Mathematical Geosciences
https://doi.org/10.1007/s11004-026-10307-8

Nyquist–Shannon Guidance for Linking Sampling Adequacy and Predictive Uncertainty in Geospatial Mapping

Meng Lu[1] · **Jiong Wang**[2]

Received: 5 November 2025 / Accepted: 10 May 2026
© The Author(s) 2026

Abstract

Modeling geospatial patterns of environmental variables at increasingly fine resolutions and broader scales highlights the reliability of map products built from sparse samples as a critical issue in scientific research and practical applications. It is desirable to attribute predictive uncertainty to identifiable sources, namely, sampling density, data noise, and model limitations. To approach this, we examine whether a spatial process imposes a minimum achievable predictive variance as a function of sampling density. We first present a transition from the geometric/spatial domain to the spectral domain, linking spatial autocorrelation to bandwidth frequency through the spectrum of the Gaussian radial basis function (RBF) kernel. We then apply the Nyquist–Shannon sampling theorem to link sampling rate and predictive uncertainty. The Gaussian RBF kernel is used as a demonstration vehicle instead of a restrictive assumption. We derive the effective bandwidth as a function of the kernel length scale and subsequently generalize this to a tolerance-based prediction error floor that quantifies unresolved high-frequency variance captured by the kernel for a given sampling rate. A practical implication is that sampling density induces fundamental, model-independent limits on achievable accuracy and resolution. Furthermore, this work provides a quantitative interpretation of kernel length scale, linking predictive variance to resolvable spectral energy, and provides guidance on sampling sufficiency, resolution limits, and uncertainty interpretation. Using both simulated fields and remote sensing imagery as real-world data, one- and two-dimensional experiments with and without covariates show that the Nyquist-based limits provide practical guidance for assessing sampling adequacy and quantifying predictive uncertainty in geospatial modeling.

✉ Meng Lu
meng.lu@uni-bayreuth.de

Jiong Wang
j.wang-4@utwente.nl

[1] Geoinformatics – Spatial Big Data Research Group, Faculty of Biology, Chemistry and Earth Science, University of Bayreuth, Universitätsstraße 30, 95440 Bayreuth, Germany

[2] Department of Urban and Regional Planning and Geo-Information Management, ITC, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

Published online: 15 June 2026

Keywords Geostatistics · Nyquist–Shannon theorem · Sampling density · Predictive uncertainty · Geospatial modeling · Spectral analysis

1 Introduction

Probabilistic inference at unsampled spatial locations, widely applied in domains such as environmental science and geohealth, is an important approach for producing interpolations and maps that serve as information infrastructure for subsequent geographical analyses and decisions. Point measurements, such as noise, weather, climate, air quality, and soil-property observations, are among the most common geoscientific data types (Moraga 2023). Point-to-point interpolation generalizes to point-to-areal or areal-to-areal inference via change-of-support methods (Cressie 1989; Isaaks and Srivastava 1990).

Quantifying uncertainty is integral to probabilistic inference. Uncertainty arises from both data and models (Turley et al. 2009). Careful treatment of these sources is crucial for model development and interpretation, and several studies provide frameworks to identify and quantify modeling uncertainty (e.g., Wallentin and Car 2013). In practice, uncertainty from data and prediction models is quantified and integrated (Zhu et al. 2025). In hierarchical modeling frameworks (Cressie and Wikle 2011), the uncertainty from the process is explicitly represented, which differs from more classical statistics, where models are specified only for data and parameters. Predictive uncertainty refers to uncertainty in the model's probabilistic prediction (Tyralis 2024); it integrates multiple sources of uncertainty through the modeling hierarchy and closely corresponds to Zhu et al.'s (2025) notion of overall uncertainty. Recent probabilistic frameworks extend to machine learning, with Gaussian processes and Bayesian neural network among the most widely applied methods (Iwata et al. 2017; Biggio et al. 2021; Gonçalves and Wellmann 2025; Kwon et al. 2020).

Predictive uncertainty is not necessarily summarized by a single number, as it is scale-dependent: a model may estimate a broad, low-frequency structure with high confidence while remaining highly uncertain about high-frequency components such as local fluctuations. In addition, low uncertainty in a model does not necessarily indicate model reliability, but rather reflects a level of confidence given the model assumption and data: If the sample locations are too far apart relative to a spatial process with high-frequency variation, an "incorrect" assumption of a large range parameter would fail to encode uncertainty associated with unresolved high-frequency variation in the prediction.

Uncertainty propagation does not necessarily allow predictive uncertainty to be attributed to, for example, data limitations versus model structure. Uncertainty attribution requires clearly specified modeling assumptions and asymptotic regimes. For example, not all covariance parameters are consistently estimable under fixed-domain asymptotics (Stein 1999, p. 11 and Sect. 4.3; Zhang 2004). In addition, in practice, replicated measurements are rarely available to distinguish microscale variability from measurement error (Cressie 2015). In geostatistics, the microscale component of the nugget reflects unresolved variability induced by insufficient sampling resolution. However, decomposition of the nugget into measurement error, intrinsic microscale

variability, and sampling-induced effects is generally infeasible without replicated measurements. One example is the hierarchical model in Shaddick et al. (2018), which decomposed posterior predictive variance into expected conditional predictive variance, that is, process uncertainty given parameters, and parameter uncertainty, but could not isolate measurement noise.

In sampling design, where the placement of samples is to be optimized (e.g., Rogerson 2004; Senarathne et al. 2023), the relationship between sampling density and predictive uncertainty remains unresolved. Design-based criteria (Grafström and Schelin 2014) compute uncertainty from known inclusion probabilities and do not isolate sampling-induced uncertainty without additional diagnostics. Consequently, the relationship between sampling density and predictive uncertainty, conditional on the underlying spatial process, remains understudied. As a result, it is difficult to determine whether the predictive uncertainty arises from under-sampling or from the limitations of the model itself.

Sensitivity analysis or scenario generation compares the uncertainty quantified under alternative models or inputs. For example, Qin et al. (2013) propagated data uncertainty through a model sensitivity analysis. Pereira et al. (2024) used a jackknife approach to generate scenarios by sequentially removing one observation at a time and subsequently applied Bayesian model averaging to compute weighted uncertainty envelopes. However, this kind of method is specific to a given set of sample locations and a given target process.

Sampling density can be viewed as one component of the broader problem of sample representation. In digital soil mapping and related spatial prediction studies, representativeness has been evaluated through geographical coverage, environmental covariate-space coverage, and environmental similarity between sampled and unsampled locations (Zhu et al. 2015, 2018). Related work has identified representative samples using fuzzy membership to environmental cluster centroids (An et al. 2018) and quantified sample-set representativeness by comparing the covariate distribution of sampled locations with that across the full mapping area (Zhang and Zhu 2019). Zhu et al. (2026) further investigated uncertainty quantification methods that integrate similarity measures between samples and prediction locations. These studies show that sample representation is closely related to predictive uncertainty, but they do not establish a general relationship between sampling density, the spatial scale of the target process, and the minimum achievable predictive variance.

We focus on the sampling density component of this broader problem by examining how sampling density relates to predictive uncertainty in geospatial modeling. Specifically, we ask whether, for a given spatial process, sampling density imposes a minimum achievable predictive variance that cannot be reduced by improved statistical inference alone. This research question lays the foundation for further investigation of the relationship among sampling density, the spatial structure of the target process, predictive uncertainty, and restorable spatial scale in geospatial modeling. One main challenge in linking all these factors is to establish an explicit relationship between sample configuration and the amount of information that can be captured from the underlying spatial process. Our approach is inspired by sampling theory, which relates the highest recoverable frequency to the sampling frequency. The connection between

the spatial and frequency domains can be established through spectral analysis of the covariance function.

Linking sampling distribution and process recovery through the lens of frequency naturally connects to the Nyquist–Shannon sampling theorem (Nyquist 1928; Shannon 1949; Unser et al. 2000), which was originally discovered for converting between an analog signal and a sequence of numbers, and provides a fundamental link between sampling rate and recoverable frequency content. It states that a continuous, band-limited signal can be fully reconstructed from discrete samples only if the sampling rate exceeds twice the maximum frequency present in the signal. Our rationale is that if a mathematical link could be established between the kernel length scale and effective bandwidth, the Nyquist–Shannon sampling theorem (hereafter referred to as the Nyquist principle) could be applied to derive a relationship between the recoverable signal and the sampling rate. This connection enables us to express the relationship between sampling density and the lower bound on predictive variance. It further allows us to map a user-chosen tolerance and covariance structure, such as the kernel length scale or variogram range, to effective bandwidth.

Given a true or estimated spatial or feature-space process, this provides (1) the required sampling rate for recovering a specified amount of signal energy, and (2) the prediction error floor, that is, the lower bound on prediction error, defined as the unrecovered energy of the process, given a sampling rate. As a baseline, we focus on regular sampling situations. We interpret the error floor as a diagnostic complement to conventional uncertainty metrics, indicating whether prediction and estimation errors arise primarily from sampling limitations or model capability.

The link between length scale l and the effective bandwidth under the Nyquist principal, f_{eff}, has several key implications (I1–I3):

1.1 I1 Direct Nyquist Guidance from the Kernel Length Scale

It bridges familiar geostatistical scale parameters, such as the variogram range, and frequency-domain sampling theory.

1.2 I2 Closed-Form, Model-Based Link from Length Scale to Frequency

If l is estimated from prior knowledge or model fitting, its posterior can be used to derive an effective boundary frequency and implied sampling-rate recommendations. In addition, the closed form allows us to choose tolerance levels depending on application needs and derive associated sampling rate or error floors, aligning with the view that robust data analysis requires critical synthesis of domain semantics and statistical assumptions (Scheider et al. 2017).

1.3 I3 Diagnostic Understanding of Uncertainty

Considering the error floor together with the root mean square error (RMSE), obtained from cross-validation or test data, and other error or uncertainty quantities can support

the interpretation of model performance and the identification of the uncertainty or error sources. It also supports a scale-aware interpretation of sampling adequacy.

The applicability of these implications is demonstrated in the analysis, and I3 is further elaborated in discussion. The objectives are as follows:

1. Link radial basis function (RBF) kernel length scale to effective bandwidth in the Nyquist–Shannon sampling theorem.
2. Quantify the prediction error floor.
   Derive the minimum achievable prediction error, or "error floors," caused by unresolved variance in the latent process modeled as a Gaussian process (GP) with an RBF kernel.
3. Test how the Nyquist principle guides sampling rules under regular sampling.
   Show how the effective bandwidth and the associated error floors relate to the sampling rate needed for reliable prediction and estimation of kernel parameters, in particular the process length scale.

2 Background

Statistical and geostatistical methods, which learn patterns from data, have enabled fine-resolution, large-area mapping across domains. Examples include malaria risk (Morlighem et al. 2023), air pollution concentration (Lu et al. 2020b, 2023; Shen et al. 2022), traffic emissions (Yang et al. 2019), air temperature (Meyer et al. 2016), land surface temperature (Zhang et al. 2021), urban dynamics (Li et al. 2021), and human population distribution (Linard et al. 2012).

As the relationship between predictor variables, also referred to as covariates or features, and the target variable is commonly unknown, sampling adequacy is critical. The consequences of inadequate sampling are twofold: (1) important spatial variation may be missed, and (2) the uncertainty estimates may become biased, with poor ability to capture rare or extreme phenomena. To illustrate, consider a one-dimensional GP with zero mean and a squared-exponential (or Gaussian RBF) kernel (Eq. 1), with variance $\sigma_f{}^2 = 2$ and length scale $l = 1$ over the interval $[-5, 15]$. We fit a GP to 10 random samples out of 500 simulated locations (Fig. 1) and observe that the prediction interval broadens in data-poor regions while the true extremes remain uncovered. The mismatch echoes a classical question: how much information can be obtained from sparse measurements? The RBF kernel provides a spectral view of spatial autocorrelation, allowing us to apply the Nyquist principle to answer this question.

The Nyquist principle states that, for a band-limited signal, accurate reconstruction requires a sampling rate, or Nyquist rate, f_{Nyq_rate}, at least twice the signal's effective bandwidth, or maximum frequency $f_{\max}$; equivalently, the maximum recoverable frequency, or Nyquist frequency f_{Nyq} is half the sampling rate (Jerri 1977; Oppenheim and Schafer 1998). While originally formulated for temporal signals, the rate–bandwidth logic extends to spatial signals. Prior work has applied it to image reconstruction and compression (Jerri 1977), sampling sufficiency of the human brain electrical field (Srinivasan et al. 1998), and information compression and quality control in remote

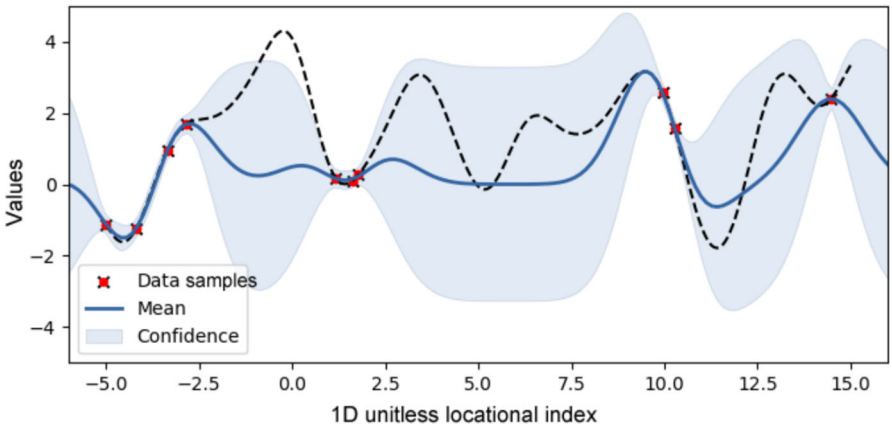


Fig. 1 A one-dimensional ground-truth pattern, shown as the dashed line, was simulated using a Gaussian RBF kernel with zero mean, variance $\sigma_f^2 = 2$, and length scale $1 = 1$. Ten training points (red crosses) are used to fit the GP. The posterior mean and prediction interval (blue band) show how inadequate sampling misses key variations and produces predictive uncertainty that is broad in data-poor regions but fails to capture extremes

sensing (Qin et al. 2016; Garma et al. 2017), indicating its relevance for discretely sampled spatial fields. More recently, Dong et al. (2024) developed filtered kriging and proved that when measurements satisfy the Nyquist principle, perfect reconstruction is achieved from filtered Kriging interpolation under certain assumptions. Miao and Wang (2025) used the Nyquist principle to determine the number of basis functions to be truncated in a sparse spectrum representation of the covariance matrix when modeling nonstationary processes.

Because the Gaussian RBF spectrum is not band-limited, the Nyquist principle is not directly applicable. We adopt a tolerance-based effective cutoff for the Gaussian RBF spectrum. In the Hilbert space L_2 setting, ideal low-pass filtering is equivalent to orthogonal projection onto a band-limited subspace, which yields the minimum-error band-limited approximation (Unser 2000). Under this perspective, the discarded spectral tail of the Gaussian RBF spectrum constitutes the approximation error relative to the original space, while the retained band-limited component determines the sampling requirement.

3 Method

3.1 RBF Spectrum and the Irreducible Error

In this study, we focus on the squared-exponential (Gaussian RBF) kernel, whose length scale l controls how rapidly correlations decay with distance. A small l means rapid fluctuations, whereas a large l corresponds to smooth, slowly varying fields

$$K(\boldsymbol{r}) = \sigma_f^2 \exp\left(-\frac{r^2}{2l^2}\right), \tag{1}$$

where r is the spatial distance, ${\sigma_f}^2$ is the variance (amplitude), and l is the length scale (controls how fast correlations decay). This means that the kernel assumes that correlations decay smoothly with distance.

The relationship between l and frequency becomes clear through the spectral representation of a kernel (Stein 1999, p. 24). By Bochner's theorem, $K(\boldsymbol{r})$ is a valid stationary covariance function if and only if it can be represented as the Fourier transform of a positive finite spectral measure F. If F has a density with respect to the Lebesgue measure, the density is called the spectral density, denoted by $S(\omega)$ (Stein 1999, pp. 24–25)

$$K(\boldsymbol{r}) = \int_{\mathbb{R}^d} e^{i\omega t} S(\omega) \mathrm{d}\omega. \tag{2}$$

This shows that every kernel corresponds to a spectrum, that is, variance distributed over spatial frequencies, where $S(\omega)$ is the amount of variance per unit frequency.

For the Gaussian RBF kernel, the spectral density is proportional to an exponential function

$$S(\omega) \propto \exp\left(-\frac{l^2\omega^2}{2}\right). \tag{3}$$

The Gaussian form implies that the process has no hard frequency cutoff; that is, its spectral energy decays smoothly. The kernel's width in the spatial domain (l) is inversely related to its width in the frequency domain. A large l concentrates energy at lower frequencies (narrow spectrum), and vice versa. Thus, l provides a direct bridge between the spatial kernel formulation of Gaussian processes (GP) and the Nyquist sampling criterion.

Because this spectrum is unbounded, we define an effective angular bandwidth $\Omega_\varepsilon(l)$ that captures $(1-\varepsilon)$ of the total spectral energy, which is equivalent to $(1-\varepsilon)$ of the kernel variance. We consider the one-dimensional case, which suffices to illustrate the derivation of the relationship between the kernel length scale and effective bandwidth.

$$\int_{-\Omega_\varepsilon(\ell)}^{+\Omega_\varepsilon(\ell)} S(\omega)\mathrm{d}\omega = (1-\varepsilon) \int_{-\infty}^{+\infty} S(\omega)\mathrm{d}\omega. \tag{4}$$

As the exponential function is even, that is, $A\int_{-\infty}^{+\infty} \exp\left(-\frac{l^2\omega^2}{2}\right)\mathrm{d}\omega = 2A\int_{0}^{+\infty} \exp\left(-\frac{l^2\omega^2}{2}\right)\mathrm{d}\omega$, where A is a constant, we can obtain

$$\Omega_\varepsilon(l) = \frac{\sqrt{2}}{l}\,\mathrm{erf}^{-1}(1-\varepsilon), \tag{5}$$

where $\mathrm{erf}^{-1}(x)$ is the inverse error function $\mathrm{erf}(x)$. The $\mathrm{erf}(x)$ is a special function that arises when integrating the Gaussian distribution. The RBF kernel has a closed form via $\mathrm{erf}^{-1}(x)$. For example, if $\varepsilon = 0.1$, $\mathrm{erf}^{-1}(0.9) \approx 1.163$. If $\varepsilon = 0.01$, $\mathrm{erf}^{-1}(0.99) \approx 1.821$.

This means that $\Omega_\varepsilon(l)$ defines the effective angular bandwidth that is inversely proportional to the l with a tolerance-dependent constant. With Eq. 6, we can calculate it at different tolerance levels; for example, $\Omega_{\varepsilon=0.01} = 2.547/l$. Expressed in cycles per unit length,

$$\begin{aligned} f_\varepsilon &= \Omega_\varepsilon/2\pi \\ &= \frac{\sqrt{2}}{2\pi}\mathrm{erf}^{-1}(1-\varepsilon)/l \\ &= c(\varepsilon)/l. \end{aligned} \tag{6}$$

At tolerance levels of 0.01 and 0.1, $c(0.01) \approx 0.41$, and $\mathrm{c}(0.1) \approx 0.26$; $f_{\varepsilon=0.01} \approx 0.41/l$; $f_{\varepsilon=0.1} \approx 0.26/l$. The effective bandwidth, f_{eff}, is therefore from $\frac{0.26}{l}$ to $\frac{0.41}{l}$. This result can be validated with fast Fourier transformation (FFT).

The tolerance ε is the unresolved variance fraction (tail energy) of the kernel's spectrum. If the process variance is σ^2, the minimum achievable prediction error of the GP is bounded by $\varepsilon\sigma^2$. The statistical interpretation is direct: the tail variance, $\varepsilon\sigma^2$, is unrecoverable even when sampling at the Nyquist rate; it sets a prediction error floor. The $\varepsilon\sigma^2$, which corresponds to mean squared error (MSE), represents the irreducible contribution of unresolved high-frequency components to the prediction error. Note that for kernels that do not have a closed form for $\Omega_\varepsilon(\ell)$, it can be calculated numerically.

3.2 Relating RBF Kernel to Sampling Theorem

The RBF kernel provides a direct link between spatial correlation and the frequency content of a geospatial process. In principle, once the kernel's spectral form is known, (1) given a sampling rate ($f_s = 1/\Delta x$), we can express how much variance remains unresolved, which is the MSE ($\varepsilon\sigma^2$); and (2) given a target error tolerance ε, we can solve for the sampling rate that achieves it. Note that, in practice, ℓ and σ^2 are commonly estimated with prior knowledge or from samples.

3.2.1 Unresolved Variance at a Given Sampling Rate

Let $f_s = 1/\Delta x$, where Δx is the sampling interval, and the Nyquist frequency is

$$f_{Nyq} = f_s/2. \tag{7}$$

The Nyquist angular frequency (which is a property of the signal) is

$$\Omega_{Nyq} = 2\pi f_{Nyq} = \frac{\pi}{\Delta x}. \tag{8}$$

This means that angular frequencies larger than Ω_{Nyq} can be aliased.

For the Gaussian RBF spectrum, from Eq. (5) we can obtain the unresolved variance at spacing Δx

$$\varepsilon(l, \Delta x) = 1 - \mathrm{erf}\Big(l\pi/\sqrt{2}\Delta x\Big). \tag{9}$$

This means that ε depends on the ratio between l and Δx, and coarser sampling intervals (larger Δx) leave more (high-frequency) variance unresolved.

3.2.2 Required Sampling Rate for a Given Tolerance

Conversely, for a target tolerance ε, the Nyquist rate ($f_{Nyq_rate} = \frac{1}{\Delta x_{Nyq_rate}} \geq 2 f_{\max}$) implies

$$\Delta x_{Nyq_rate} \leq \left(\frac{\pi}{\Omega_\varepsilon(\ell)}\right) \Rightarrow \Delta x_{Nyq_rate} \leq \frac{\pi l}{\sqrt{2}\mathrm{erf}^{-1}(1-\varepsilon)}. \tag{10}$$

Over a one-dimensional domain of length L, this corresponds to sample size

$$s \geq \frac{L}{\Delta x_{Nyq_rate}} \Rightarrow s \geq \frac{\sqrt{2}L}{\pi\ell}\mathrm{erf}^{-1}(1-\varepsilon). \tag{11}$$

3.2.3 General Form

More generally, the ε can be expressed as

$$\varepsilon(l, \Omega_c) = 1 - \ \mathrm{erf}\big(l\Omega_c/\sqrt{2}\big), \tag{12}$$

where the Ω_c is the angular frequency we can capture (e.g., the angular Nyquist frequency Ω_{Nyq}, or any cutoff we would like to set).

These relations bridge the kernel view (length scale l) and the sampling view (frequency), showing that sampling sufficiency can be understood as capturing the dominant frequencies implied by the kernel. Equations (9) and (12) quantify the unresolved variance at a given sampling rate. This yields a bound on the fraction of variance any estimator can recover under the assumed spectrum (kernel) and the given sampling interval. In practice, the true process length scale l_0 is commonly unknown. If we estimate l from finite samples, when sampling is sparse, the process appears smoother than it truly is, biasing $\widehat{l}$ upward. The overestimated $\widehat{l}$ reduces $\varepsilon\big(\widehat{l}, \Delta x\big)$ in Eq. (9), underestimating the error floor and creating the impression of adequacy.

The derivations are exact for the RBF kernel, whose spectrum is Gaussian. Otherwise, the spectral cumulative distribution function (CDF) of a kernel (e.g., Matérn) is required. In multidimensional cases, Eqs. (10) and (11) can be applied to each axis. For irregular sampling (e.g., in feature space), Δx should be replaced by an effective sampling distance (Bretthorst 2001), which we leave for future study.

3.3 Validation

The theoretical relationships developed in this section are validated through four experiments, comprising three simulation studies and one real-world experiment. The validation focuses on three aspects: (1) the relationship between length scale and spectral energy, (2) the relationship between sampling rate and the derived error floor, and (3) the recovery of kernel hyperparameters as a function of sampling rate. In Experiments 1 to 3, data are generated from spatial processes with known kernels, enabling direct validation of theoretical predictions. Experiment 4 extends to satellite imagery and with covariates, where performance is evaluated using absolute error, entropy of the latent process, and RMSE between the posterior mean prediction and the ground truth. Detailed experimental designs are presented in Sect. 4.1.

4 Experiment

To clearly demonstrate how our objectives are achieved, we formulate two tests: Test 1, error floor validation, where we validate that the observed prediction error is bounded by the spectral error floor implied by the sampling rate; and Test 2, length scale identifiability, in which we quantify the sampling interval and sample size needed to recover the process length scale. The experimental designs for these tests are detailed below.

4.1 Experimental Design

To incrementally explore the sampling theorem for geospatial point-measurement regression modeling, we proceed in four steps, with steps 1 to 3 using one-dimensional simulations and step 4 using two-dimensional satellite imagery:

1. FFT validation of the relationship between length scale and frequency (referred to as the "$l - f_{\text{eff}}$ link");
2. Test 1, error floor validation as a function of sampling rate;
3. Test 2, sampling rate for length scale estimation and prediction errors;
4. Extension of Test 2 using real-world data, in two dimensions and with covariates.

For the simulation, we adopt the setup illustrated in Fig. 1: a one-dimensional index $[-5, 15]$ of length $L = 20$, which can be interpreted as arbitrary geographical or temporal locations or as a covariate (feature) axis. Along the index, ground-truth realizations are generated on a fine grid and then subsampled at varying sampling

rates. For consistency, we use the spacing ratio

$$r = \Delta x / l_0, \tag{13}$$

where l_0 is the true process length scale. Using Eq. (10), and using the values of $f_{\varepsilon=0.01}$ and $f_{\varepsilon=0.1}$, the tolerance-based Nyquist spacing ratios are

$$r_{nyq}(0.1) \approx 1.91, \quad r_{nyq}(0.05) \approx 1.60, \quad r_{nyq}(0.01) \approx 1.22.$$

Thus, for example, when the spacing ratio $r \approx 1.22$, 99% of the kernel's spectral energy is retained.

4.1.1 FFT Validation of the $l - f_{\text{eff}}$ Link

We fix $l_0 = 1$ and variance $\sigma^2 = 2$ in the Gaussian RBF kernel and simulate 500 points with a tiny jitter (1e−6 for numerical stability) for the ground-truth pattern. We generate 100 realizations on the fine grid and compute the power spectrum of FFT after mean removal and Hann windowing. For $\varepsilon \in \{0.01, 0.05, 0.1\}$, that is, the effective bandwidth at the retain levels 99%, 95%, and 90%, we locate the cutoff where the cumulative spectral energy reaches $1 - \varepsilon$ and report cycles per unit length (mean ± standard deviation [SD]) across realizations. Then we compare these bandwidths to the derived $f_{\text{eff}}(l_0)$.

4.1.2 Test 1

The same simulation setting as in 4.1.1 is used to simulate 100 realizations. We test the spacing ratio $r = \{0.2, 0.5, 0.9, 1.2, 1.6, 1.9, 2.5, 3.0, 4.0, 5.0, 9.0\}$. For each r, we fit a GP regression model to the subsampled data and compute the root mean square error (RMSE) of the posterior mean against the ground truth. To disentangle sampling spacing from hyperparameter estimation, we compare the RMSE to three error floors:

Benchmark floor $\sigma\sqrt{\varepsilon(l_0, \Delta x)}$ (true l_0, σ).

Marginal floor $\sigma\sqrt{\varepsilon\left(\hat{l}, \Delta x\right)}$ (fitted $\hat{l}$, true σ).

Practical floor $\hat{\sigma}\sqrt{\varepsilon\left(\hat{l}, \Delta x\right)}$ (fitted $\hat{l}$, $\hat{\sigma}$).

4.1.3 Test 2

We simulated 500,000 realizations for different combinations of length scale $l \in (0, 5]$ and variance $\sigma_f{}^2 \in (0, 9]$ and examined the relationship between the sample size s and the error in l estimation $E_l = |\widehat{l} - l_0|$. We focused on l instead of the predictive error between the simulated and fitted signals because the length scale directly reflects the dominant frequency of the process, whereas RMSE is confounded by the variance of the realizations.

4.1.4 Real-World Experiment

The Landsat land surface temperature (LST) level 2 (L2) product (Sayler 2023) covers the city of Amsterdam (path 198, row 24), the Netherlands, obtained via the satellite Thermal Infrared Sensor (TIRS) onboard Landsat-8 on July 18, 2021, at 10:33 local time. Data are resampled to 30 m pixels. There are many factors contributing to the observed LST pattern at a certain time point, and the main factors are the angle of observation, solar radiation, and land surface material with albedo and specific heat capacity as important features, and the geometry of land surface elements such as buildings that can be observed partially with their roofs and facades (Voogt and Oke 2003). Within one pixel, depending on the resolution, different elements are involved, making the interpretation of the mixed information potentially difficult. However, when zoomed out with a sufficiently large scale (e.g., the scale of the entire city of Amsterdam), the discrete measurement resembles continuous patterns of the LST.

For the one-dimensional scenario, as denoted by the red dashed line in Fig. 2, a subset of the LST data obtained along the central transect of the LST map consisting of 576 pixels as an LST profile is used to examine the number of samples needed to recover the one-dimensional transect as the ground-truth pattern. The one-dimensional test can be considered a replication of the test conducted on the simulated data. The two-dimensional test is intuitive for real-world mapping activities. As highlighted by

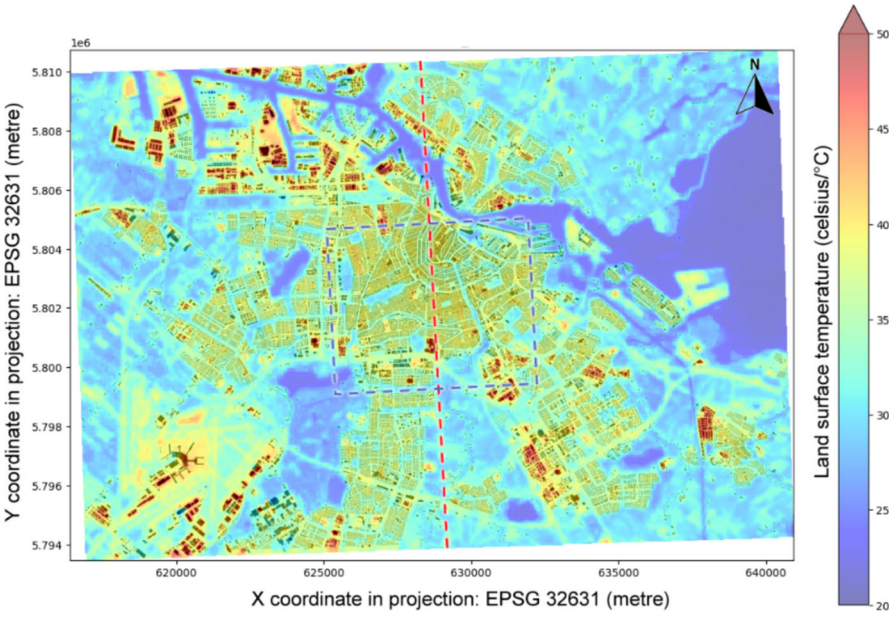


Fig. 2 Land surface temperature (LST) over the city of Amsterdam. The transect for the one-dimensional real-world experiment and the bounding box for the two-dimensional real-world experiment are denoted by the red dashed line and blue dashed box, respectively

the blue dashed box in Fig. 2, a two-dimensional subset of 230 × 180 pixels is taken for the two-dimensional scenario.

4.2 Model Configuration and Modeling

The GP model provides the flexibility of machine learning models and quantifies uncertainty based on prior knowledge and variance in the observations. The GP regression defines patterns or function values f along a domain (e.g., a spatial domain) indexed by $\boldsymbol{x}$ as

$$f(\boldsymbol{x}) \sim \mathrm{GP}\big(m(\boldsymbol{x}), k\big(\boldsymbol{x}, \boldsymbol{x}'\big)\big), \tag{14}$$

where GP stands for Gaussian process to emphasize that the function values f are considered a random process with a mean function m and a kernel k specifying the joint Gaussian distribution of the function values. In particular, the kernel encodes the prior distribution over the function values and is updated into the posterior distribution of the function value (Rasmussen and Williams 2006).

In geostatistics, $\boldsymbol{x}$ commonly encodes geographical locations or coordinates. While one can choose from a large set of potential kernels k, the form of k is highly context-dependent (MacKay 1998; Rasmussen and Williams 2006). In the case of an isotropic process, we can use a single length scale, for example, using the Gaussian RBF kernel (Eq. 1). More generally, the latent process includes noise

$$K\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sigma_f^2 \exp\left(-\frac{d\left(\boldsymbol{x}, \boldsymbol{x}'\right)^2}{2l^2}\right) + \sigma_n^2 \delta, \tag{15}$$

where σ_f, l are defined in Eq. (1), δ is the Kronecker delta, and ${\sigma_n}^2$ is the noise variance. The noise variance describes the degree to which the function f constitutes noisy observations across the domain indexed by $\mathbf{x}$, which, in geostatistics, corresponds to geographical location. These parameters correspond to the geostatistical modeling variogram parameters range, sill, and nugget. Specifically, l corresponds to the range parameter (operational scale), that is, the scale of the observation needed to effectively characterize changes across space. The process variance ${\sigma_f}^2$ corresponds to the sill, and ${\sigma_n}^2$ to the nugget.

To account for possible anisotropy in LST, we adopted an anisotropic Gaussian RBF kernel with automatic relevance determination (ARD; Rasmussen and Williams 2006) for the two-dimensional real-world case (Test 2)

$$K\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sigma_f^2 \exp\left(-\sum_{h=1}^{d} \frac{d(\boldsymbol{x}_h, \boldsymbol{x}_{h'})^2}{2l_h^2}\right) + \sigma_n^2 \delta, \tag{16}$$

where h denotes the input dimension and l_h is the dimension-specific length scale. In this study, $d = 2$, corresponding to the east–west and north–south dimensions.

A covariate can be considered as a set of function values in addition to the target f and can be modeled jointly with the target. In our study, we use a separable multi-task GP (MTGP; Bonilla 2008)

$$f(\boldsymbol{x}) \sim \mathrm{GP}\Big(m(\boldsymbol{x}),\, K^{f} K^{\boldsymbol{x}}\Big), \tag{17}$$

where the kernel K^{f} represents inter-functional (inter-task) covariance, and $K^{\boldsymbol{x}}$ represents covariance across input locations within each task. The MTGP has been applied to predict LST using LST at different time points (Wang et al. 2016). In geostatistics, the formalization links to cokriging. The multiple geospatial processes involved may not share the same mean, but the processes can be reduced to zero mean with additional assumptions. For interpretability, we removed the mean of each task (target or covariate); that is, the kernel operates on the residuals.

The parameters estimated by applying the GP model to the full image are considered the true parameters. The covariate is constructed as a noisy transform of the original LST field (with additive Gaussian noise chosen to achieve an R-squared of around 0.8), ensuring interpretability by reducing uncertainty introduced by unknown variation or noise. We randomly split the samples so that half of them are sampled from the covariate and the other half from the target variable, expecting that the information sampled from a proper covariate can reasonably complement the information sampled from the target variable.

5 Results and Discussion

5.1 One-Dimensional Simulated Data

5.1.1 Validation of the l - f_{eff} Link

Across 100 GP realizations, the FFT-based effective bandwidths (cycles per unit length; mean $\pm$ SD) closely match the theoretically derived f_{eff} values shown in parentheses: 0.410 ± 0.061 (0.409) at 99% retained spectral energy ($\varepsilon = 0.01$); 0.321 ± 0.056 (0.312) at 95% ($\varepsilon = 0.05$); 0.270 ± 0.057 (0.262) at 90% ($\varepsilon = 0.10$).

5.1.2 Test 1: Error Floor and Sampling Rate

Figure 3 compares observed GP interpolation errors with different theoretical error floors across 100 realizations. The blue curve shows the mean observed RMSE, with the gray band marking ± 1 standard deviation across realizations. The "benchmark floor" systematically lies below the observed RMSE. The gap quantifies the inaccuracy of parameter estimation and the use of finite samples.

The "marginal floor" closely follows the observed RMSE at a spacing ratio below 1.9. It slightly overestimates RMSE due to bias in the estimated length scale, the uncertainty of which becomes large from $r = 1.6$ onward (Fig. 4). If the model suggests a large length scale, the theoretical unresolved variance fraction is close to zero,

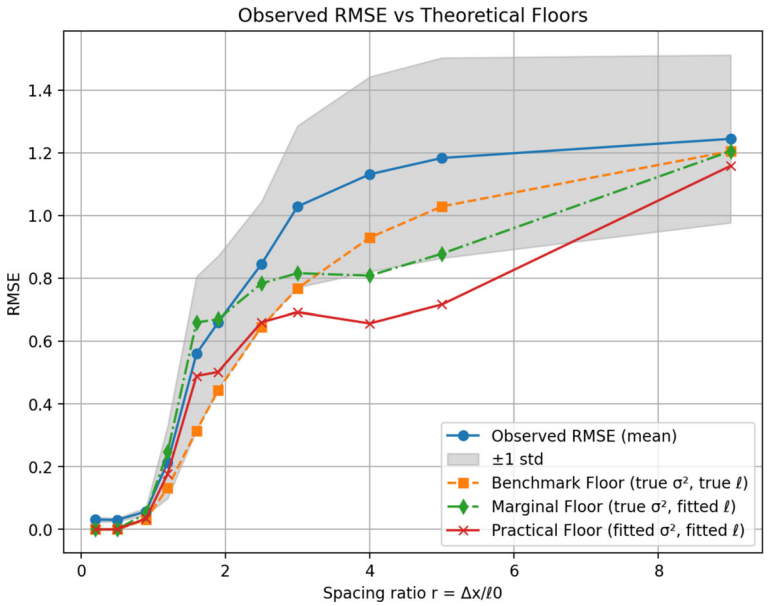


Fig. 3 Observed RMSE and error floors calculated with true or fitted variance and length scale. The gray area indicates ±1 standard deviation around the estimated mean

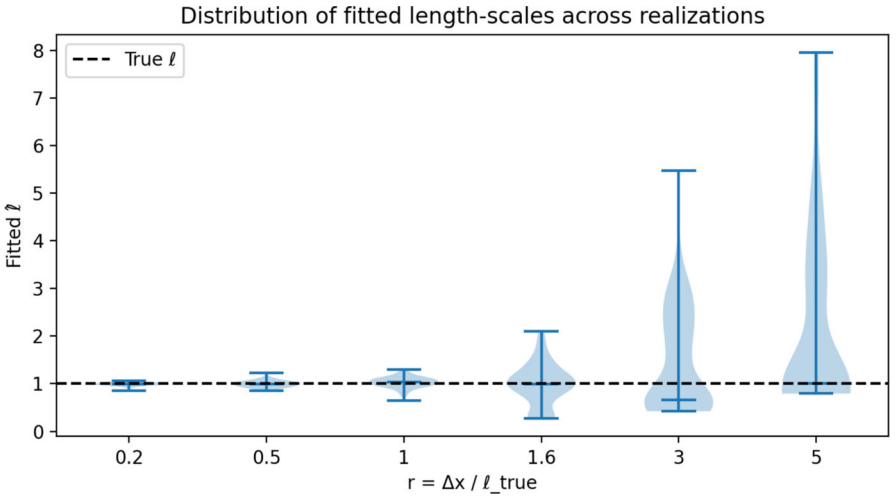


Fig. 4 Fitted length scale across 100 realizations generated in Test 1 at different sampling intervals

producing an error floor that underestimates the true reconstruction error. This explains why the curve falls below the benchmark floor at moderate to large spacing ratios.

The "practical floor" diverges further from the observed RMSE for sampling ratios $r > 1.6$ and starts to become smaller than the benchmark floor for $r > 2.5$. Interestingly, for $r \leq 2.5$, the opposing biases—overestimation of the length scale (Fig. 4) and underestimation of the variance—partially cancel, so the practical floor happens to lie closer to the observed RMSE than the benchmark floor. This apparent agreement at small r stems from bias cancellation rather than improvement in uncertainty estimation. When the sampling interval is larger than the distance over which any correlation exists, all information about the process between points is lost. In this limit, every error floor and the observed RMSE converge to the process standard deviation, and reconstruction is impossible beyond using the sample mean and variance.

The simulation confirmed that when the sampling spacing ratio met the Nyquist rate, $r \leq 0.5$, the RMSE was near zero, and nearly 100% of the variance was recovered. As spacing increased, the RMSE gradually increased, closely following the theoretical error floor until $r = 1$. Beyond this point, the error increased sharply. Beyond $r = 1.6$ ($\varepsilon = 0.05$), the discrepancy between the observed RMSE and the benchmark floor also increased more rapidly. By $r = 4$, only about 57% of the process variance could be recovered.

5.1.3 Test 2: Error Reduction with Sampling Rate

The heatmap (Fig. 5) shows the length scale absolute estimation errors as $E_l = \widehat{|l} - l_0|$, fitted by GP with the Gaussian RBF kernel. The spacing corresponding to the Nyquist rate at $\varepsilon = 0.01$ is $r_{nyq}(0.01) \approx 1.22$. This corresponds to around $s = \frac{17}{l_0}$ samples. From Fig. 3, we observe that when $r \leq 1$, the benchmark, practical, and marginal floors almost coincide, indicating very little estimation error. This requires around 20 samples. The red line indicates $r = 0.5$, that is, $l = \frac{40}{s}$ (Fig. 5); above the

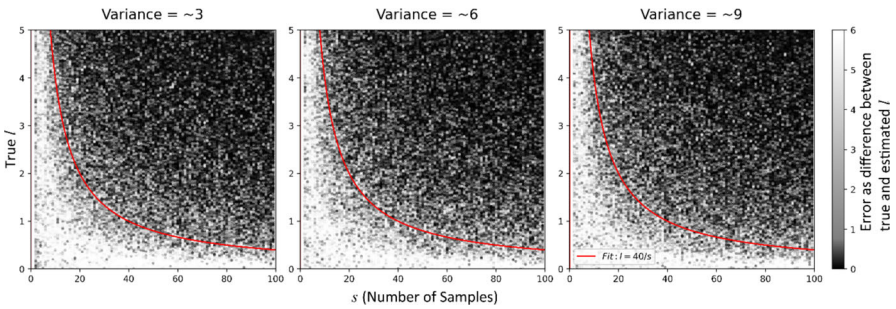


Fig. 5 Relationship between the GP kernel length scale l, sample size, and absolute recovery error in one-dimensional simulations. Shorter length scales (rougher signals) require denser sampling to avoid aliasing, consistent with the Nyquist condition. The fitted red line indicates a spacing ratio equal to 0.5, the most conservative line above which the errors are negligible. This applies to various variances (≈ 3, 6, and 9) in the signals

line, E_l becomes negligible, indicating full recovery of the signal. The variance does not affect these relationships.

5.2 One-Dimensional Demonstration of Test 2 with Real-World Data

For the one-dimensional real-world data from the LST transect in Fig. 2, we expect not only reduced error in estimating important kernel parameters with increasing number of samples same as the simulation but also increasingly reliable uncertainty quantification, as shown in Fig. 6. In the example below, it is a "lucky shot" that some of the relatively extreme low values are sampled even in the case of a very small sample size; hence, the model is able to encode this information about the large variance in the posterior variance, especially for locations distant from the samples, and the quantified uncertainty seems to reasonably encompass a large range that can

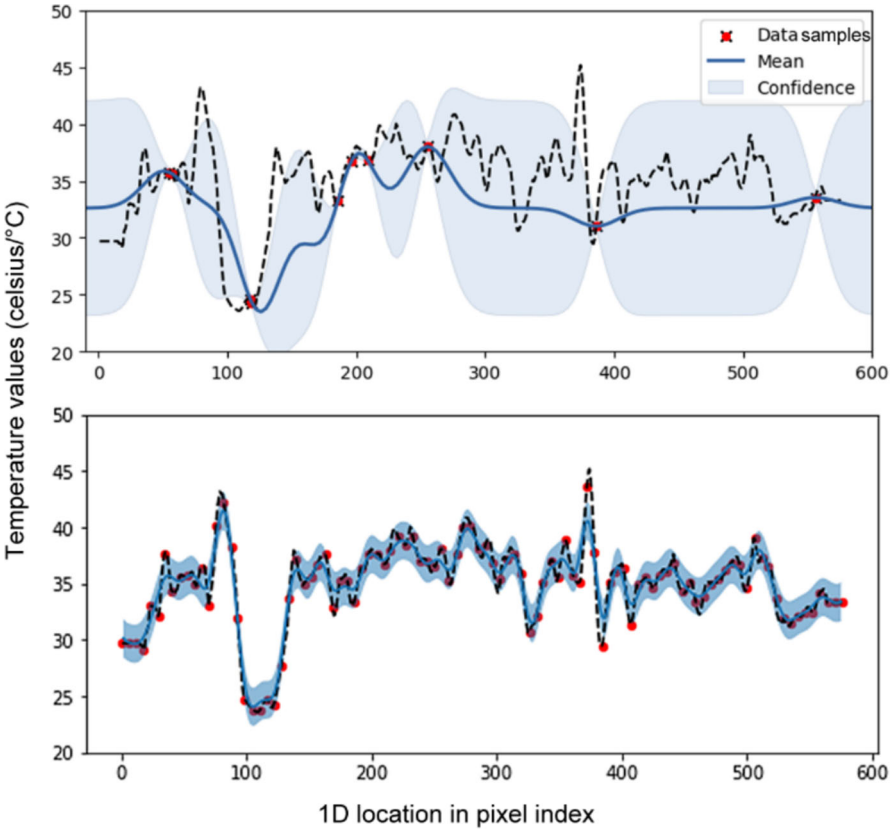


Fig. 6 The ground truth (dashed line), model prediction (dark blue line), and prediction interval (lighter blue band). For the one-dimensional real-world data of the LST transect, the modeled LST with a few samples is nothing close to the ground-truth pattern, while the uncertainty is large, especially at locations distant from the samples. The uncertainty could be mischaracterized if extreme values are not sampled (top). When the sample size is larger, the modeled pattern and uncertainty quantification both improve (bottom)

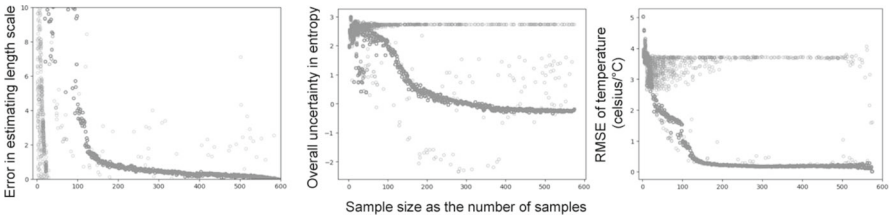


Fig. 7 From left to right: The absolute error of the modeled length scale, entropy of the latent process, and RMSE. We can observe that all three metrics decrease steeply with increasing sample size, with a tipping point at a sample size of around 150, where the modeled pattern starts to improve more slowly with further increases in sample size

include possible variations. As the number of samples increases with the decrease of distances among the samples, the model encodes confidence about nearby variance in the posterior variance. Consequently, the quantified uncertainty decreases everywhere. Figure 6 provides an intuitive visualization of how the modeled pattern and quantified uncertainty change with increasing sample size.

Further, for Test 2, the following metrics are calculated for different sample sizes and are visualized in Fig. 7 (from left to right).

1. Absolute error of the estimated length scale $E_l = |\widehat{l} - l_0|$
2. Entropy of the latent process (assumes a Gaussian variable) as an indicator of overall uncertainty, $H(X) = \frac{1}{2}\ln\left(2\pi e\sigma^2\right)$, where e is Euler's number, and σ^2 is the posterior variance of the latent process ($\sigma^2 = \text{var}(f_*|D)$, where f_* indicates the estimated latent process and D indicates the data).
3. RMSE: the RMSE between the observed value (LST) and the predicted values.

We observe that these quantities each approach zero as sample size increases (Fig. 7). Most interestingly, there is a clear tipping point at which all the quantities approach zero very quickly when the sample size reaches around 150 within the one-dimensional pattern under the given geospatial bounds. As the sample size increases beyond 150, the uncertainty and errors still decrease but at a flatter rate.

We apply FFT to examine the frequency distribution of the original pattern (Cooley and Tukey 1965). Given the fact that the original LST pattern consists of 576 pixels, the highest resolvable frequency is 0.5 cycles/sample, corresponding to 288 cycles across the 576 pixels. Therefore, it is sufficient to visualize frequencies up to this range. The FFT reveals that most of the spectral energy is concentrated around 70–80 cycles across the 576 pixels (Fig. 8). According to the Nyquist principle, at least twice this frequency is required for reliable reconstruction, implying a sampling rate of about 140–160 out of 576, or roughly 150 samples. This aligns with the tipping point observed in Fig. 7, where prediction errors and uncertainty drop markedly, and is consistent with the results shown in Fig. 3.

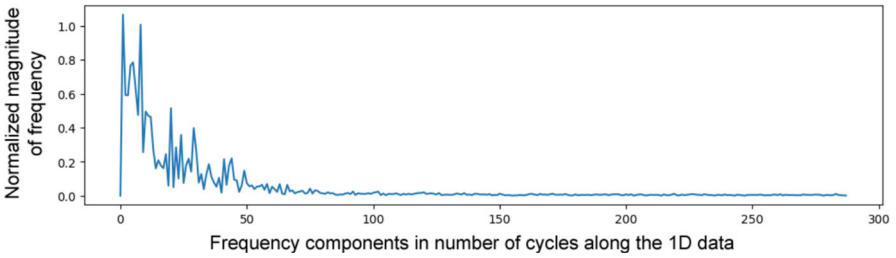


Fig. 8 The power spectrum obtained using the fast Fourier transform (FFT) shows that the main spectral energy in the original one-dimensional LST pattern extends up to around 70–80 cycles across the original 576 pixels, which requires 70–80 $\times$ 2 samples to reliably recover the original pattern. This explains the pattern in Fig. 7 as the smallest sample size to approach minimal prediction error

5.3 Two-Dimensional Demonstration of Test 2 with Real-World Data

Figure 9 shows the errors of the estimated length scales $E_l = \widehat{l} - l_0$ for the east–west and north–south directions, respectively, and the overall uncertainty indicated by the entropy of the latent process in the two-dimensional real-world case. We again observe the tipping point in the improvement of estimating the parameters of the ground-truth

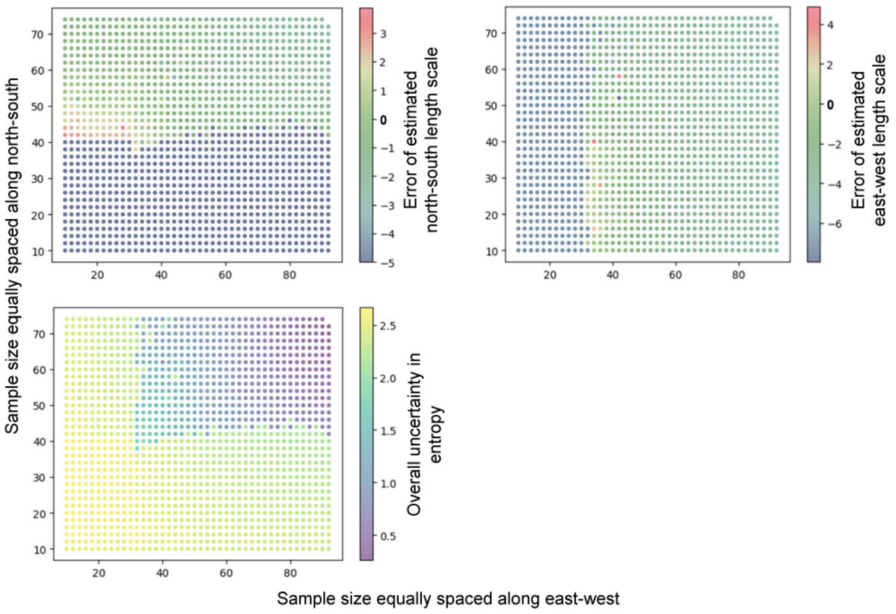


Fig. 9 The x-axes indicate the sample size in the east–west direction (the samples are equally spaced); the y-axes indicate the sample size in the north–south direction. The colors respectively indicate the error in the estimated length scale in the north–south direction (top left), the error in the estimated length scale in the east–west direction (top right), and the entropy of the estimated latent process (bottom left)

geographical pattern (Fig. 9). Figure 9 shows that the north–south length scale parameter estimation is only sensitive to the number of the samples along this direction, and at around 40–45 equally spaced samples, the error in the estimated length scale drops very close to zero. A similar situation can be observed for the east–west length scale estimation when around 30–40 samples are taken. The overall uncertainty quantified as entropy drops significantly when the number of samples taken along both directions reaches the threshold values corresponding to those for estimating the length scale parameters along both directions.

The geographical pattern in the frequency domain for the two-dimensional case further confirms the above findings as reasonable. This time, the frequency is normalized by the number of pixels along both directions, as only those frequencies below 0.5 times the total pixel sampling rate along both directions can be resolved according to the definition of the Nyquist frequency mentioned in Sect. 3. Hence, it is only necessary to visualize frequencies below the normalized value of 0.5 (Fig. 10). Along both directions, the highest frequencies seem to cease around 0.1. That is, the highest frequencies along the two directions correspond to $0.1 \times 230 = 23$ and $0.1 \times 180 = 18$ cycles within the given geospatial extent, respectively. The Nyquist sample size is therefore 46 and 36, aligning with the pattern in Fig. 9.

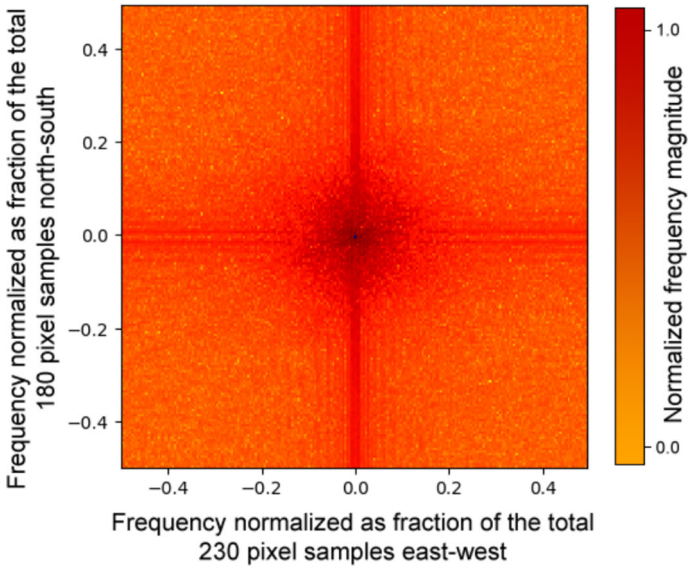


Fig. 10 Frequency analysis with a two-dimensional FFT applied to the two-dimensional subset showing that the highest frequency along both east–west and north–south directions is around 0.1 times the pixel sampling rates, that is, $0.1 \times 230 = 23$ and $0.1 \times 180 = 18$ cycles along the two directions within the given geospatial extent, respectively. Please note that the frequency is normalized by the number of pixels, or equivalently by the total pixel sampling rate, along each direction

5.4 Two-Dimensional Demonstration of Test 2 with Real-World Data: Involving Covariates

Covariates can improve modeling by providing statistically correlated information beyond the target variable itself. This can be thought of as sampling observations that contain indirect information—that is, not direct observations of the target variable, but measurements that are correlated with it to some degree. In this context, the sampling simulations described above can be extended to cases with covariates: geographically equally spaced samples are taken, but values are drawn either from the target variable or from a correlated covariate. To represent such a covariate in a controlled way, we simulate it by adding Gaussian noise to the target variable, thereby constructing an "ideal" covariate with imperfect correlation. In our experiment, the simulated covariate achieves an ordinary least squares (OLS) R-squared of 0.81 with the target two-dimensional LST subset (Fig. 11).

When applying the MTGP, a very similar pattern in the modeling outcome can be observed. As long as there is a good correlation between the target variable and the covariate, the error in estimating the ground-truth pattern—for instance, the RMSE—drops significantly when the number of samples approaches the Nyquist sampling rate in both the east–west and north–south directions, as shown in Fig. 12. This time, the RMSE does not drop very close to zero, which is reasonable, as the information in the simulated covariate is intentionally made not to correlate 100% with the original LST.

6 Discussion

We established a theoretical basis for sampling adequacy and the error floor (the irreducible prediction error due to unresolved high-frequency variance) by linking geospatial modeling with the Nyquist–Shannon sampling theorem. From the Gaussian RBF kernel spectrum, we derived a tolerance-indexed error floor and expressed it as

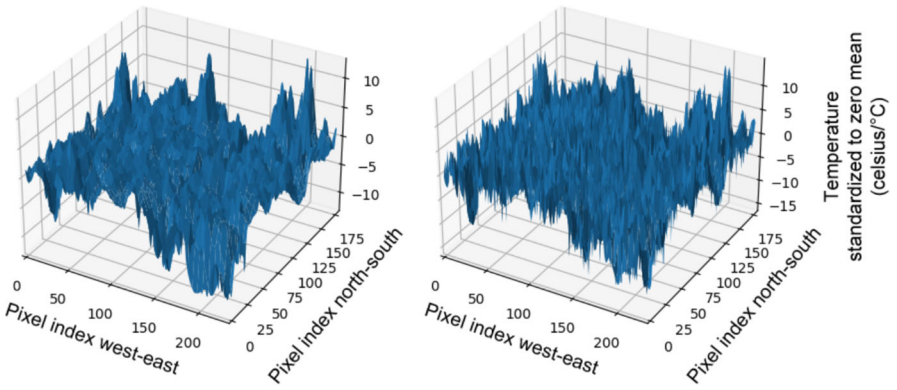


Fig. 11 The original two-dimensional subset ground-truth LST and the simulated covariate as a noisy version of the original LST

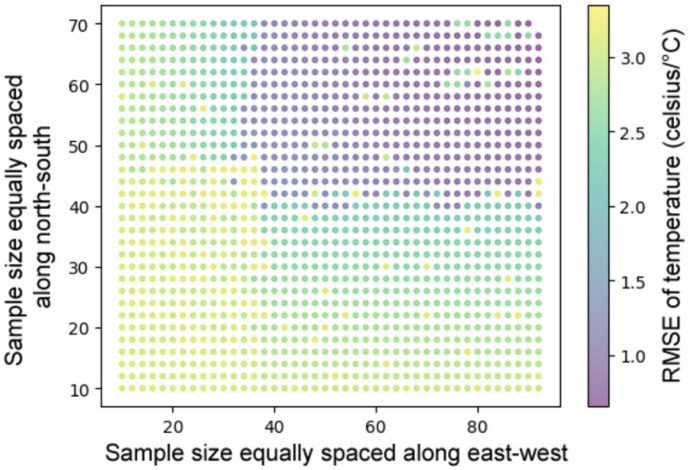


Fig. 12 The RMSE in estimating the ground-truth LST significantly decreases when the number of samples reaches around 40 along both the east–west and north–south directions, which conforms to the Nyquist sampling rate

a function of the sampling interval. Two tests with simulations and real-world data confirmed the theory, wherein (1) the observed errors conform to the error floor, and (2) the sampling interval required for estimating the length scale and predicting at a certain level of accuracy follows the Nyquist principle. This yields actionable rules linking sampling rate f_s, tolerated unresolved variance (ε), and the Gaussian RBF kernel length scale (l) that characterizes spatiotemporal autocorrelation. Instead of treating the effective bandwidth (f_{eff}) as a hard cutoff wavelength, which is not strictly valid for the RBF kernel whose spectrum is unbounded, we derive the exact relationship between the length scale and the sampling rate.

Our analysis has shown that establishing the link between l and f_{eff} can directly lead to I1 and I2. For I3, it is clear that the following additional diagnostics can be provided by combining the error floor with the RMSE:

If the l_0 and σ are known

RMSE $\approx \sigma\sqrt{\varepsilon(l_0, \Delta x)}$, the error or uncertainty is limited by sampling.

RMSE $> \sigma\sqrt{\varepsilon(l_0, \Delta x)}$, the model can be improved by, for example, using another kernel.

RMSE $< \sigma\sqrt{\varepsilon(l_0, \Delta x)}$, evaluation bias, that is, the training and test sets are dependent.

If we have estimated $\hat{l}$ and $\hat{\sigma}$

RMSE $< \hat{\sigma}\sqrt{\varepsilon\left(\hat{l}, \Delta x\right)}$, evaluation bias, or there is leakage or over-smoothing.

For irregular and high-dimensional sampling, we can replace Δx with an effective spacing Δ_{eff} (Qureshi 2012). This represents the distance between samples such that a regular grid with the Δ_{eff} would yield the same average information density as

the given (possibly irregular) sampling configuration. In practice, the Δ_{eff} can be approximated by, for example, the mean nearest-neighbor distance or other coverage metrics in feature space.

It has long been recognized that denser sampling decreases predictive uncertainty and that a random field with integrable spectral density has a spectral representation (Stein 1999). Classical works such as Cressie (2015) show that kriging variance decreases with denser samples and depends on the assumed covariance function. The benefit of analyzing spatial correlation in the frequency domain has been well recognized, and recent work has also used spectral representations in GP regression (e.g., Miao and Wang 2025). However, uncertainty is primarily expressed through kriging variance, posteriors, and effective range. Our contribution is to make the connection explicit between the parameters of the covariance function and the Nyquist principle by analyzing the process in the frequency domain.

Although the statistical model is continuous as in spline or GP formulations, the scale of meaningful prediction is bounded by the sampling resolution and the smoothness of the underlying process. We refer to this limit as the effective output resolution, which is illuminated by how we apply the Nyquist principle in this study. Accessing the effective resolution of the outcome map, that is, the spatial prediction scale that is reliably supported by the measurements and the adopted model, is an important question that has been neglected in most environmental mapping studies. Many mapping products are delivered at an arbitrarily chosen resolution, such as the resolution of the predictors. For example, using official ground station measurements together with covariates such as road density in different buffers and populations, Larkin et al. (2017) produced a global NO_2 map at 100 m resolution; the map is used as an input in various important geo-health studies. Subsequent work (e.g. Lu et al. 2020a) produced NO_2 maps at 25 m resolution. Given the nonlinear and geographically heterogeneous relationships between covariates and NO_2, and the sparsity of the ground stations, it is not obvious that fine-scale variability is recoverable at those resolutions. Current geospatial modeling or mapping activities rarely explicitly assess whether the data support the outcome resolution. We argue that output (map) resolution should align with recoverable variability. Our study takes an important step toward addressing this problem; a full treatment of how sampling rate (or effective spacing in high-dimensional feature space), kernel choice, and length scale estimation jointly determine effective output resolution is left for future work.

7 Conclusion

7.1 Gap we address

Uncertainty quantification does not provide an explicit link between sample adequacy and the expected minimum achievable predictive error. In sampling design, the relationship between sample size and predictive uncertainty is commonly established by minimizing variance. Our approach complements current variance-minimizing methods by providing the bound given the sampling density and the tolerance level, which is different from optimizing where to place the samples to approach that bound. When

combined with other uncertainty or error quantities, our bound enables further analysis of the sources of error—whether the error is due to parameter estimation, inadequate sampling, or cross-validation design.

7.2 Intellectual contribution

Conceptually, we link geostatistics (the length scale of a covariance function) to signal processing (the Nyquist rate and frequency) and obtain a closed-form mapping from $f_\varepsilon(l) = \mathrm{c}(\varepsilon)/l$ and the unresolved variance bound $\varepsilon(l, \Delta x)$. It also allows us to disentangle different sources of error, which complements existing quantities such as cross-validation error and predictive variance. By making explicit how error tolerance, sampling interval, and process length scale interact, we offer a practical guide for designing sampling campaigns and evaluating model reliability.

7.3 How to use it

Several practical implications emerge. First, covariates should be carefully selected to support interpretation and uncertainty quantification. Informative covariates can explain part of the signal, which potentially reduces the effective bandwidth of the residual process to be modeled by the GP. This may reduce the sample size needed to reach a given level of predictive performance. Future studies should further investigate how covariates affect adequate sample size. Second, experimental mapping in small areas can be an effective strategy when sampling resources (e.g., sensors, personnel) are limited. Detailed campaigns, such as high-resolution air pollution monitoring (Apte et al. 2017), provide knowledge about spatial variability that can then inform broader sampling configurations and model design.

7.4 Limitations and future directions

As a baseline study, this work is restricted to noise-free situations and regular sampling, but the derivations and test results invite broader extensions. In future work, we will extend the analysis to explicitly examine the role of the signal-to-noise ratio and to include irregular and high-dimensional sampling schemes, where effective spacing must be defined differently. In addition, future studies should systematically link existing sampling theorems to geospatial modeling. Several relevant extensions of the Nyquist–Shannon sampling theorem exist, such as Landau's generalization, which addresses recovery at a given average sampling rate (Landau 1967; Marvasti 2001). Developments in compressed sensing demonstrate that signals can be recovered with fewer samples than the Nyquist rate if they are sparse in the frequency domain (Candes and Wakin 2008; Candès et al. 2006; Donoho 2006).

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was used to conduct this research or prepare this manuscript.

Data and Codes Availability All code and data used in this study are available below to fully reproduce the results presented in this paper: https://figshare.com/articles/software/codes_for_results_of_the_Nyquist_

Shannon_Guidance_for_Linking_Sampling_Adequacy_and_Predictive_Uncertainty_in_Geospatial_Mapping/30405754.

Declarations

Conflict of interest The authors declare that there are no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Declaration of generative AI in the writing process The conceptualizations, theoretical derivations, analyses, and interpretations presented in this paper were developed entirely by the authors. During manuscript preparation, the authors used *OpenAI's ChatGPT (GPT-5, 2025)* to proofread text and checked mathematical consistency. All content and conclusions remain the sole intellectual responsibility of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

An Y, Yang L, Zhu A-X, Qin C-Z, Shi J (2018) Identification of representative samples from existing samples for digital soil mapping. Geoderma 311:109–119. https://doi.org/10.1016/j.geoderma.2017.03.014

Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, Marshall JD, Portier CJ, Vermeulen RCH, Hamburg SP (2017) High-resolution air pollution mapping with Google Street View Cars: exploiting big data. Environ Sci Technol. https://doi.org/10.1021/acs.est.7b00891

Biggio L, Wieland A, Chao MA, Kastanis I, Fink O (2021) Uncertainty-aware prognosis via deep Gaussian process. IEEE Access 9:123517–123527. https://doi.org/10.1109/ACCESS.2021.3110049

Bonilla EV, Chai K, Williams C (2008) Multi-task Gaussian process prediction. In: Advances in neural information processing systems 20, pp 153–160. NIPS Foundation

Bretthorst GL (2001) Nonuniform sampling: bandwidth and aliasing. AIP Conf Proc 567:1–28. https://doi.org/10.1063/1.1381847

Candès EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math. https://doi.org/10.1002/cpa.20124

Candes EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag. https://doi.org/10.1109/MSP.2007.914731

Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comput. https://doi.org/10.1090/S0025-5718-1965-0178586-1

Cressie N (1989) Geostatistics. Am Stat. https://doi.org/10.1080/00031305.1989.10475658

Cressie N (2015) Statistics for spatial data. Wiley, New York

Cressie N, Wikle CK (2011) Statistics for spatio-temporal data. Wiley, New York

Dong Z, Jia S, Shao C (2024) Filtered kriging for improved interpolation of periodic manufacturing surfaces. J Manuf Process 131:1–12. https://doi.org/10.1016/j.jmapro.2024.09.002

Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory. https://doi.org/10.1109/TIT.2006.871582

Garma RJD, Schott JR, Fiete RD, Qiao J, McKeown D (2017) Image quality modeling and characterization of Nyquist-sampled framing sensors with operational considerations for remote sensing. Opt Eng. https://doi.org/10.1117/1.OE.56.1.013102

Gomes Gonçalves Í, Wellmann F (2025) Uncertainty propagation in deep Gaussian process networks. Math Geosci 57(6):1115–1133. https://doi.org/10.1007/s11004-025-10187-4

Grafström A, Schelin L (2014) How to select representative samples. Scand J Stat. https://doi.org/10.1111/sjos.12016
Isaaks EH, Srivastava RM (1990) An introduction to applied geostatistics. Oxford University Press, Oxford
Iwata T, Ghahramani Z (2017) Improving output uncertainty estimation and generalization in deep learning via neural network Gaussian processes. arXiv preprint arXiv:1707.05922
Jerri AJ (1977) The Shannon Sampling Theorem—its various extensions and applications: a tutorial review. Proc IEEE. https://doi.org/10.1109/PROC.1977.10771
Kwon Y, Won JH, Kim BJ, Paik MC (2020) Uncertainty quantification using Bayesian neural networks in classification: application to biomedical image segmentation. Comput Stat Data Anal 142:106816. https://doi.org/10.1016/j.csda.2019.106816
Landau HJ (1967) Necessary density conditions for sampling and interpolation of certain entire functions. Acta Math. https://doi.org/10.1007/BF02395039
Larkin A, Geddes JA, Martin RV, Xiao Q, Liu Y, Marshall JD, Brauer M, Hystad P (2017) Global land use regression model for nitrogen dioxide air pollution. Environ Sci Technol 51(12):6957–6964. https://doi.org/10.1021/acs.est.7b01148
Li X, Zhou Y, Hejazi M, Wise M, Vernon C, Iyer G, Chen W (2021) Global urban growth between 1870 and 2100 from integrated high resolution mapped data and urban dynamic modeling. Commun Earth Environ. https://doi.org/10.1038/s43247-021-00273-w
Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS ONE. https://doi.org/10.1371/journal.pone.0031743
Lino Pereira J, Varouchakis EA, Karatzas GP (2024) Uncertainty quantification in geostatistical modelling of saltwater intrusion at a coastal aquifer system. Math Geosci 56:867–885. https://doi.org/10.1007/s11004-023-10120-7
Lu M, Schmitz O, de Hoogh K, Kai Q, Karssenberg D (2020a) Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale. Environ Int 142:105856. https://doi.org/10.1016/j.envint.2020.105856
Lu M, Soenario I, Helbich M, Schmitz O, Hoek G, van de Molen M, Karssenberg D (2020b) Land use regression models revealing spatiotemporal co-variation in NO2, NO, and O3 in the Netherlands. Atmos Environ 223:117238. https://doi.org/10.1016/j.atmosenv.2019.117238
Lu M, Cavieres J, Moraga P (2023) A comparison of spatial and nonspatial methods in statistical modeling of NO2: prediction accuracy, uncertainty quantification, and model interpretation. Geogr Anal 55:703–727. https://doi.org/10.1111/gean.12356
MacKay DJC (1998) Introduction to Gaussian processes. In: Bishop CM (ed) Neural networks and machine learning. Springer, Berlin
Marvasti F (ed) (2001) Nonuniform sampling: theory and practice. Kluwer Academic, Amsterdam
Meyer H, Katurji M, Appelhans T, Müller MU, Nauss T, Roudier P, Zawar-Reza P (2016) Mapping daily air temperature for Antarctica based on MODIS LST. Remote Sens. https://doi.org/10.3390/rs8090732
Miao C, Wang Y (2025) Sparse spectrum representation of Gaussian process regression for modeling nonstationary geo-data from limited measurements. Math Geosci. https://doi.org/10.1007/s11004-025-10243-z
Moraga P (2023) Spatial statistics for data science: theory and practice with R. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781032641522
Morlighem C, Chaiban C, Georganos S, Brousse O, van Lipzig NPM, Wolff E, Dujardin S, Linard C (2023) Spatial optimization methods for malaria risk mapping in Sub-Saharan African cities using Demographic and Health Surveys. Geohealth. https://doi.org/10.1029/2023GH000787
Nyquist H (1928) Certain topics in telegraph transmission theory. Trans Am Inst Electr Eng 47(2):617–644. https://doi.org/10.1109/T-AIEE.1928.5055024
Oppenheim AV, Schafer RW (1998) Discrete-time signal processing, 2nd edn. Prentice Hall, Hoboken
Qin CZ, Bao LL, Zhu AX, Wang RX, Hu XM (2013) Uncertainty due to DEM error in landslide susceptibility mapping. Int J Geogr Inf Sci 27(7):1364–1380. https://doi.org/10.1080/13658816.2013.770515
Qin Z, Gao Y, Parini CG (2016) Data-assisted low complexity compressive spectrum sensing on real-time signals under Sub-Nyquist rate. IEEE Trans Wirel Commun 15(2):1174–1185. https://doi.org/10.1109/TWC.2015.2485992
Qureshi MA, Schmidt CH, Eibert TF (2012) An effective non-redundant sampling representation for planar near-field antenna measurements. In: 34th ESA antenna workshop, Noordwijk, the Netherlands
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge

Rogerson PA, Delmelle E, Batta R, Akella M, Blatt A, Wilson G (2004) Optimal sampling design for variables with varying spatial importance. Geogr Anal 36(2):177–194. https://doi.org/10.1111/j.1538-4632.2004.tb01131.x

Sayler K (2023) Landsat 8–9 collection 2 (C2) level 2 science product (L2SP) guide. U.S. Geological Survey, New York

Scheider S, Ostermann FO, Adams B (2017) Why good data analysts need to be critical synthesists: determining the role of semantics in data analysis. Future Gener Comput Syst 72:11–22. https://doi.org/10.1016/j.future.2017.02.046

Senarathne SGJ, Müller WG, McGree JM (2023) Bayesian design for minimizing prediction uncertainty in bivariate spatial responses with applications to air quality monitoring. Biom J 65(4):2100386. https://doi.org/10.1002/bimj.202100386

Shaddick G, Thomas ML, Green A, Brauer M, Donkelaar A, Burnett R, Chang HH, Cohen A, Dingenen RV, Dora C, Gumy S (2018) Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. J R Stat Soc C Appl Stat 67(1):231–253. https://doi.org/10.1111/rssc.12227

Shannon CE (1949) Communication in the presence of noise. Proc IRE 37(1):10–21. https://doi.org/10.1109/JRPROC.1949.232969

Shen Y, de Hoogh K, Schmitz O, Clinton N, Tuxen-Bettman K, Brandt J, Christensen JH, Frohn LM, Geels C, Karssenberg D, Vermeulen R, Hoek G (2022) Europe-wide air pollution modeling from 2000 to 2019 using geographically weighted regression. Environ Int 168:107485. https://doi.org/10.1016/j.envint.2022.107485

Srinivasan R, Tucker DM, Murias M (1998) Estimating the spatial Nyquist of the human EEG. Behav Res Methods Instrum Comput 30(1):8–19. https://doi.org/10.3758/BF03209412

Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer, New York. https://doi.org/10.1007/978-1-4612-1494-6

Turley MC, Ford ED (2009) Definition and calculation of uncertainty in ecological process models. Ecol Model 220(17):1968–1983. https://doi.org/10.1016/j.ecolmodel.2009.04.046

Tyralis H, Papacharalampous GA (2024) A review of predictive uncertainty estimation with machine learning. Artif Intell Rev 57:94. https://doi.org/10.1007/s10462-023-10698-8

Unser M (2000) Sampling—50 years after Shannon. Proc IEEE 88(4):569–587. https://doi.org/10.1109/5.843002

Voogt JA, Oke TR (2003) Thermal remote sensing of urban climates. Remote Sens Environ 86(3):370–384. https://doi.org/10.1016/S0034-4257(03)00079-8

Wallentin G, Car A (2013) A framework for uncertainty assessment in simulation models. Int J Geogr Inf Sci 27(2):408–422. https://doi.org/10.1080/13658816.2012.715163

Wang J, Zhan Q, Guo H (2016) The morphology, dynamics and potential hotspots of land surface temperature at a local scale in urban areas. Remote Sens 8(1):18. https://doi.org/10.3390/rs8010018

Yang D, Zhang S, Niu T, Wang Y, Xu H, Zhang KM, Wu Y (2019) High-resolution mapping of vehicle emissions of atmospheric pollutants based on large-scale, real-world traffic datasets. Atmos Chem Phys 19(13):8831–8843. https://doi.org/10.5194/acp-19-8831-2019

Zhang H (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J Am Stat Assoc 99(465):250–261. https://doi.org/10.1198/016214504000000241

Zhang G, Zhu A-X (2019) A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping. Geoderma 351:130–143. https://doi.org/10.1016/j.geoderma.2019.05.024

Zhang T, Zhou Y, Zhu Z, Li X, Asrar GR (2021) A global seamless 1 km resolution daily land surface temperature dataset (2003–2020). Iowa State University, Ames. https://doi.org/10.25380/iastate.c.5078492

Zhu A-X, Liu J, Du F, Zhang SJ, Qin C-Z, Burt J, Behrens T, Scholten T (2015) Predictive soil mapping with limited sample data. Eur J Soil Sci 66(3):535–547. https://doi.org/10.1111/ejss.12244

Zhu A-X, Lu G, Liu J, Qin C-Z, Zhou C (2018) Spatial prediction based on Third Law of Geography. Ann GIS 24(4):225–240. https://doi.org/10.1080/19475683.2018.1534890

Zhu A-X, Ma T, Zhao F, Yang X, Xia Y (2025) Uncertainty quantification for digital soil mapping: an overview. Pedosphere. https://doi.org/10.1016/j.pedsph.2025.12.011

Zhu A-X, Liu J, Yang X, Brungard C (2026) Quantification of prediction uncertainty due to sample representation for digital soil mapping based on Third Law of Geography. Pedosphere. https://doi.org/10.1016/j.pedsph.2026.03.003