Applied Physics B
Lasers and Optics

Process of measurement error treatment using model selection and local intensive smoothing and application to refractive index estimation of water

Kyungeun Lee[1] · Joong Bae Kim[1] · Jeong Woo Park[1] · Bong Jae Lee[1] · Ikjin Lee[1]

Received: 22 May 2019 / Accepted: 12 February 2020 / Published online: 6 March 2020

Abstract

In this paper, a new process to deal with measurement error is proposed using smoothing, regression, and model selection. The main objectives of this research are to construct a theoretically reliable process to deal with the measurement errors and to validate the process with a case study for refractive index estimation of water. The proposed process for measurement error treatment consists of (1) smoothing of spiky fluctuation, (2) integration of multiple measurements into a single prediction model with statistical regression, and (3) physics-based model selection. The first and third processes enhance the nominal accuracy, and the second process improves the precision of the estimation. In particular, a methodology for intensive local smoothing with a new criterion is proposed for physically unreasonable spikes that cannot be smoothed enough with existing criteria. Before applying all the proposed processes to refractive index estimation of water, three candidate models of the refractive index are generated according to their physics-based possibility. The generated models are tested through the proposed processes, and a final model is selected according to the principle of Occam's razor. The proposed process results in much improved estimation of the refractive index of water by reducing the estimation error from 3.90% to 1.95% of absolute error. Through this study, a useful methodology to deal with measurement errors is successfully established and it can be also applied to problems with similar type of measurement data.

1 Introduction

There are several error sources in measurements due to inherent uncertainties such as distribution of physical properties, setting error, equipment resolution, and environmental effect. Certain errors affect the nominal accuracy that causes bias, and some errors disperse repeated measurements that cause deviation. Throughout the literature survey, there have been methodologies commonly used in measurement error treatment. As one of the representative methods, smoothing has been widely used and studied to squash out invalid fluctuations of data. Smoothing has been implemented by kernel smoothing [1], process convolution [2, 3], Gaussian process regression (GPR) with multi-kernel [4], GPR with random inputs [5, 6], spline smoothing [7, 8] and various filters [9]. A problem of almost all smoothing methods is to determine the level of smoothness controlled by their model parameters. For example, there have been researches to determine the frame length of the Savitzky–Golay (SG) filter [10–15] and kernel bandwidth of kernel soothing [16–19]. There also exists a criterion to determine smoothness for general purpose regardless of smoothing method [8]. In addition, studies on boundary problems [19–22] have been proposed to solve boundary errors occurring in smoothing.

Next, regression is often utilized for deviation treatment to integrate multiple measurements into a single prediction model with statistical synthesis. GPR, one of the most well-known statistical regression methods [23–28], constructs a regression model by hyperparameter optimization through likelihood maximization. Since GPR often causes overfitting by tracking all of meaningless spiky oscillations of data instead of smoothing them, smoothing must be preceded before the regression.

In spite of the developed methods, measurement error treatment is still an unresolved area. The first reason is that the aforementioned error treatment methods cannot assure whether the error is removed and the actual signal is

✉ Ikjin Lee
ikjin.lee@kaist.ac.kr

[1] Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

preserved. Second, most of the smoothing methods cannot remove noise clearly, which means there could be a case of overfitting or under-smoothness. This is because the smoothness selection criteria are based on the model accuracy such as cross-validation error or mean squared error. The final difficulty is that error treatment method depends on numerical or physical properties of data, which makes it difficult to select an appropriate method for given data.

To resolve the last problem, a case that commonly occurs but hard to handle is utilized in this study. The selected case is to estimate unmeasurable physical property from other measurable properties using an analytic function. This is hard to treat since the error of the measured properties may cause unpredictable and complex ramifications on the accuracy of the resultant estimated property. For the second problem, this study suggests a new criterion for local intensive smoothing while maintaining a specified level of accuracy. Finally, for the first problem, model selection with prior knowledge is applied. Under high-uncertainty situation, several hypothetical models can be established and the most valid model among the candidates must be selected. In this case, we utilize physical knowledge and Occam's razor-based model selection [29–31]. The whole process is illustrated with a problem of refractive index estimation of water.

The remainder of this article is organized as follows: introduction on the optical property, characteristic of measured raw data, and problems of the existing method are explained in Sect. 2. The proposed process to remove measurement noise including smoothing, GPR, and model selection is described in Sect. 3. Application of the proposed process to refractive index estimation of water as a case study is illustrated in Sect. 4. Finally, Sect. 5 summarizes the application results.

2 Optical property estimation and measured data characteristics

This section presents the fundamental physics of optical property estimation, the investigation of the measured data, and problems of the existing method in the optical property estimation without preprocess [32]. Especially, refractive index estimation of water and its experimental conditions that will be used as a case study in Sect. 4 are described.

2.1 Introduction of optical property estimation

Optical properties of materials have practical importance in many engineering fields such as solar thermal collector or glass fibers. Recently, various studies to tune the optical properties of materials with nanoparticles have been intensively reported [33, 34]. Plasmonic nanofluid is a suspension employing the plasmonic nanoparticles whose electron can couple with the light. It is possible to improve the absorption efficiency of a solar thermal collector and to minimize pumping loss simultaneously using an extremely small amount of metal nanoparticles.

In applications of the plasmonic nanofluids, it is important to accurately predict absorption and scattering phenomena of the nanofluids that depend on material, shape, and size of the nanoparticles and properties of the base fluid [35]. For the prediction, a refractive index of the base fluid is required, but there has been little information of optical constants about the base fluid for the direct solar thermal collector. Therefore, the measurement of the refractive index of the unknown fluid is essential for the prediction of absorption and scattering in nanofluids.

There have been two methods to obtain the refractive index: (1) to measure refraction angle or beam displacement [36–38] and (2) to solve an inverse problem of Airy's formulae using measured transmission (T) and reflection (R) spectra [39, 40]. Using the second method, the refractive index (n) is directly determined through a simple calculation from ultraviolet–visible to the infrared range as

$$n(\lambda) = \text{function}(T(\lambda), R(\lambda)) \tag{1}$$

at a certain wavelength λ. Detailed explanation on Eq. (1) is found in Kim et al. [32].

If T and R are accurately measured, the refractive index of the base fluid can be precisely estimated. However, estimation of the refractive index using Eq. (1) is quite sensitive to measurement errors as can be seen in the subsequent section. Consequently, it is essential to develop an uncertainty treatment method to estimate an accurate refractive index with this method when there exist measurement errors in T and R.

2.2 Experimental conditions and problems of the existing method

This section presents measurement results of T and R of water, base fluid for the nanofluid, and the refractive index estimated using Eq. (1) with the mean of the measured T and R without any preprocess according to the existing method [32]. For the refractive index estimation of water, T and R are measured 30 times along the wavelength, as shown in Fig. 1a and b. As can be seen from Fig. 1b, the fluctuation is severe in short wavelength, and the deviation between measurements is very large in long wavelength in R. As shown in Fig. 1a, T has a large interval between the maximum and minimum values compared with R, and T shows relatively small fluctuation along the wavelength and small deviation between measurements.

In the previous study [32], the mean of repetitive measurements with no preprocess on T and R is used for the

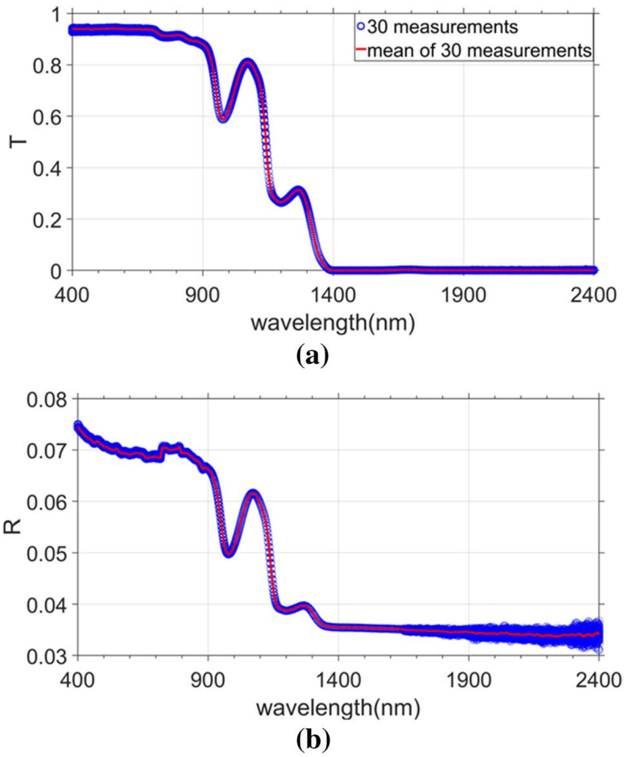

Fig. 1 Measured data of **a** *T* and **b** *R* from 30 experiments

refractive index estimation that results in large errors, as shown in Fig. 2. The estimation accuracy of the refractive index can be quantified using Palik's data that is known as the exact refractive index of water. The estimation error between the exact refractive index and the mean of 30 measurements is 3.95% at the wavelength of 575 nm that is a very low level of accuracy even the multiple measurements cancel out most deviation errors. Hence, additional treatment is needed to reduce the estimation error under 2%.

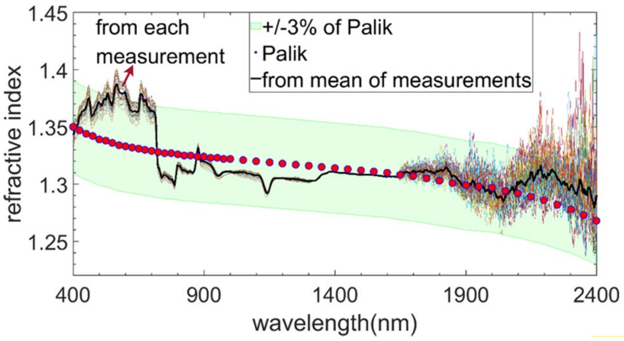

Fig. 2 Estimated refractive index of water using the existing method [32]

2.3 Measured data characteristics

The reasons for the large estimation error come from (1) measurement bias and (2) noisy fluctuation of measured data. To examine the causes of errors, first, the region where $T + R > 1$ as shown in Fig. 3 must have a measurement bias that needs correction. According to the nature of physics, $T + R$ must be equal to or less than one since $T + R +$ absorptance must be equal to one and the absorptance is always non-negative.

Second, as shown in Fig. 1b, spiky fluctuations along the wavelength are observed in *R*. Estimated refractive index is sensitive to the measurement error of *R* since the scale of *R* is small compared with *T* and the effect of the same amount of value change of *R* becomes larger than that of *T*.

However, the measurement error cannot be reduced by careful manipulation since it is originated from nature and experimental equipment. Therefore, a process to resolve the problems above is proposed by introducing appropriate error treatment methods in Sect. 3.

3 Methods and process for data treatment

Proper data treatment to reduce measurement errors in *T* and *R* enhances the accuracy of refractive index estimation. For the purpose, a data treatment process is proposed that consists of smoothing of fluctuation, statistical regression of multiple measurements, and physics-based model selection explained in Sects. 3.2, 3.3, and 3.4, respectively. Before performing the process, data segmentation for local intensive smoothing according to data characteristics needs to be carried out first that will be explained in Sect. 3.1.

3.1 Data segmentation for local intensive smoothing

Data segmentation is essential to decide regions to perform local intensive smoothing. The third derivative of data is

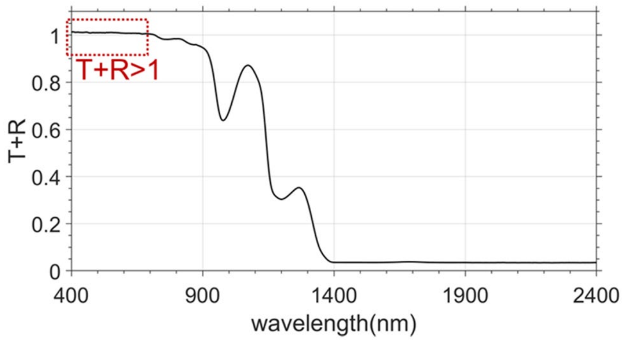

Fig. 3 $T + R$ obtained using the mean of measured *T* and *R*

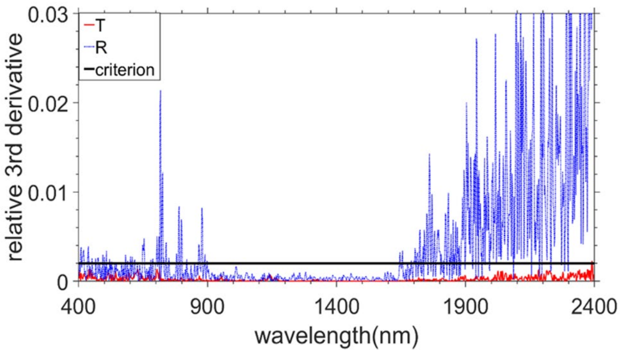

Fig. 4 Relative third derivatives of T and R

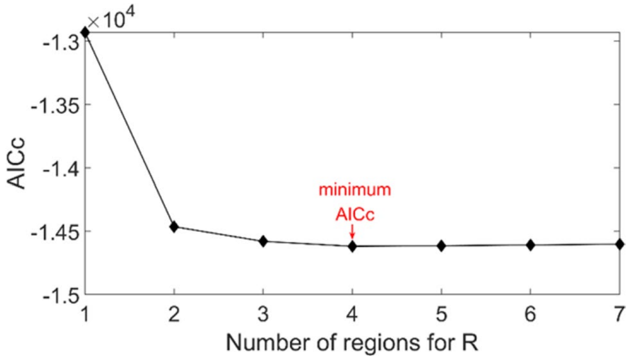

Fig. 5 AICc results according to the number of regions of R

used as a fluctuation level index in this research. The relative third derivatives of T and R are shown in Fig. 4, where the relative third derivative of $\mathbf{Y}$ is calculated as $\left|\frac{\text{3rd derivative of }\mathbf{Y}}{\max(\mathbf{Y})-\min(\mathbf{Y})}\right|$. As shown in Fig. 4, T shows small fluctuation in the whole wavelength, whereas R shows very high fluctuation in short- and long-wavelength regions compared with the criterion of 0.002. Therefore, region segmentation must be preceded to properly pick out regions to be smoothed intensively in R.

To establish the data segmentation standard, Gaussian mixture [41, 42] is utilized in this research. Gaussian mixture, a kind of data clustering methods according to data similarity and dissimilarity, is applied to region segmentation according to the likelihood of the third derivative level of R with a specified number of regions. Using the Gaussian mixture, split positions with the highest loglikelihood (ln$L$) are selected for the best positions whose number is determined by AICc (Akaike information criterion corrected) [43, 44]. ln L and AICc are calculated as

$$\ln L = \ln p(\mathbf{Y}|t_1, t_2, \ldots, t_{k-1}) = \ln\left(\sum_{j=1}^{k} N(\mathbf{Y}_j|\mu_j, \sigma_j)\right) \quad (2)$$

and

$$AICc = -2\ln L + 2k + \frac{2k(k+1)}{n-k-1}, \quad (3)$$

where $\mathbf{Y}$ means data to be segmented, k is the number of segment groups, t_{k-1} is the $(k-1)$th split position, $\mathbf{Y}_j$ is the jth segment of $\mathbf{Y}$, μ_j and σ_j are parameters of normal distribution fit for $\mathbf{data}_j$, and n is the number of elements in vector $\mathbf{Y}$. In the region segmentation process, only information about the split positions ($t_1 \sim t_{k-1}$) is used in the next procedure. AICc results obtained using the relative third derivative of measured R are shown in Fig. 5 which shows that the best number of regions is four due to the minimum AICc value. Detailed results from the region segmentation are shown in Sect. 4 using a case study of refractive index estimation of water.

3.2 Strategies for local intensive smoothing

As aforementioned, local intensive smoothing is required for locally and highly fluctuating regions as observed in R. The SG filter, which is one of the existing smoothing methods and has been extensively studied including the bandwidth selection [16–19], is introduced in Sect. 3.2.1. However, since the SG filter has limitations in local intensive smoothing, a new smoothing method is proposed in Sect. 3.2.2 to apply exclusively to high-fluctuation regions.

3.2.1 Theoretical background of SG filter

An SG filter, which is a least square fitting method inside of moving frame with a predetermined frame length and polynomial order, is selected to be the basic smoothing method in this research. It follows the form of a weighted sum of nearby neighbor's data within the frame with the frame length of len. With the determined len that must be an odd number and raw data before filtering $\mathbf{y} = [y_1, y_2, y_3, \ldots]$, filtered data $\mathbf{Ynew}$ at the jth point is calculated with a coefficient matrix $\mathbf{C}$ as

$$\mathbf{Ynew}_j = (\mathbf{C} \otimes \mathbf{y})_j = \sum_{i=\frac{-\text{len}+1}{2}}^{\frac{\text{len}-1}{2}} \mathbf{C}_i y_{j+i}. \quad (4)$$

For example, if len = 5, then $i = -2, -1, \ldots, 2$ where the minus i implies that the point y_{j+i} is located in the left of y_j. The coefficient matrix $\mathbf{C}$ is obtained using the least square fit of $y_{j-2}, y_{j-1}, \ldots, y_{j+2}$ with a low order of polynomials. For a case of equally spaced data, when $\mathbf{z} = [-2, -1, 0, 1, 2]^T$, the mth order polynomial approximation of the data within the frame can be expressed as

$$\mathbf{Ynew} = a_0 + a_1\mathbf{z} + a_2\mathbf{z}^2 + \cdots + a_m\mathbf{z}^m. \quad (5)$$

$\mathbf{a} = [a_0, a_1, \ldots, a_m]^T$ in Eq. (5) is obtained by solving the normal equation of

$$\mathbf{y} = a_0 + a_1\mathbf{z} + a_2\mathbf{z}^2 + \cdots + a_m\mathbf{z}^m \tag{6}$$

whose solution is given by

$$\mathbf{a} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \tag{7}$$

with

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}^0_{(1)} & \mathbf{z}^1_{(1)} & \cdots & \mathbf{z}^m_{(1)} \\ \mathbf{z}^0_{(2)} & \mathbf{z}^1_{(2)} & \cdots & \mathbf{z}^m_{(2)} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{z}^0_{(\mathrm{len})} & \mathbf{z}^1_{(\mathrm{len})} & \cdots & \mathbf{z}^m_{(\mathrm{len})} \end{bmatrix}.$$

Therefore, the convolution coefficient **C** is expressed as

$$\mathbf{C} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}} \tag{8}$$

As shown in Fig. 6, as the frame length becomes larger, smoothing effect becomes higher, which means that the decision of the frame length directly affects noise removal.

3.2.2 Proposed intensive smoothing method

The SG filter is appropriate due to its smoothness control ability by frame length and not distorting the front and rear parts of the data. The important problem is how to determine the frame length to control the smoothing level. To resolve the problem, the existing spline smoothing method [7, 8] utilizes the regularization equation [43, 44] given by

$$g = \left(\sum \left(\mathbf{Y}_i - f(\mathbf{x}_i)\right)^2 + \alpha \int \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2}\mathrm{d}\mathbf{x}\right), \tag{9}$$

where the first term is conformity between true observation and prediction f and the second term is a roughness penalty. This regularization has been widely used to impose smoothness for other types of smoothing models such as bandwidth selection for kernel smoothing. In Eq. (9), α is determined by minimizing the cross-validation error [8]. Since the smoothing parameter of the SG filter is defined by len, Eq. (9) can be rewritten in terms of two parameters, len and α, as

$$g(\mathrm{len}, \alpha) = \left(\sum \left(\mathbf{Y}_i - f(\mathbf{x}_i;\mathrm{len}, \alpha)\right)^2 + \alpha \int \frac{\partial^2 f(\mathbf{x};\mathrm{len}, \alpha)}{\partial \mathbf{x}^2}\mathrm{d}\mathbf{x}\right) \tag{10}$$

which needs to be minimized to obtain the optimal len with a specified α. α and len that minimize the cross-validation error (cve) given by

$$\mathrm{cve}(\mathrm{len}) = \sum_i^N \left(\mathbf{y} - f_{\sim i}(\mathbf{x};\mathrm{len}, \alpha_*)\right)^2 \tag{11}$$

with specified $\alpha = \alpha_*$ is selected. As can be seen in Fig. 7, the smoothing effect becomes stronger as α increases. Especially, when there is no roughness penalty in Eq. (10), the smoothing result almost follows the measurement data.

The overall smoothing process is shown in Fig. 8. As shown in Fig. 8, the process includes two optimization loops: the inner loop for $\mathrm{len}_{\mathrm{opt}}$ and the outer loop for α_{opt}. That is, each α has its own $\mathrm{len}_{\mathrm{opt}}$, and α with the minimum cve is selected as α_{opt}.

However, this parameter selection process does not guarantee sufficient smoothness of spiky oscillation data since the cve criterion in Eq. (11) focuses on prediction accuracy. Moreover, in the case of R, the data oscillation along the wavelength is not even that requires locally adaptive smoothing. Existing locally adaptive smoothing methods have limitations that the frame length cannot be larger than the user-defined predetermined maximum, and the smoothing level is still under-smoothness. To resolve these problems, a new criterion for smoothness control is suggested in this study as

$$\mathrm{len}_{\mathrm{opt}} = \underset{\mathrm{len}}{\arg\min} \left| \frac{\sqrt{\frac{1}{n}\sum \left(\mathbf{Y}_i - \mathrm{pred}^{-i}(\mathbf{x}_i, \mathrm{len}, \alpha_*)\right)^2}}{\max(\mathbf{Y}) - \min(\mathbf{Y})} - S_{\mathrm{cve}} \right| \tag{12}$$

with a specified cve criterion $S_{\mathrm{cve.}}$ This criterion maximizes smoothness as long as S_{cve} is satisfied. The difference between the proposed and existing criteria is shown in Fig. 9. The existing method selects α that minimizes cve while the proposed method selects α that makes normalized cve closest to the given S_{cve}. α_{opt} and $\mathrm{len}_{\mathrm{opt}}$ of the proposed method are larger than those of the existing method that results in higher smoothing. $S_{\mathrm{cve}} = 0.03$ is adopted in the proposed local intensive smoothing. As the polynomial order, intensive smoothing regions adopts one and the other adopts six.

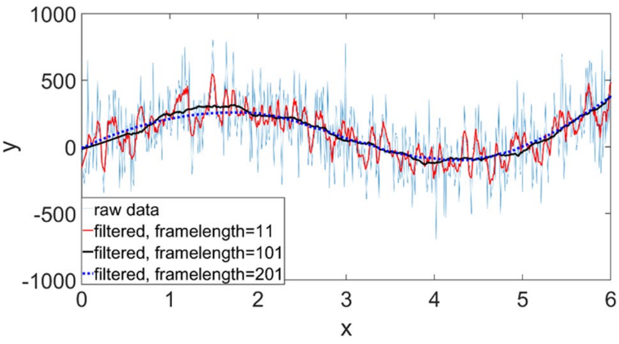

Fig. 6 Effect of frame length in SG filter

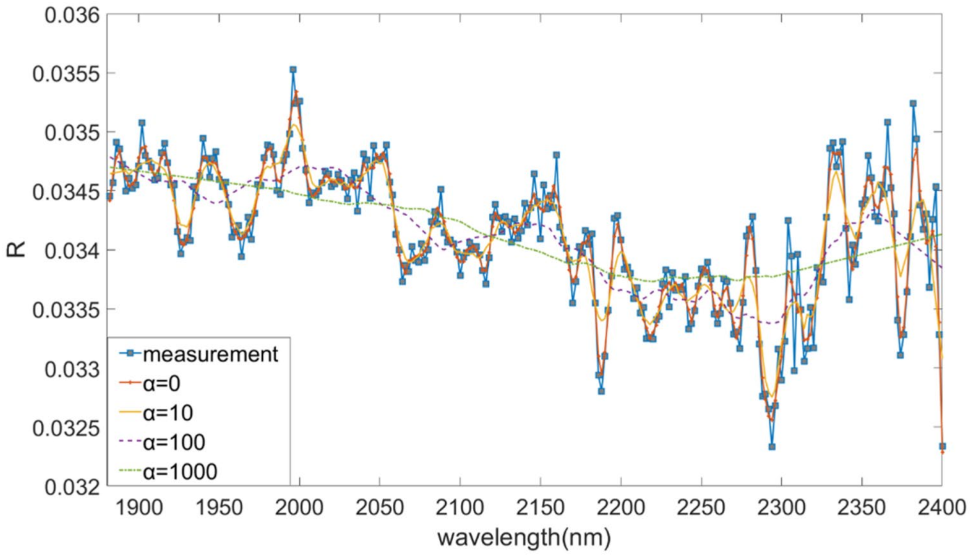

Fig. 7 Effect of α in the proposed intensive smoothing method

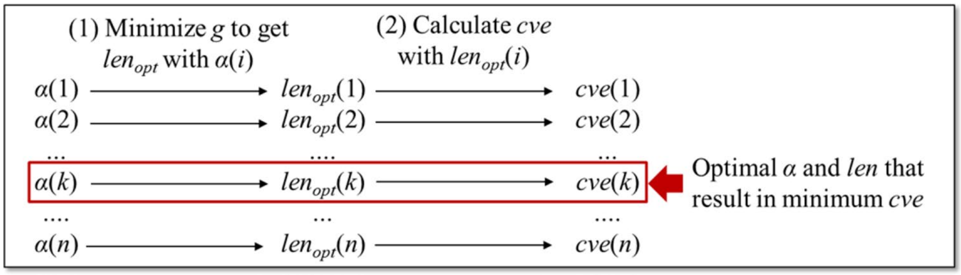
(1) Minimize g to get len_opt with α(i)
(2) Calculate cve with len_opt(i)
α(1)
len_opt(1)
cve(1)
α(2)
len_opt(2)
cve(2)
....
α(k)
len_opt(k)
cve(k)
Optimal α and len that result in minimum cve
α(n)
len_opt(n)
cve(n)

Fig. 8 The parameter selection process of the existing method using Eqs. (10) and (11)

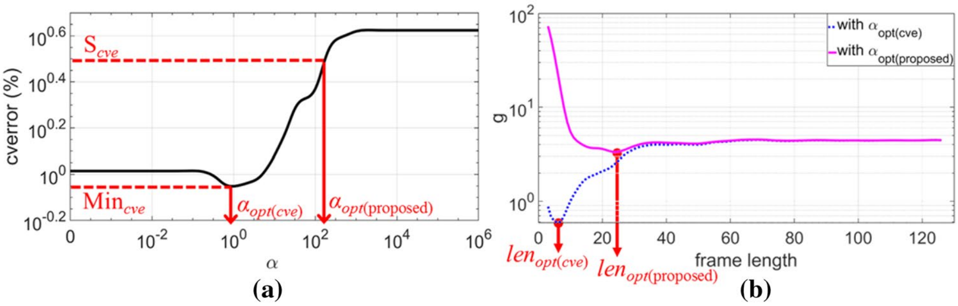

Fig. 9 The proposed method for the determination of **a** α and **b** frame length

3.3 Data integration of multiple measurements

This section explains a method to integrate multiple measurements into a single prediction model using GPR that is a well-known statistical regression modeling method with high accuracy. For derivations of GPR, it is assumed that data $\mathbf{y}$ follows the form of $\mathrm{y}(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$, $\mathbf{f} = f(\mathbf{x})$, where $\varepsilon \sim N(0,\sigma^2)$ with Gaussian noise σ^2 and latent function f [23–28]. This is because it can converge to a normal distribution according to the central limit theorem if the number of noise errors extracted from an independent process is large enough. In Bayesian approach,

$$\mathbf{y} = \mathbf{f} + \sigma^2\mathbf{I} \tag{13}$$

with an identity matrix $\mathbf{I}$ and Gaussian prior as

$$p(\mathbf{f}) \sim N(\mathbf{m}_0, \mathbf{K}_0), \tag{14}$$

where the mean function value $\mathbf{m}_0 = m(\mathbf{x})$ with mean function m, the covariance matrix $(\mathbf{K}_0)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ with a covariance function k. To calculate $p(\mathbf{f}_*|\mathbf{y})$ that is a prediction on new input $\mathbf{x}_*$ with given training data $\{\mathbf{x},\mathbf{y}\}$, the marginalization of joint posterior $p(\mathbf{f},\mathbf{f}_*|\mathbf{y})$ along the latent function value $\mathbf{f}$ is implemented as

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f},\mathbf{f}_*|\mathbf{y})\mathrm{d}\mathbf{f} = \frac{1}{p(\mathbf{y})}\int p(\mathbf{f},\mathbf{f}_*)p(\mathbf{y}|\mathbf{f})\mathrm{d}\mathbf{f} \tag{15}$$

since the joint posterior according to the Bayes' rule is written as

$$p(\mathbf{f},\mathbf{f}_*|\mathbf{y}) = \frac{p(\mathbf{f},\mathbf{f}_*)p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y})}. \tag{16}$$

Each term in the integrand of Eq. (15) is expressed as

$$p(\mathbf{f},\mathbf{f}_*) \sim N\left(\begin{bmatrix}\mathbf{m}_0\\ \mathbf{m}_*\end{bmatrix}, \begin{bmatrix}\mathbf{K}_0 & \mathbf{K}_*\\ \mathbf{K}_*^{\mathrm{T}} & \mathbf{K}_{**}\end{bmatrix}\right) \tag{17}$$

and

$$p(\mathbf{y}|\mathbf{f}) \sim N(\mathbf{f},\sigma^2\mathbf{I}), \tag{18}$$

respectively, where $(\mathbf{K}_*)_{ij} = k(\mathbf{x}_i, \mathbf{x}_{*j})$, $(\mathbf{K}_{**})_{ij} = k(\mathbf{x}_{*i}, \mathbf{x}_{*j})$, and $\mathbf{m}_* = m(\mathbf{x}_*)$ with new input $\mathbf{x}_*$.

Using the previous research [45], Eq. (15) with Eqs. (17) and (18) is rewritten as

$$\begin{aligned}p(\mathbf{f}_*|\mathbf{y}) &= N(\mathbf{m}_{\mathrm{post}}, \mathbf{K}_{\mathrm{post}})\\ &= N(\mathbf{m}_* + \mathbf{K}_*^{\mathrm{T}}(\mathbf{K}_0+\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{m}_0),\ \mathbf{K}_{**} - \mathbf{K}_*^{\mathrm{T}}(\mathbf{K}_0+\sigma^2\mathbf{I})^{-1}\mathbf{K}_*).\end{aligned} \tag{19}$$

When the mean function is defined using polynomials, the mean function $\mathbf{m}_0$ and $\mathbf{m}_*$ are expressed as $\mathbf{H}_0{}^{\mathrm{T}}\boldsymbol{\beta}$ and $\mathbf{H}_*{}^{\mathrm{T}}\boldsymbol{\beta}$ with basis matrices $\mathbf{H}_0$ and $\mathbf{H}_*$, respectively.

All the hyperparameters for the model including $\boldsymbol{\beta}$, σ, and parameters for covariance function are obtained by maximizing marginal likelihood $p(\mathbf{y})$ over the latent function value $\mathbf{f}$ given by

$$\begin{aligned}\log p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\mathrm{d}\mathbf{f}\\ &= -\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{K}_0+\sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\left|\mathbf{K}_0+\sigma^2\mathbf{I}\right| - \frac{n}{2}\log 2\pi\end{aligned} \tag{20}$$

using Eqs. (14) and (18). A more detailed calculation process is found in references [23–28].

3.4 Model selection through physical validity-based bias correction

This section presents a method for bias correction for cases of $T+R>1$ where there is no way to find out which out of T or R causes the bias. Therefore, three cases are considered as shown in Fig. 10: Case 1 assumes that all the bias is caused by R, case 2 assumes that all the bias is caused by T, and case 3 assumes that the bias is caused equally by both T and R. After smoothing and GPR for each case, the best case is selected using the final model simplicity from each case based on the principle of Occam's razor [29–31] that selects the simplest model as the best model. In the proposed process, the model with the highest linearity of the final estimated refractive index among the three cases is selected.

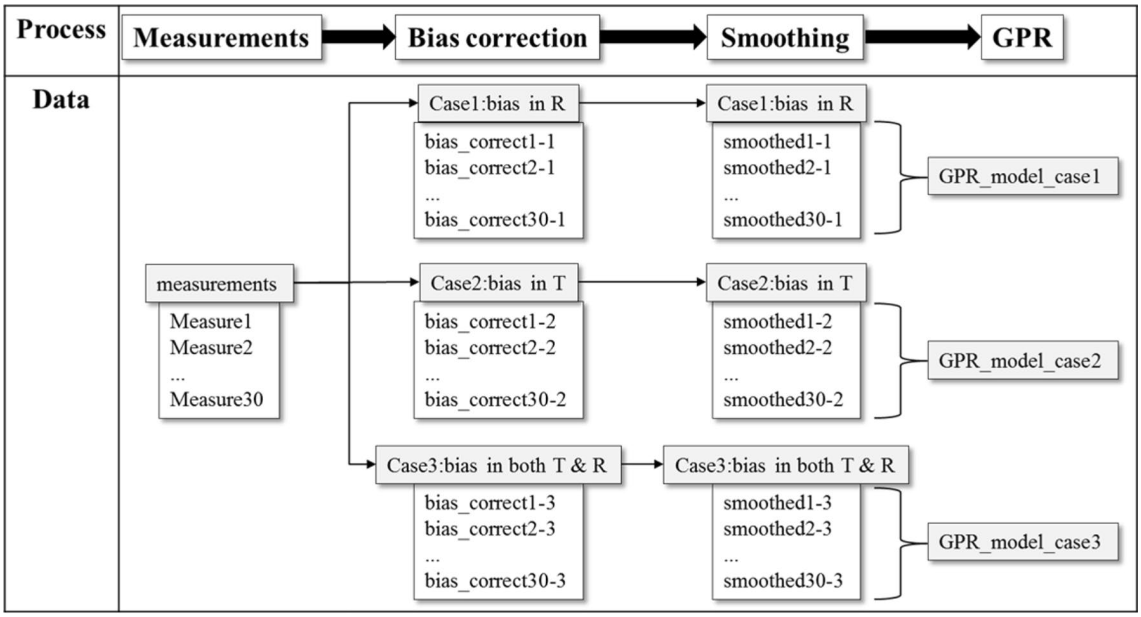

Fig. 10 The overall process of data treatment

According to the process in Fig. 10, Eq. (1) is modified as

$$n(\text{case}, \lambda) = \text{function}(\text{GPR\_T}(\text{case}, \lambda), GPR\text{\_R}(\text{case}, \lambda)), \quad \text{case} = 1,2,3. \tag{21}$$

whose inputs are noise removed and statistically synthesized regression data obtained from the proposed process.

4 Case study: refractive index estimation of water

This section illustrates the estimation results of the refractive index of water using (1) mean of 30 measurements, (2) smoothing only, and (3) bias correction + smoothing. All the smoothed data apply GPR, and the most appropriate bias case among three hypothetical cases assumed in Fig. 10 is selected in the final step. The region segmentation results with the best split points using the method in Sect. 3.1 are shown in Fig. 11. In the case of R, the best number of regions is four as shown in Fig. 5. Figure 11a and b shows the probability density function (PDF) and segmentation results of the third derivatives of four regions of the second data with the smallest value of AICc, respectively. Figure 11b also shows that four regions are appropriately divided according to the fluctuation level. In addition, Fig. 11c shows the division of the regions at similar locations. In this figure, the 31st data are the result of applying the Gaussian mixture method using 30 data averages. Since regions 1, 3, and 4 exceed the criterion of the 3rd derivative as shown in Fig. 4, the regions are intensively smoothed using the proposed method in Sect. 3.2.

Smoothing results of T and R are shown in Fig. 12a and b, respectively. From the region segmentation result in Fig. 11, local intensive smoothing is applied using the proposed criterion to regions 1, 3, and 4 for R. Meanwhile, only region 2 for R is smoothed using the existing criterion because of its low third derivative value. On the other hand, T does not need intensive smoothing over all regions, and thus the existing criterion is adopted as shown in Fig. 12a.

For comparison, the Nadaraya–Watson method [1] and the spline smoothing [7, 8] are applied in regions 1 and 4

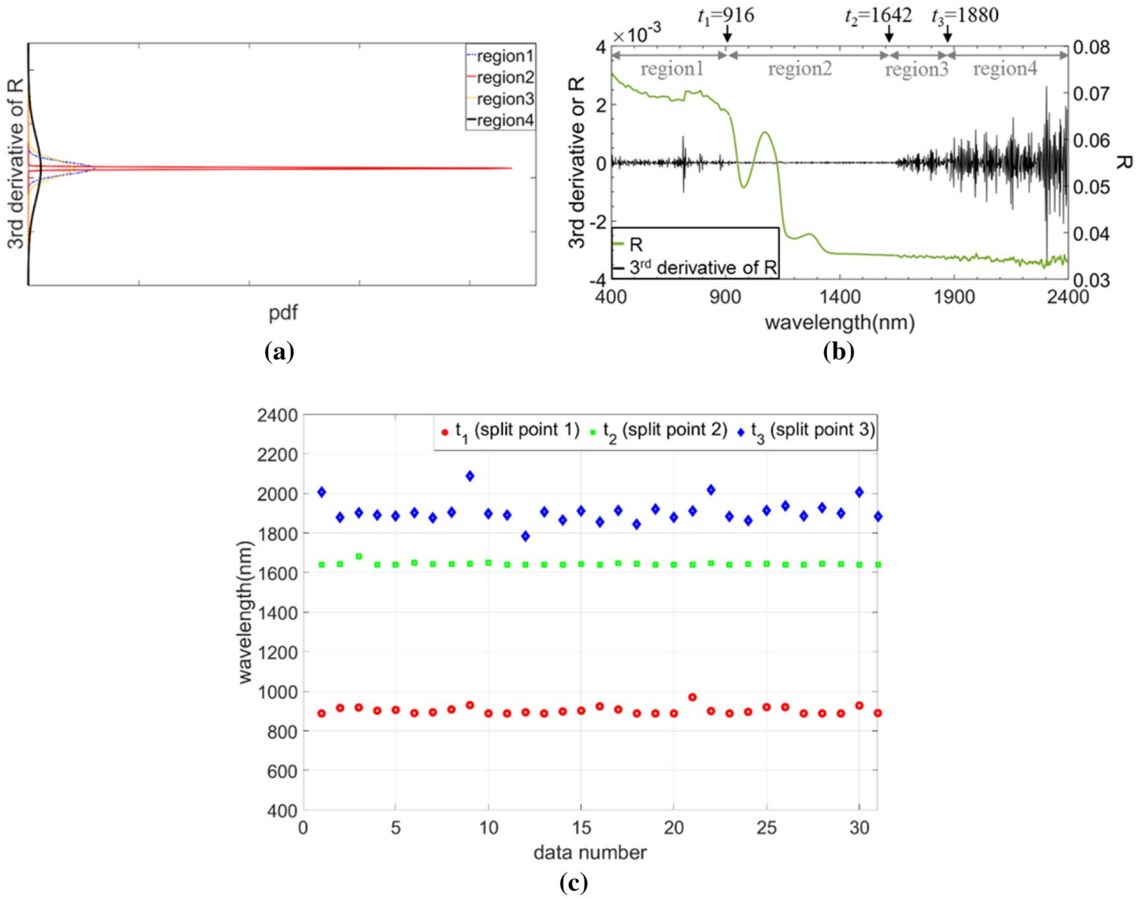

Fig. 11 **a** Distributions of 3rd derivative values in each region of the second data, **b** region segmentation of R according to the 3rd derivative of the second data, **c** region segmentation of R according to the 3rd derivative of all 30 data

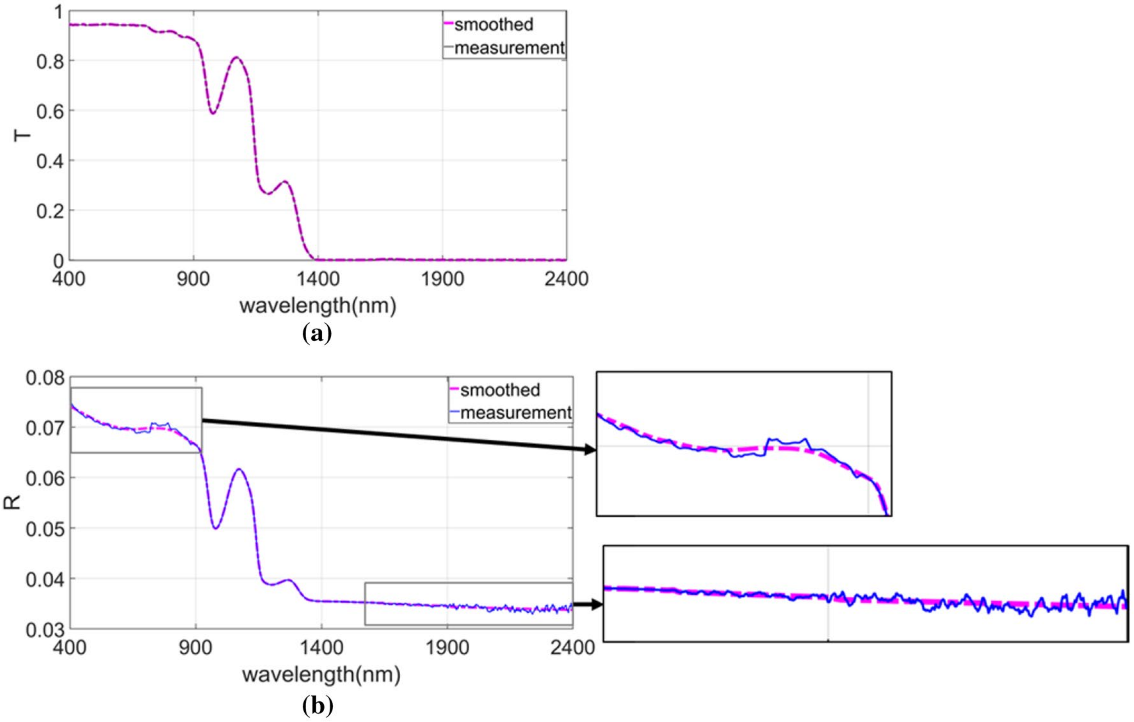

Fig. 12 Data smoothing results of **a** *T* and **b** *R* using the proposed smoothing in Sect. 3.2

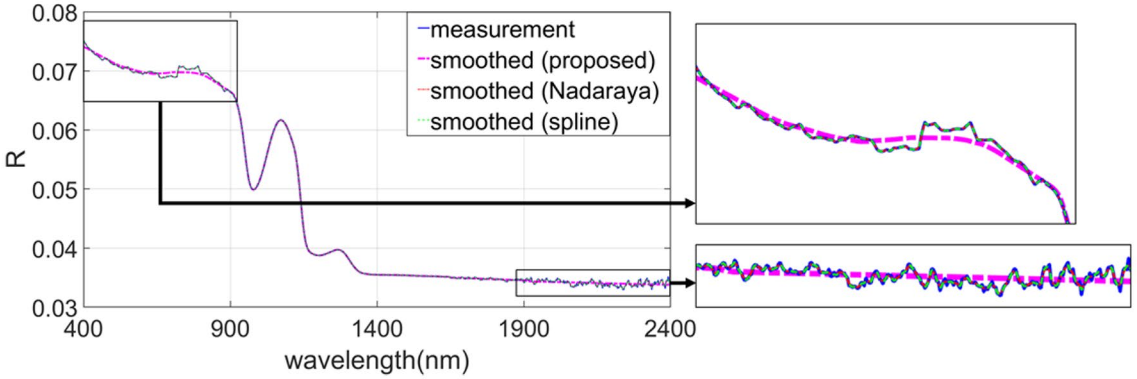

Fig. 13 Comparison of other methods in regions 1 and 4 of *R*

of *R* where the 3rd derivative value is high. Given that the total range of *R* is from 0.03 to 0.08, regions 1 and 4 need to be intensively smoothed because these regions have very small range of *R*. However, as shown in Fig. 13, the spline smoothing and Nadaraya–Watson method do not smooth the given data. On the other hand, the proposed method shows clear smoothing results because users can adjust the smoothing intensity.

Figure 14 illustrates refractive index estimation results of three cases in bias correction as assumed in Fig. 10 after smoothing and GPR. Case 2 is selected as the most probable bias correction case since it shows the highest linearity according to the principle of Occam's razor.

It can be seen from Fig. 15 that the error is gradually reduced according to the process of smoothing, and bias correction and smoothing. Smoothing squashes the noisy peaks and enables stable prediction. Bias correction shows the best result by relocating physically invalid shifted values close to the correct position using the model selection. Moreover, the error estimated from a single measurement is very high in the long-wavelength domain as shown in Fig. 15. If the refractive index was estimated with only one

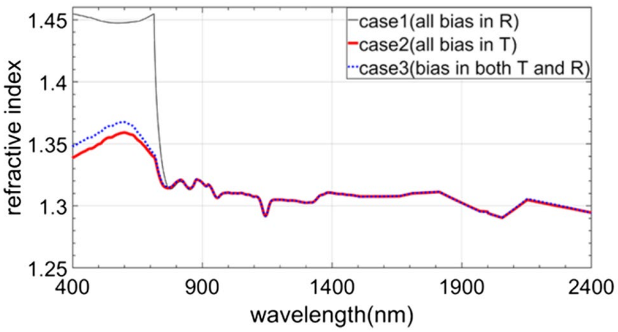

Fig. 14 Refractive index estimation from bias correction according to three cases

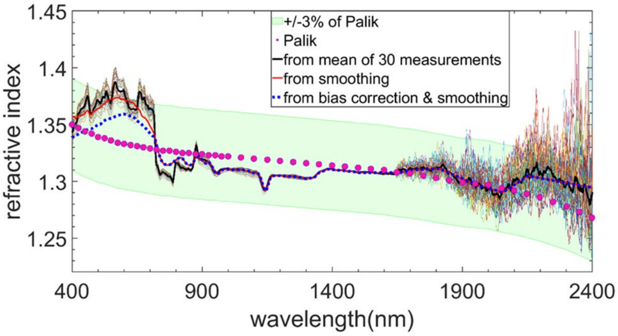

Fig. 15 Comparison of the refractive index estimation accuracy using mean of measurements, smoothing only, and bias correction and smoothing

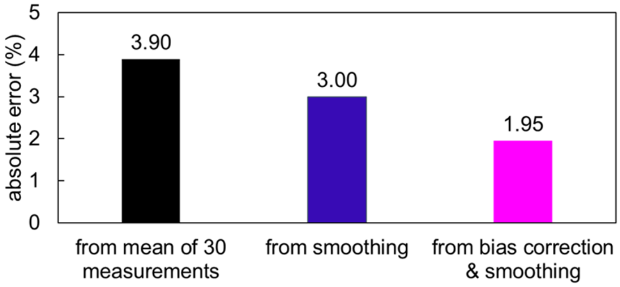

Fig. 16 Estimation error of the refractive index of water using the proposed method

measurement data, the error would be very high, which shows the importance of the statistical approach.

The level of error reduction is quantified in Fig. 16. Bias correction and smoothing shows the highest improvement that the absolute error compared with the result from the mean of 30 measurements is reduced from 3.90% to 1.95%.

5 Conclusion

In this paper, a theoretically reliable process for measurement error treatment is proposed and applied to refractive index estimation of water which cannot be directly measured but can be estimated using other measurable properties such as transmittance T and reflectance R. It is apparent that the measurement errors of T and R are propagated to refractive index, and an appropriate treatment of the errors enhances the resultant estimation accuracy of refractive index. A series of processes consisting of (1) smoothing of spiky fluctuation, (2) data synthesis of multiple measurements into a single prediction model using GPR, and (3) bias correction based on physical validity is proposed for the error treatment method. In addition, a local intensive smoothing criterion with data segmentation method based on data fluctuation level is proposed. The process is validated using a case study of refractive index estimation of water whose true refractive index is known. From the validation, it is shown that the proposed method reduces estimation error by 50% compared with the existing method. In addition, it is shown that the region requiring intensive smoothing is well selected, showing clean smoothing results. Therefore, the novelty of the proposed method is that it is capable of smoothing any type of data by analyzing characteristics of the data and optimizing smoothing parameters. However, there exists weakness of the proposed method in that users need to empirically select S_{cve} according to the characteristics of data. S_{cve} selection method will be a new topic that can be studied in the near future.

Acknowledgements This research was supported by the development of reliability validation method on computational simulation of gas turbine funded by Korea Electric Power Corporation (R17XA05-06). This research was also supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (no. 20172010000850).

References

1. E.A. Nadaraya, On estimating regression. Theory Probab. Appl. **9**(1), 141–142 (1964)
2. C.W. Anderson, V. Barnett, P.C. Chatwin, A.H. El-Shaarawi (eds.), *Quantitative methods for current environmental issues* (Springer, Berlin, 2012)
3. D. Higdon, in *Quantitative Methods for Current Environmental Issues*, ed. by C. Anderson, V. Barnett, P. Chatwin, A. El-Shaarawi. Space and space-time modeling using process convolutions (Springer, London, 2002), pp. 37–56
4. A. Melkumyan, F. Ramos, Multi-kernel Gaussian processes. Twenty-Second International Joint Conference on Artificial Intelligence, 2011

5. A. McHutchon, C.E. Rasmussen, in *Advances in Neural Information Processing Systems*, ed. by J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, K.Q. Weinberger. Gaussian process training with input noise (2011), pp. 1341–1349
6. A. Girard, C.E. Rasmussen, J.Q. Candela, R. Murray-Smith, in *Advances in neural information processing systems*, ed. by S. Becker, S. Thrun, K. Obermayer. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting (2003) pp. 545–552
7. R.L. Eubank, *Nonparametric regression and spline smoothing* (CRC Press, Boca Raton, 1999)
8. B.W. Silverman, Spline smoothing: the equivalent variable kernel method. Ann. Stat. **12**(3), 898–916 (1984)
9. L.R. Rabiner, B. Gold, *Theory and application of digital signal processing* (Prentice-Hall Inc, Englewood Cliffs, 1975), p. 777
10. A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simnplified least squared procedures. Anal. Chem. **36**(8), 1627–1639 (1964)
11. M. Adeghi, F. Behnia, Optimum window length of Savitzky-Golay filters with arbitrary order (2018), arXiv preprint arXiv:1808.10489.
12. J. Li, H. Deng, P. Li, B. Yu, Real-time infrared gas detection based on an adaptive Savitzky-Golay algorithm. Appl. Phys. B **120**(2), 207–216 (2015)
13. D. Acharya, A. Rani, S. Agarwal, V. Singh, Application of adaptive Savitzky-Golay filter for EEG signal processing. Perspect. Sci. **8**, 677–679 (2016)
14. B. Zimmermann, A. Kohler, Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Appl. Spectrosc. **67**(8), 892–902 (2013)
15. S.R. Krishnan, C.S. Seelamantula, On the selection of optimum Savitzky-Golay filters. IEEE Trans. Signal Process. **61**(2), 380–391 (2013)
16. B. Clarke, E. Fokoue, H.H. Zhang, *Principles and theory for data mining and machine learning* (Springer, Berlin, 2009)
17. C.M. Hurvich, J.S. Simonoff, C.L. Tsai, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **60**(2), 271–293 (1998)
18. J. Fan, I. Gijbels, Variable bandwidth and local linear regression smoothers. Ann. Stat. **20**(4), 2008–2036 (1992)
19. J. Fan, *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* (Routledge, London, 2018)
20. M.Y. Cheng, J. Fan, J.S. Marron, *Minimax efficiency of local polynomial fit estimators at boundaries* (University of North Carolina, Chapel Hill, 1993)
21. D. Ruppert, M.P. Wand, Multivariate locally weighted least squares regression. Ann. Stat. **22**(3), 1346–1370 (1994)
22. P. Hall, P. Qiu, Discrete-transform approach to deconvolution problems. Biometrika **92**(1), 135–148 (2005)
23. J. Quiñonero-Candela, C.E. Rasmussen, A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. **6**, 1939–1959 (2005)
24. C.E. Rasmussen, in *Gaussian processes in machine learning*, ed. by O. Bousquet, U. von Luxburg, G. Rätsch. Summer school on machine learning (Springer, Berlin, 2004), pp. 63–71
25. L.S. Bastos, A. O'Hagan, Diagnostics for Gaussian process emulators. Technometrics **51**(4), 425–438 (2009)
26. J.E. Oakley, A. O'Hagan, Probabilistic sensitivity analysis of complex models: a Bayesian approach. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **66**(3), 751–769 (2004)
27. P.D. Kirk, M.P. Stumpf, Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. Bioinformatics **25**(10), 1300–1306 (2009)
28. K. Lee, H. Cho, I. Lee, Variable selection using Gaussian process regression-based metrics for high-dimensional model approximation with limited data. Struct. Multidiscip. Optim. **59**(5), 1439–1454 (2019)
29. A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Occam's razor. Inf. Process. Lett. **24**(6), 377–380 (1987)
30. D. Madigan, A.E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. **89**(428), 1535–1546 (1994)
31. A.E. Raftery, Bayesian model selection in social research. Sociol. Methodol. **25**, 111–164 (1995)
32. J.B. Kim, S. Lee, K. Lee, I. Lee, B.J. Lee, Determination of absorption coefficient of nanofluids with unknown refractive index from reflection and transmission spectra. J. Quant. Spectrosc. Radiat. Transf. **213**, 107–112 (2018)
33. S.K. Das, S.U. Choi, W. Yu, T. Pradeep, *Nanofluids: science and technology* (Wiley, Hoboken, 2007)
34. B.J. Lee, K. Park, T. Walsh, L. Xu, Radiative heat transfer analysis in plasmonic nanofluids for direct solar thermal absorption. J. Sol. Energy Eng. **134**(2), 021009 (2012)
35. J. Jeon, S. Park, B.J. Lee, Optical property of blended plasmonic nanofluid based on gold nanorods. Opt. Express **22**(104), A1101–A1111 (2014)
36. E. Moreels, C. De Greef, R. Finsy, Laser light refractometer. Appl. Opt. **23**(17), 3010–3013 (1984)
37. J.E. Saunders, C. Sanders, H. Chen, H.P. Loock, Refractive indices of common solvents and solutions at 1550 nm. Appl. Opt. **55**(4), 947–953 (2016)
38. M. Daimon, A. Masumura, Measurement of the refractive index of distilled water from the near-infrared region to the ultraviolet region. Appl. Opt. **46**(18), 3811–3820 (2007)
39. K. Lamprecht, W. Papousek, G. Leising, Problem of ambiguity in the determination of optical constants of thin absorbing films from spectroscopic reflectance and transmittance measurements. Appl. Opt. **36**(25), 6364–6371 (1997)
40. T.P. Otanicar, P.E. Phelan, J.S. Golden, Optical properties of liquids for direct absorption solar thermal energy systems. Sol. Energy **83**(7), 969–977 (2009)
41. D. Reynolds, Gaussian mixture models. Encycl. Biom. **741** (2015)
42. G. Xuan, W. Zhang, P. Chai: *EM algorithms of Gaussian mixture model and hidden Markov model*. Proceedings 2001 international conference on image processing IEEE (Cat. no. 01CH37205), vol. 1 (2001), pp. 145–148
43. J. Kuha, AIC and BIC: Comparisons of assumptions and performance. Sociol. Methods Res. **33**(2), 188–229 (2004)
44. N.Z. Sun, A. Sun, *Model calibration and parameter estimation: for environmental and water resource systems* (Springer, Berlin, 2015)
45. R. Von Mises, *Mathematical theory of probability and statistics* (Academic Press, Cambridge, 2014)

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