. We can see that some of the properties were predicted very poorly, such as Cr, Fe, Ni, Cu, NH_4, NO_2 with the best average score of 0.635 for SO_4. One of the reasons for the low accuracy may be a high level of noise in the data. It is accounted in the model and estimated during the training phase as an additive Gaussian component with the covariance matrix D from Eq. (4). Large noise values tend to correspond to a poor fit of the model given the training data for a particular water property. Unfortunately, it is not possible to illustrate the noise values in the original state space straightforwardly. The reason is that the modeling is done in normalized parameter state space (see in "Data normalization" section), therefore, in the original space normal noise is not obliged to be normal anymore. We illustrate it with Fig. 5b in a more comprehensive way, dividing the inter-quartile range for each noise component by the

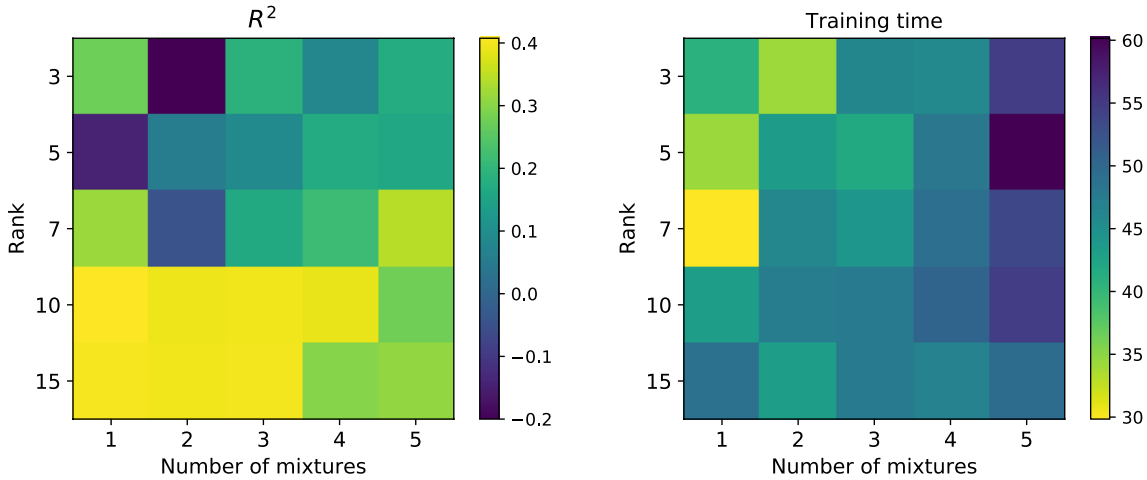


Figure 4. R^2-score averaged over every split and well-modeled property (left) and complete training time of a model in seconds (right) for different numbers of mixture components Q and ranks r of the inter-feature covariance matrix K^f. Increase in number of mixtures leads to over-parametrization and degradation in both accuracy and computational speed. Whereas, increase of the rank improves accuracy with a reasonable increase of computational time.

respective normalized normative range. For water properties without established restrictions, i.e., K, HCO_3, Ca, relative noise is zero, and for poorly modeled properties relative noise is very high.

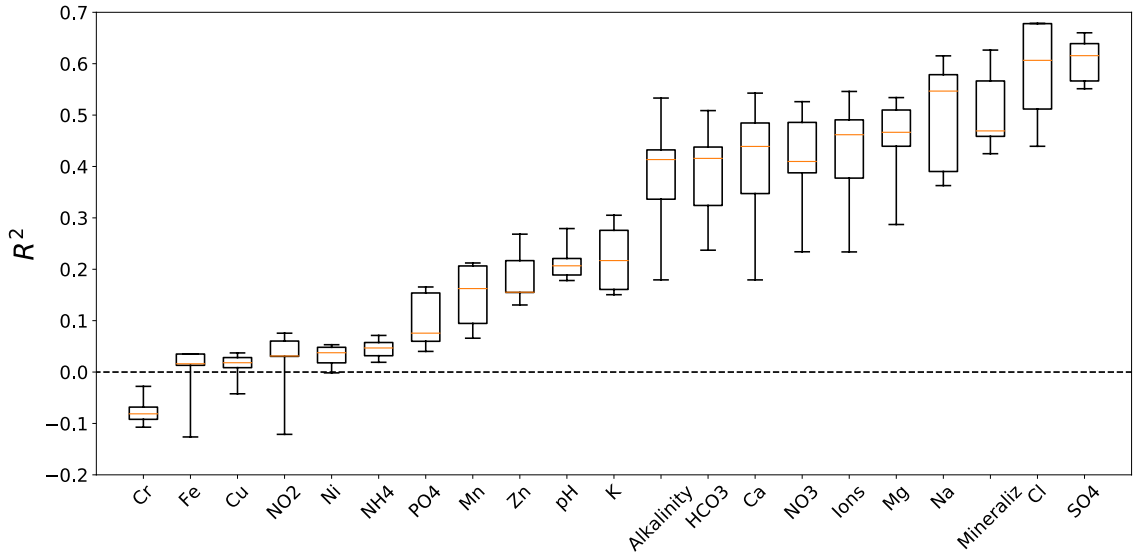
As a "side-effect" of the model construction we obtained the optimized K^f covariance matrix describing interconnections of water properties, for which, corresponding correlation matrix is shown in Fig. 6. General characteristics such as Alkalinity, Hardness and Mineralization rates are highly correlated with HCO_3, Ca, Mg, correlation coefficients (ρ) lay in diapason from 0.63 to 0.99, while the presence of SO_4 mostly correlates with ions of Na, Mg, correlations are 0.79 and 0.65, respectively. Apart from that, highest correlations (more than 0.6) are observed in pairs: ρ(Cr & Fe) = 0.81, ρ(Cr & Cu) = 0.73, ρ(Cr & Ni) = 0.75, ρ(Cr & Mn) = 0.74, ρ(Cr & K) = 0.68, ρ(NH_4 & Fe) = 0.63, ρ(Cu & Ni) = 0.85, ρ(Cu & Zn) = 0.77.

PSQI. In order to calculate the PSQI values, we have used admissible bounds reflected in local regulations of the Russian Federation (see Table 1). Some of the parameters do not have any regulatory restrictions (Ca, K and HCO_3), thus, we excluded them from the calculation of PSQI to avoid overestimation. As could be noted earlier from Fig. 5, the prediction quality of our model differs across the properties and for some of them even yield negative R^2-scores. Calculation of PSQI at a single point comprises two steps: (a) evaluation of the predictive distribution; (b) computations from Eqs. (11) and (12) (see in "Probabilistic substance quality index" section). We chose the best performing model during the validation stage, fixed all of its hyper-parameters, and used the whole dataset to make predictions at locations of interest. They were uniformly selected across the New Moscow region with 100 m^2 resolution, giving 151 447 points. The evaluation of the predictive distribution took 5.8 hours with approximately 130 ms per point. The subsequent computation of PSQI took only 95 seconds, which can be considered negligible. Figure 7 shows the spatial distribution of PSQI values obtained from predicted distributions, where outlined points denote collected samples. To additionally validate that PSQI indeed corresponds to the fraction of measured parameters being in admissible bounds, we, (i) evaluated PSQI at each sampling location, (ii) calculated such fraction directly for each measurement, and finally (iii) computed Pearson correlation coefficient between them, resulting in the reasonably large value of 0.68. The spatial distribution of PSQI confidence appeared to be of no practical interest, due to its very small scatter from 0.935 to 0.956 with 55% points yielding values less than 0.936.

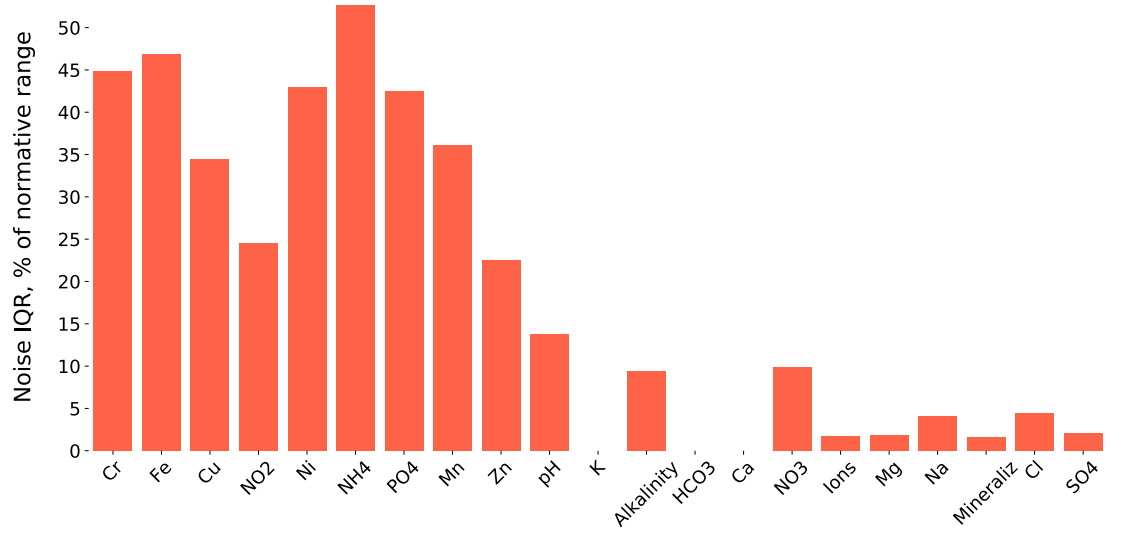
Discussion

A typical modeling task consists of several pre-processing steps, one of which may be dimensionality reduction used to disregard the non-informative features or find the most "independent" combinations of the features to build the model for. It may help to decrease the computational complexity, but requires a careful dataset pre-processing and leads to information losses. Multi-task GPR allows to perform simultaneous geospatial modeling and capture the inter-feature dependencies while being able to control complexity using parametrization techniques. However, the computational overhead may reach as much as $O(N^3M^3)$, thus, requiring further considerations of performance improvements[50,51,68–71].

One of key issues for geo-spatial modeling is the model construction process itself. On the one hand, manual selection of the kernel function based on domain knowledge allows to adapt to different areas of application. On the other hand, it limits automatization and scalability of the modeling significantly and causes some difficulty for integrating it into support-decision systems. Spectral Mixture Kernel facilitates the task of model construction, giving the ability to approximate arbitrary stationary kernels and control the accuracy with the number of mixtures. Since the number of mixtures is discrete, it can not be effectively optimized using MLE. In this case



(a) R^2-score box plot for the best model predictions obtained during the training phase across different data splits, where caps denote minimum and maximum values. Prediction accuracy varies across the parameters, where some of them yield small or even negative R^2-scores.



(b) The inter-quartile ranges for estimated normal noise D relative to the respective normative ranges, in percent. For K, HCO_3 and Ca, there are no restrictions, thus, relative noise is depicted as zero. As expected, large estimated noise values correspond to poorly modeled parameters.

Figure 5. Modeling results.

multiple models (for the different number of mixtures) can be trained and compared using the Cross-Validation technique to pick the best overall solution. However, the increase in the model complexity may lead not only to accuracy improvements, but on the contrary to the over-parametrization and degradation of both computational speed and the quality of predictions. Therefore, effective model construction still requires thorough consideration of both domain knowledge and the dataset structure.

Freshwater characteristics usually depend on various factors such as the intensity of geological and hydro-geological settings due to dissolution processes and ion exchanges; seasonal fluctuations and climate change in global, being also affected by anthropogenic loads[72]. The modeled properties show reasonable correlations (Fig. 6), adequate for the natural freshwater resources and explainable for those influenced by urbanization and agricultural activity widespread across the territory of sample net. Ions of HCO_3, Ca, Mg, SO_4, Na, Mg, being major macro constituents, are always presented in groundwater, their concentrations depend mostly on the mineral composition of rocks, although surrounding lands as well[73]. The presence of nitrogen forms in the ground and surface water can accompany the agricultural and landfill sites, relating to the migration of products of fertilizers, pesticides or domestic sewage decay[74–77]. The observed correlations between NH_4, NO_2, NO_3 are ranged according to the stages of oxidation transformation of NH_4. Additionally, substantial correlations

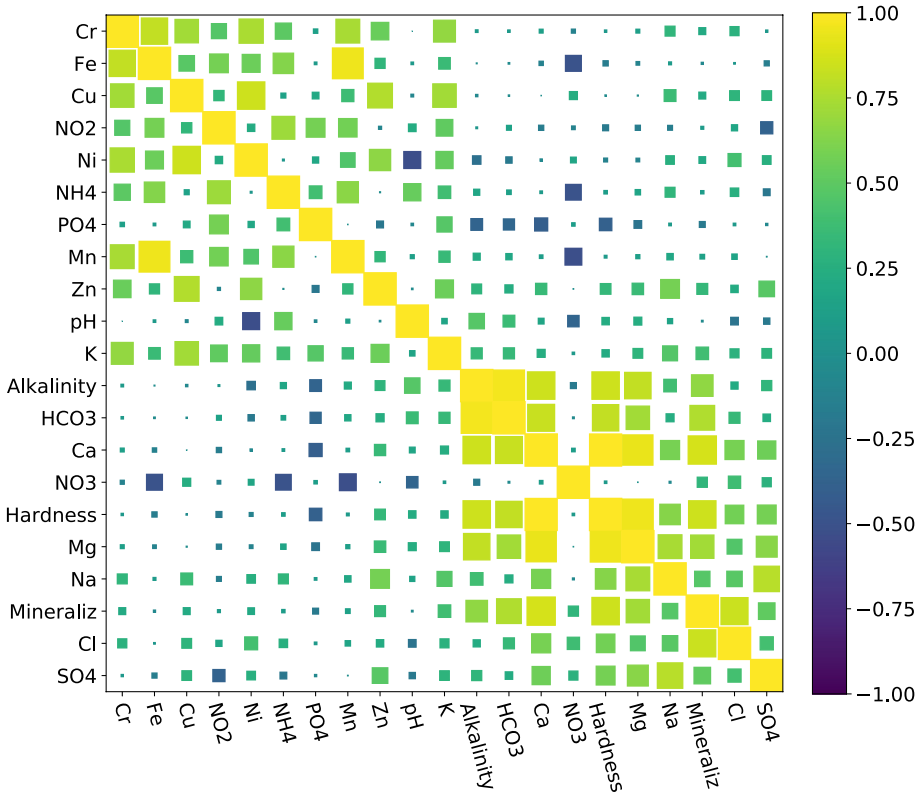


Figure 6. Pearson correlation matrix for the modeled properties derived from the optimized K^f covariance matrix. Yellow (dark blue) color denotes positive (negative) correlation, whereas wide (narrow) boxes represent strong (weak) correlation.

between Cr, Fe, Cu, Ni, Mn, K, as well as between PO_4 and K also often are explained by the migration of fertilizers' degradation residuals across the landscape and into water sources. Apart from that, high intercorrelation between trace elements in the water samples can be linked to the objects with a high pollution potential, such as landfills, transport systems, industry. Among them, Ni demonstrates a remarkable negative correlation of -0.52 with pH, which can be explained by the reduced migration ability in alkaline conditions and can be supported by the noticeable correlation of -0.48 between pH and Alkalinity.

Although more sophisticated variations of WQI have been recently proposed (e.g. based on multi-criteria decision analysis (MCDA)[28], entropy-weighted indices[30], modified by principal component analysis for optimized parameter selection[29]), most of the existing WQI solutions have limitations, e.g. low independence of expertise and, as a consequence, site and case specificity as well as low robustness. In general, estimation of WQI includes the following steps: scaling of the selected parameters if they have different dimensions; selection of the most important parameters according to some rule, including an a priori knowledge; determining the relative weight of each parameter; calculating the sub-index of each parameter from the relative weight, and, finally, summarizing the results and determining the quality rating scale[78–80]. However, subjective judgments may cause certain confusion, e.g., in the determination of parameters importance, in the choice of the weight values for each parameter, in comparing the sum with expert-opinion-based ranges, as well as different limits scales of parameters.

As compared to the methodologies discussed above the proposed PSQI is directly linked to the established water quality guideline standards for each characteristic itself. PSQI does not rely on the subjective judgments about parameters' importance neither on the structure of the input data. Additionally, PSQI covers admissible limits not only as single values, e.g., when properties must be lower than specific upper bound, but allows to consider an optimal range, such as for pH or alkalinity. This is an important improvement of the currently used techniques and it makes the proposed solution applicable for the assessment of other environmental media, e.g. in the case of soils favorable content of macro and micro-nutrients or physical properties are expressed as optimum ranges[81,82].

Materials

Study area. The data was collected in the 2017–2018 years. A detailed description of the territory and the sample net is provided in Shadrin et al[83]. In a nutshell, the sample net is mostly located across the New Moscow region, Russia. According to the official statistics and research reports New Moscow is characterized by the rapid rise of both urban areas and density during the last 10 years[84,85]. The New Moscow is located close to the bottom boundary of southern taiga, in the Central European part of Russia (55°N, 37°E) and extends over 1480 km^2 in area. The mean annual temperature of this region is about 3-4° C. The territory includes all of the common

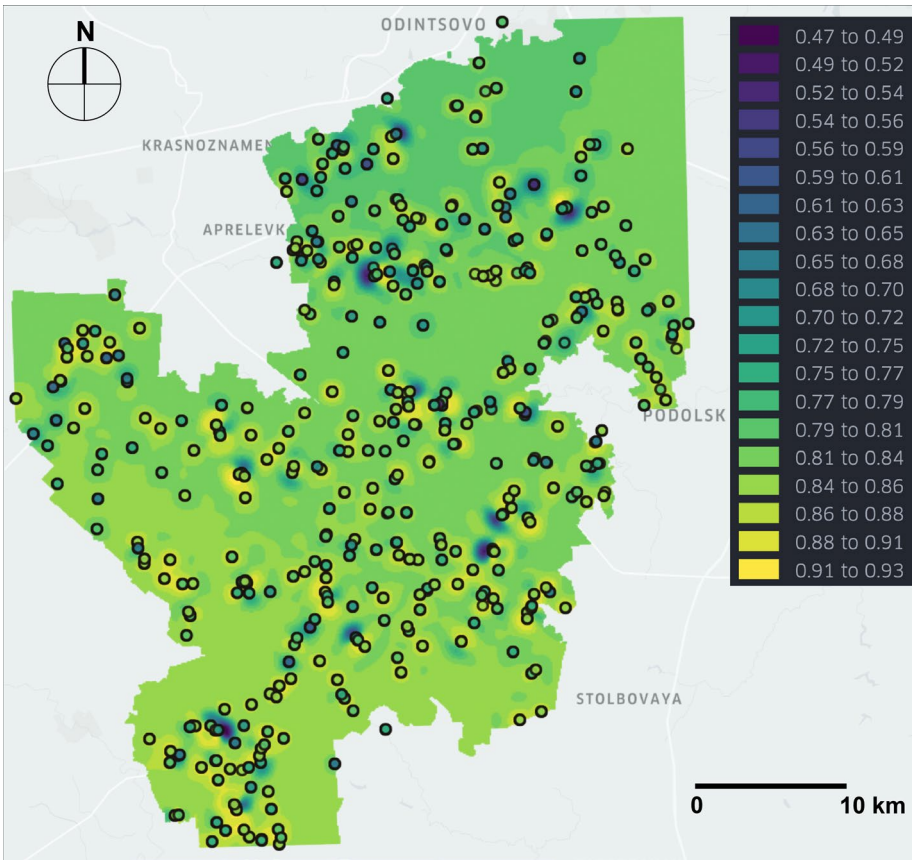


Figure 7. Geo-spatial map of predicted PSQI values. Dark violet color represents low PSQI values, whereas light yellow - high PSQI values. Outlined points are sampled measurements with color representing the fraction of properties that appear to be in admissible bounds. The map was created with Kepler.gl platform (v2.5.1).

Parameter	Dimension	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Normative range	
	pH	–	5.50	6.73	7.10	7.04	7.40	8.40	6–9
Alkalinity	mg-eq/L	0.50	3.45	4.50	4.68	5.84	12.00	0.5-6.5	
Hardness	mg-eq/L	0.60	4.20	5.60	5.74	6.90	21.90	7	
Mineraliz	mg-eq/L	37.00	281.80	366.00	402.60	481.00	1586.00	1000	
Ca	mg/L	8.85	63.60	82.17	85.74	101.90	340.00	–	
Mg	mg/L	1.48	12.01	16.81	17.67	22.04	60.44	50	
Na	mg/L	0.00	9.99	16.19	26.22	31.22	245.00	200	
K	mg/L	0.00	1.09	2.59	8.56	6.94	181.80	–	
NH_4	mg/L	0.00	0.00	0.06	0.54	0.40	38.00	2	
HCO_3,	mg/L	31.00	210.00	275.00	285.70	356.20	732.00	–	
Cl	mg/L	0.00	12.26	26.07	53.96	59.29	748.41	350	
NO_3	mg/L	0.00	5.18	17.16	27.20	37.61	352.81	45	
NO_2	mg/L	0.00	0.00	0.00	0.02	0.00	2.25	3	
PO_4	mg/L	0.00	0.00	0.00	0.35	0.00	15.32	3.5	
SO_4	mg/L	0.80	20.53	34.08	40.44	52.20	246.14	500	
Cr	mg/L	0.00	0.00	0.00	0.00	0.00	0.04	0.05	
Cu	mg/L	0.00	0.00	0.00	0.00	0.00	0.13	1	
Fe	mg/L	0.00	0.04	0.14	0.34	0.32	18.52	0.3–1	
Mn	mg/L	0.00	0.00	0.01	0.06	0.04	3.12	0.1	
Ni	mg/L	0.00	0.00	0.00	0.00	0.00	0.27	0.1	
Zn	mg/L	0.00	0.01	0.04	0.15	0.11	4.90	5	

Table 1. List of parameters of study dataset and their basic statistics: minimum (**Min.**), first quartille (**1st Qu.**), median, mean, third quartile (**3rd Qu.**), maximum (**Max.**) and normative range, represented according to sanitary regulations in Russia.

types of land-uses, including urban fabric, forests, and green urban areas, arable lands, industrial cites. The predominant types of natural vegetation are coniferous and broad-leaved forests, while agricultural lands include pastures and arable land mostly growing feed crops and cereals[86]. The mean temperature in the coldest month of the year (January) ranges between -9.5° C and -11.5° C, while in the warmest month, July, mean temperatures are between +17° C and +18.5° C. The average annual precipitation is approximately 400-500 mm, with around two-thirds by rainfall and the rest by snow, according to recent observations from weather stations' net across the territory (available at https://www.ncdc.noaa.gov/cdo-web/datatools/findstation). The territory has a plain topographic relief, and the bedrock consists of glacial and fluvioglacial loams and sands with the inclusion of sandy alluvial deposits.

Dataset description. The analytical samples were collected from different sources of freshwater, namely: wells, rivers, and springs from the territories of private households. Some of the sample points included the replicated measurements, which have been taken into account at both modeling and results' interpretation stages. Overall, the dataset includes 1569 samples at 460 unique points, each sample consists of longitude, latitude, and list of chemical compounds content and properties, commonly used for water quality assessment. The normative ranges for measured properties are given according to the Russian regulation documents—SanPiN (Sanitary Rules and Norms), number 1.2.3685-21 being in force at the time of the manuscript preparation. These values were given as an example for study support and can be changed according to any other guideline source. Although being measured, Hg, Cd, Co, Pb were eliminated from further work as having very low variability across the dataset. The data points sampled too far from the main research area were removed. Finally, for further modeling, we used 1526 data vectors of 21 properties, namely Alkalinity, Ca, Cl, Cr, Cu, Fe, HCO_3, Hardness, K, Mg, Mineralization, Mn, NH_4, NO_2, NO_3, Na, Ni, PO_4 SO_4, Zn, pH (see Table 1).

Code availability

The source code, data, and results are provided in our publicly accessible repository[67] with available interactive visualization via kepler.gl platform and step-by-step instructions.

Received: 18 February 2021; Accepted: 16 November 2021
Published online: 10 December 2021

References

1. Abell, R. & Harrison, I. J. A boost for freshwater conservation. *Science* **370**, 38–39 (2020).
2. Dudgeon, D. *et al.* Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biol. Rev.* **81**, 163–182 (2006).
3. Boulton, A. J., Fenwick, G. D., Hancock, P. J. & Harvey, M. S. Biodiversity, functional roles and ecosystem services of groundwater invertebrates. *Invert. Syst.* **22**, 103–116 (2008).
4. Tait, P., Baskaran, R., Cullen, R. & Bicknell, K. Nonmarket valuation of water quality: Addressing spatially heterogeneous preferences using GIS and a random parameter logit model. *Ecol. Econ.* **75**, 15–21 (2012).
5. Siebert, S. *et al.* Groundwater use for irrigation—A global inventory. *Hydrol. Earth Syst. Sci.* **14**, 1863–1880 (2010).
6. Álvarez-Cabria, M., Barquín, J. & Peñas, F. J. Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. *Sci. Total Environ.* **545**, 152–162 (2016).
7. Gu, Q. *et al.* Characterizing the spatial variations of the relationship between land use and surface water quality using self-organizing map approach. *Ecol. Indicat.* **102**, 633–643 (2019).
8. Mirzaei, M. *et al.* Mitigating environmental risks: Modeling the interaction of water quality parameters and land use cover. *Land Use Policy* **95**, 103766 (2020).
9. Horton, R. K. An index number system for rating water quality. *J. Water Pollut. Control Fed.* **37**, 300–306 (1965).
10. Jha, M. K., Shekhar, A. & Jenifer, M. A. Assessing groundwater quality for drinking water supply using hybrid fuzzy-GIS-based water quality index. *Water Res.* **179**, 115867 (2020).
11. Tyagi, S., Sharma, B., Singh, P. & Dobhal, R. Water quality assessment in terms of water quality index. *Am. J. Water Resour.* **1**, 34–38 (2013).
12. Saeedi, M., Abessi, O., Sharifi, F. & Meraji, H. Development of groundwater quality index. *Environ. Monitor. Assess.* **163**, 327–335 (2010).
13. Katyal, D. Water quality indices used for surface water vulnerability assessment. *Int. J. Environ. Sci.* **2** (2011).
14. Sutadian, A. D., Muttil, N., Yilmaz, A. G. & Perera, B. Development of river water quality indices—A review. *Environ. Monitor. Assess.* **188**, 58 (2016).
15. Uddin, M. G., Nash, S. & Olbert, A. I. A review of water quality index models and their use for assessing surface water quality. *Ecol. Indicat.* **122**, 107218 (2021).
16. Bünemann, E. K. *et al.* Soil quality—A critical review. *Soil Biol. Biochem.* **120**, 105–125 (2018).
17. Raiesi, F. A minimum data set and soil quality index to quantify the effect of land use conversion on soil quality and degradation in native rangelands of upland arid and semiarid regions. *Ecol. Indicat.* **75**, 307–320 (2017).
18. Bandyopadhyay, S. & Maiti, S. K. Application of statistical and machine learning approach for prediction of soil quality index formulated to evaluate trajectory of ecosystem recovery in coal mine degraded land. *Ecol. Eng.* **170**, 106351 (2021).
19. Kachroud, M., Trolard, F., Kefi, M., Jebari, S. & Bourrié, G. Water quality indices: Challenges and application limits in the literature. *Water* **11**, 361 (2019).
20. Misaghi, F., Delgosha, F., Razzaghmanesh, M. & Myers, B. Introducing a water quality index for assessing water for irrigation purposes: A case study of the Ghezel Ozan River. *Sci. Total Environ.* **589**, 107–116 (2017).
21. de Andrade Costa, D., de Azevedo, J. P. S., Dos Santos, M. A. & Assumpção, R. D. S. F. V. Water quality assessment based on multivariate statistics and water quality index of a strategic river in the Brazilian Atlantic Forest.. *Sci. Rep.* **10**, 1–13 (2020).
22. Lumb, A., Sharma, T. & Bibeault, J.-F. A review of genesis and evolution of water quality index (WQI) and some future directions. *Water Quality Exposure Health* **3**, 11–24 (2011).
23. Mukate, S., Wagh, V., Panaskar, D., Jacobs, J. A. & Sawant, A. Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water. *Ecol. Indicat.* **101**, 348–354 (2019).
24. Ewaid, S. H., Abed, S. A., Al-Ansari, N. & Salih, R. M. Development and evaluation of a water quality index for the Iraqi rivers. *Hydrology* **7**, 67 (2020).
25. Alver, A. Evaluation of conventional drinking water treatment plant efficiency according to water quality index and health risk assessment. *Environ. Sci. Pollut. Res.* **26**, 27225–27238 (2019).

26. Gao, Y. *et al.* Hydrogeochemical characterization and quality assessment of groundwater based on integrated-weight water quality index in a concentrated urban area. *J. Cleaner Prod.* **260**, 121006 (2020).
27. Mohammadpour, R. *et al.* Prediction of water quality index in constructed wetlands using support vector machine. *Environ. Sci. Pollut. Res.* **22**, 6208–6219 (2015).
28. Jhariya, D., Kumar, T., Dewangan, R., Pal, D. & Dewangan, P. K. Assessment of groundwater quality index for drinking purpose in the Durg district, Chhattisgarh using geographical information system (GIS) and multi-criteria decision analysis (MCDA) techniques. *J. Geol. Soc. India* **89**, 453–459 (2017).
29. Tripathi, M. & Singal, S. K. Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India. *Ecol. Indicat.* **96**, 430–436 (2019).
30. Islam, A. R. M. T., Ahmed, N., Bodrud-Doza, M. & Chu, R. Characterizing groundwater quality ranks for drinking purposes in Sylhet district, Bangladesh, using entropy method, spatial autocorrelation index, and geostatistics. *Environ. Sci. Pollut. Res.* **24**, 26350–26374 (2017).
31. Pak, H. Y., Chuah, C. J., Tan, M. L., Yong, E. L. & Snyder, S. A. A framework for assessing the adequacy of water quality index-quantifying parameter sensitivity and uncertainties in missing values distribution. *Sci. Total Environ.* **751**, 141982 (2021).
32. Huan, H. *et al.* Quantitative evaluation of specific vulnerability to nitrate for groundwater resource protection based on process-based simulation model. *Sci. Total Environ.* **550**, 768–784 (2016).
33. Schenk, E. R., O'Donnell, F., Springer, A. E. & Stevens, L. E. The impacts of tree stand thinning on groundwater recharge in aridland forests. *Ecol. Eng.* **145**, 105701 (2020).
34. Anderson, M. P., Woessner, W. W. & Hunt, R. J. *Applied groundwater modeling: Simulation of flow and advective transport* (Academic press, 2015).
35. Hayley, K. The present state and future application of cloud computing for numerical groundwater modeling. *Groundwater* **55**, 678–682 (2017).
36. Clark, M. P. *et al.* The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrol. Earth Syst. Sci.* **21**, 3427–3440 (2017).
37. Wang, X., Zhang, F. & Ding, J. Evaluation of water quality based on a machine learning algorithm and Water Quality Index for the Ebinur Lake Watershed. *China. Sci. Rep.* **7**, 12858 (2017).
38. Chen, K. *et al.* Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **171**, 115454 (2020).
39. Bindal, S. & Singh, C. K. Predicting groundwater arsenic contamination: Regions at risk in highest populated state of India. *Water Res.* **159**, 65–76 (2019).
40. Lu, H. & Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **249**, 126169 (2020).
41. Chen, C., He, W., Zhou, H., Xue, Y. & Zhu, M. A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Sci. Rep.* **10**, 1–13 (2020).
42. Kisi, O., Keshavarzi, A., Shiri, J., Zounemat-Kermani, M. & Omran, E.-S.E. Groundwater quality modeling using neuro-particle swarm optimization and neuro-differential evolution techniques. *Hydrol. Res.* **48**, 1508–1519 (2017).
43. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **721**, 137612 (2020).
44. Belkhiri, L., Tiri, A. & Mouni, L. Spatial distribution of the groundwater quality using kriging and Co-kriging interpolations. *Groundwater Sustain. Develop.* **11**, 100473 (2020).
45. Ruybal, C. J., Hogue, T. S. & McCray, J. E. Evaluation of groundwater levels in the Arapahoe aquifer using spatiotemporal regression kriging. *Water Resour. Res.* **55**, 2820–2837 (2019).
46. Pouladi, N., Møller, A. B., Tabatabai, S. & Greve, M. H. Mapping soil organic matter contents at field level with Cubist. *Random Forest kriging. Geoderma* **342**, 85–92 (2019).
47. Nori-Sarma, A. *et al.* Low-cost NO_2 monitoring and predictions of urban exposure using universal kriging and land-use regression modelling in Mysore India. *Atmospheric Environ.* **226**, 117395 (2020).
48. Ingram, M., Vukcevic, D. & Golding, N. Multi-output Gaussian processes for species distribution modelling. *Methods Ecol. Evolut.* **11**, 1587–1598 (2020).
49. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (The MIT Press, 2006).
50. Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian Processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, 282–290 (AUAI Press, Arlington, Virginia, USA, 2013).
51. Wilson, A. & Nickisch, H. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, 1775–1784 (PMLR, 2015).
52. Wilson, A. G., Hu, Z., Salakhutdinov, R. & Xing, E. P. Deep kernel learning. In *Artificial intelligence and statistics*, 370–378 (PMLR, 2016).
53. Xu, S., An, X., Qiao, X., Zhu, L. & Li, L. Multi-output least-squares support vector regression machines. *Pattern Recognit. Lett.* **34**, 1078–1084 (2013).
54. Isazadeh, M., Biazar, S. M. & Ashrafzadeh, A. Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. *Environ. Earth Sci.* **76**, 1–14 (2017).
55. Taghizadeh-Mehrjardi, R. *et al.* Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. *Geoderma* **376**, 114552 (2020).
56. Yang, J., Wang, X., Wang, R. & Wang, H. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using vis-nir spectroscopy. *Geoderma* **380**, 114616 (2020).
57. Pukalchik, M. *et al. Freshwater chemical properties for New Moscow region.*https://doi.org/10.6084/m9.figshare.10283225.v2 *(2020).*
58. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning* Vol. 112 (Springer, 2013).
59. Bonilla, E. V., Chai, K. M. & Williams, C. Multi-task Gaussian process prediction. In *Advances in neural information processing systems*, 153–160 (2008).
60. Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q. & Wilson, A. G. Gpytorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems* (2018).
61. Paszke, A. *et al.* Automatic differentiation in PyTorch. In *NIPS-W* (2017).
62. Wilson, A. & Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, 1067–1075 (2013).
63. Snelson, E., Ghahramani, Z. & Rasmussen, C. E. Warped gaussian processes. In *Advances in neural information processing systems*, 337–344 (2004).
64. Jensen, B. S., Nielsen, J. B. & Larsen, J. Bounded gaussian process regression. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6 (IEEE, 2013).
65. Vovk, V. Kernel ridge regression. In *Empirical inference*, 105–116 (Springer, 2013).
66. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).
67. Nikitin, A. Data and Source Code Repository. https://github.com/tzoiker/psqi.

68. Titsias, M. Variational learning of inducing variables in sparse Gaussian Processes. In *Artificial intelligence and statistics*, 567–574 (PMLR, 2009).
69. Wilson, A. G., Hu, Z., Salakhutdinov, R. & Xing, E. P. Stochastic variational deep kernel learning. arXiv preprint arXiv:1611.00336 (2016).
70. Pleiss, G., Gardner, J., Weinberger, K. & Wilson, A. G. Constant-time predictive distributions for Gaussian processes. In *International Conference on Machine Learning*, 4114–4123 (PMLR, 2018).
71. Wang, K. *et al.* Exact Gaussian Processes on a million data points. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 32, 14648–14659 (Curran Associates, Inc., 2019).
72. Burri, N. M., Weatherl, R., Moeck, C. & Schirmer, M. A review of threats to groundwater quality in the anthropocene. *Sci. Total Environ.* **684**, 136–154 (2019).
73. Tikhomirov, V. V. *Hydrogeochemistry Fundamentals and Advances, Groundwater Composition and Chemistry* Vol. 1 (Wiley, 2016).
74. Zhang, M. *et al.* Distributions and origins of nitrate, nitrite, and ammonium in various aquifers in an urbanized coastal area, south China. *J. Hydrol.* **582**, 124528 (2020).
75. Hansen, B., Thorling, L., Schullehner, J., Termansen, M. & Dalgaard, T. Groundwater nitrate response to sustainable nitrogen management. *Sci. Rep.* **7**, 1–12 (2017).
76. Lee, M.-S., Lee, K.-K., Hyun, Y., Clement, T. P. & Hamilton, D. Nitrogen transformation and transport modeling in groundwater aquifers. *Ecol. Modell.* **192**, 143–159 (2006).
77. Sajedi-Hosseini, F. *et al.* A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Sci. Total Environ.* **644**, 954–962 (2018).
78. Adimalla, N., Li, P. & Venkatayogi, S. Hydrogeochemical evaluation of groundwater quality for drinking and irrigation purposes and integrated interpretation with Water Quality Index studies. *Environ. Processes* **5**, 363–383 (2018).
79. Boateng, T. K., Opoku, F., Acquaah, S. O. & Akoto, O. Groundwater quality assessment using statistical approach and Water Quality Index in Ejisu-Juaben Municipality Ghana. *Environ. Earth Sci.* **75**, 489 (2016).
80. Ramakrishnaiah, C., Sadashivaiah, C. & Ranganna, G. Assessment of Water Quality Index for the groundwater in Tumkur Taluk, Karnataka State India. *J. Chem.* **6**, 523–530 (2009).
81. Kabata-Pendias, A. *Trace elements in soils and plants* (CRC press, 2000).
82. Reynolds, W., Drury, C., Yang, X. & Tan, C. Optimal soil physical quality inferred through structural regression and parameter interactions. *Geoderma* **146**, 466–474 (2008).
83. Shadrin, D. *et al.* An automated approach to groundwater quality monitoring-geospatial mapping based on combined application of gaussian process regression and Bayesian information criterion. *Water* **13**, 400 (2021).
84. Choudhary, K., Boori, M. S. & Kupriyanov, A. V. Mapping and evaluating urban density patterns in Moscow Russia. *Comput. Opt.* **41**, 528–534 (2017).
85. Vasenev, V., Stoorvogel, J., Leemans, R., Valentini, R. & Hajiaghayeva, R. Projection of urban expansion and related changes in soil carbon stocks in the Moscow Region. *J. Cleaner Prod.* **170**, 902–914 (2018).
86. Klimanova, O., Kolbowsky, E. & Illarionova, O. *Impacts of urbanization on green infrastructure ecosystem services: the case study of post-soviet Moscow* (Belgeo, Revue belge de géographie, 2018).

Acknowledgements

Presented research was funded by Russian Science Foundation (project No. 20-74-10102).

Author contributions

A.N., D.S., M.P., P.T. contributed to the design of the study. A.N. developed modeling concept, performed calculations, visualized results. S.M. and I.O. provided computational facilities. M.P. provided the dataset. A.N., P.T., D.S. reviewed model outputs and contributed to their interpretation. A.N., P.T., D.S., M.P. wrote original draft. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.