EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

A new method for error degree estimation in numerical weather prediction via MKDA-based ordinal regression

Takahiro Ogawa*, Shintaro Takahashi, Sho Takahashi and Miki Haseyama

Abstract

This paper presents a new method for estimating error degrees in numerical weather prediction via multiple kernel discriminant analysis (MKDA)-based ordinal regression. The proposed method tries to estimate how large prediction errors will occur in each area from known observed data. Therefore, ordinal regression based on KDA is used for estimating the prediction error degrees. Furthermore, the following points are introduced into the proposed approach. Since several meteorological elements are related to each other based on atmospheric movements, the proposed method merges such heterogeneous features in the target and neighboring areas based on a multiple kernel algorithm. This approach is based on the characteristics of actual meteorological data. Then, MKDA-based ordinal regression for estimating the prediction error degree of a target meteorological element in each area becomes feasible. Since the amount of training data obtained from known observed data becomes very large in the training stage of MKDA, the proposed method performs simple sampling of those training data to reduce the number of samples. We effectively use the remaining training data for determining the parameters of MKDA to realize successful estimation of the prediction error degree.

1 Introduction

Numerical weather prediction is one of the most important research topics in the field of meteorology. Numerical weather prediction enables quantitative calculation of time variation for meteorological elements such as atmospheric pressure, temperature, water vapor content, amount of condensation, wind direction, and wind velocity to estimate their prospective values [1,2]. Recently, the accuracy of numerical weather prediction has improved due to improvements in meteorological observation devices, meteorological models, and computer technology [3].

Many previously reported methods for numerical weather prediction have been based on time-series models. Furthermore, several time-series forecasting methods enable dynamically updating of the forecast according to observed results [4,5]. These methods construct transition models of target elements on the basis of prior knowledge discovered in the target field such as meteorology. This means that most studies are based on bottom-up approaches, and optimal models are constructed for each target element. Therefore, if target elements change, such methods must also use other time-series models. For example, if target elements are time-series with seasonal patterns, the models are constructed on the basis of their seasonal characteristics [6,7].

Although the performance of numerical weather prediction has improved, large prediction errors might still occur in leading edge technologies. In this paper, we define values obtained by subtracting observed values from predicted values calculated by numerical weather prediction as prediction errors, and they can be written by the following equation: [Prediction error] = [Predicted value] − [Observed value].

In this paper, large prediction errors mean errors for which absolute values of [Prediction error] are large. As stated above, since there might be large prediction errors in numerical weather prediction, meteorologists must constantly monitor observed data and compare them to the predicted values. Furthermore, when large prediction errors occur, meteorologists must also modify the forecast

*Correspondence: ogawa@lmd.ist.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814 Japan

based on the obtained errors. Therefore, the development of methodologies for estimating large prediction errors is important to assist meteorologists who monitor observed data and modify the forecast. Although it is preferable to minimize the prediction error instead of predicting the error, the problem of minimization of the prediction error is as difficult as numerical weather prediction. Therefore, in this paper, we first focus on the easier problem.

Generally, we can estimate whether large prediction errors will occur from known observed data based on some classifiers such as a support vector machine (SVM) [8]. This is related to the idea of outlier detection that has recently been studied [9,10]. By using this method, it becomes feasible to assist meteorologists performing early detection of large prediction errors and deciding which areas should be monitored carefully. However, since this method can only detect the occurrence of large errors, it is still difficult to estimate their details such as signs and degrees. Specifically, the estimation of 'whether large prediction errors will occur' enables detection of places where absolute differences between predicted values and observed values are large, e.g., larger than a predefined threshold. Then, by regarding such places as positive examples to construct a classifier, it becomes feasible to estimate the locations of large prediction errors. However, since the obtained classifier performs binary classification, it is difficult to determine whether predicted values are larger or smaller than observed values and how large the prediction errors are.

Therefore, we focus on prediction error degree estimation based on ordinal regression. The reason why we adopt this approach is shown below. If observed values become higher or lower than their prediction values, the prediction errors are closely related to their corresponding problems such as disasters. For example, if the precipitation of the observed data is larger than the forecast, it may cause a flood. On the other hand, if it is smaller, a drought may be caused. Thus, the categorized classes become different. Then, these classes are closely related to the error degrees, which represent the corresponding disasters and their scales. Therefore, the proposed method estimates the prediction error degrees based on the ordinal regression.

Note that specialists also perform numerical weather prediction based on their experiences. Specifically, they have many examples for performing numerical weather prediction and modify their prediction results. Therefore, the example-based approach from many training examples is also useful for prediction error degree estimation in numerical weather prediction. This approach is only based on many training examples and does not require the use of sophisticated meteorological models. Thus, the same scheme can be commonly applied to different meteorological elements. From this point of view, we focus on error degree estimation based on ordinal regression using many training examples.

Several methods have been proposed for ordinal regression that estimates categories of ordinal scale, and they are suitable for solving the above problem.

In this paper, we focus on recent ordinal regression methods based on SVM and discriminant analysis [11-13]. The basic idea for realizing ordinal regression is estimating decision surfaces separating neighboring categories including ordinal scale, and this point is common in other ordinal regression methods that do not adopt SVM and discriminant analysis. The SVM-based methods regard separating hyperplanes as decision surfaces considering ordinal scales to keep the robustness realized by its generalization characteristic. On the other hand, discriminant analysis-based methods maximize between-class variance and minimize within-class variance to separate neighboring categories with maintenance of their ordinal scales. The details of the above methods are shown below.

First, we focus on methods based on an SVM [11,12]. Specifically, SVM-based methods calculate separating hyperplanes, which are determined by only some training data, i.e., support vectors, and all other training data are irrelevant to the determination of hyperplanes. This is the biggest advantage of the SVM-based methods. On the other hand, these methods tend to not grasp the global distribution of target data. Specifically, the distributions of target data have directions. In such cases, the separating rules should be determined according to the directions of class distributions. However, SVM-based methods focus on support vectors, and separating hyperplanes are determined without considering the global distribution of each class, although this point is the biggest contribution of SVM.

Sun et al. proposed an attractive approach for ordinal regression based on kernel discriminant analysis (KDA) in [13]. KDA is a kernel version of linear discriminant analysis (LDA) [14]. It was introduced in [15] and generalized to more than two classes in [16] and [17]. The method in [13] focuses on a characteristic that discriminant analysis can concern global information of data with distribution of classes for classification. Furthermore, by introducing a rank constraint into KDA, successful ordinal regression can be realized. Specifically, LDA-based methods including their kernel versions consider the distribution of each class, i.e., the within-class covariance matrix and the between-class covariance matrix are used for taking into account the distribution direction. Therefore, the projection direction based on discriminant analysis, which can use the global distributions, is obtained for the ordinal regression.

It is expected that the KDA-based method can be extended to a multiple kernel version. Liu et al. proposed a novel unsupervised non-parametric kernel learning

method, which can seamlessly combine the spectral embedding of unlabeled data and manifold regularized least squares to learn non-parametric kernels efficiently [18]. From the above discussion, we focus on the KDA-based ordinal regression extended to a multiple kernel version for error degree estimation in numerical weather prediction in this paper. Many researchers have studied multiple kernel learning (MKL) algorithms [19-21]. Furthermore, when focusing on the parameterized combination approach, it is known that MKL algorithms do not always outperform single kernel-based algorithms if the combination parameters of the MKL algorithms and parameters of the kernels used are not appropriately determined [19]. Thus, their determination is important to guarantee the final estimation performance.

Other than the abovementioned ordinal regression methods, several attractive ordinal regression approaches have been proposed. For example, Srijith et al. proposed a leave-one-out Gaussian process ordinal regression method, which enables model selection based on the leave-one-out cross-validation technique [22]. Furthermore, sparse modeling of Gaussian process ordinal regression enables reduction of computation cost. In [23], neural network threshold ensemble models are proposed, enabling an improvement in the performance of ordinal regression of the threshold models. Seah et al. proposed a new transductive ordinal regression method that introduces the transductive approach into the general ordinal regression problem [24]. Liu et al. proposed a neighborhood preserving ordinal regression method that tries to extract multiple projection directions from the original dataset according to maximum margin and manifold preserving criteria [25].

Finally, we organize our main focus of this paper into several points. In this paper, we try to perform accurate prediction error degree estimation based on ordinal regression. From this point of view, we limit our main focus to the following points.

(i) Prediction error degrees of different kinds of meteorological elements should be estimated in the same manner.
(ii) In numerical weather prediction, there are common characteristics, e.g., use of atmospheric movements is commonly effective for different meteorological elements. Thus, it is suitable to adopt them for prediction error degree estimation.
(iii) Since several meteorological elements affect each other, prediction error degree estimation should be performed by integrating these heterogeneous elements.

The first point is the motivation for adopting an example-based method using ordinal regression instead of adopting a meteorological model-based method. The second and third points are useful for implementing ordinal regression concerning numerical weather prediction. When applying ordinal regression to prediction error degree estimation, we have to consider the above points. So far, these points have not been addressed in previously reported methods.

In this paper, we present a new method for error degree estimation in numerical weather prediction via multiple kernel discriminant analysis (MKDA)-based ordinal regression.

We use ordinal regression based on KDA, which can concern global information of the target data, and try to improve it by using a multiple kernel scheme. The proposed method represents error degrees by multiple ordinal ranks and performs error degree estimation for prospective values from known observed data by using the KDA-based ordinal regression. Furthermore, we introduce the following novel points into this approach to realize successful estimation of prediction error degrees. First, in order to accurately calculate input features used for the ordinal regression, the proposed method uses not only previously observed prediction errors in a target area but also those propagated from its neighboring areas based on atmospheric movements. Since prediction errors tend to be propagated to neighboring areas according to atmospheric movements, this characteristic should be considered in prediction error degree estimation. In the proposed method, the amount of available training data used for the KDA-based ordinal regression, which are extracted from all observed areas, becomes huge. Since the kernel methods depend only on computation involving inner products of those training data in the feature space, it is difficult to directly apply the above approach to such a huge amount of training data. Thus, a simple sampling scheme of training data is adopted, and the use of a kernel method becomes feasible. Fortunately, by using the remaining training data removed after the sampling, the proposed method performs effective parameter determination by the following novel approach. The proposed method tries to introduce the multiple kernel scheme into the KDA-based ordinal regression to merge multiple meteorological elements. Various kinds of meteorological elements can be used for estimating prediction error degrees of a target meteorological element, and their features are heterogeneous. Therefore, in order to merge such heterogeneous features, the proposed method introduces the multiple kernel scheme into the ordinal regression. As described above, a multiple kernel scheme does not always improve the performance [19]. In order to avoid performance degradation, we focus on the use of the remaining large amount of training data removed after the sampling scheme. This means that we separately use training data for performing KDA-based ordinal regression

and the combination parameter settings in the multiple kernel scheme to keep the robustness. Furthermore, not only the combination parameters but also the parameters of the kernels used can be determined by the same manner in our method.

Then, accurate prediction error degree estimation becomes feasible.

2 Prediction error degree estimation via MKDA-based ordinal regression

This section presents a method for error degree estimation in numerical weather prediction using MKDA-based ordinal regression. In this paper, we define prediction error degrees as several ordered discrete ranks. Specifically, we divide the axis of the prediction error into several intervals and assign a symbol, which corresponds to the prediction error degree, for each interval as shown in the example in Figure 1. As shown in Figure 2, the proposed method tries to estimate the unknown prediction error degree that will occur in the forecast in each area from known observed meteorological data. Specifically, from observed data obtained several hours ago, the prediction error degree is estimated for each area based on MKDA-based ordinal regression. In order to perform training of MKDA, we have to provide pairs of features extracted from the observed meteorological data and known prediction error degrees. Therefore, those training pairs are prepared from past observation times, e.g., a few days before the target day for which prediction error degrees are estimated.

Section 2.1 presents details of feature extraction from meteorological data, and Section 2.2 presents an algorithm for estimating the prediction error degree based on ordinal regression using MKDA.

2.1 Feature extraction from meteorological data

This subsection presents details of feature extraction from meteorological data. Using the proposed method, we try to estimate the prediction error degree in each area by using features calculated from 'previously observed errors of a target meteorological element and some other related elements' and 'their time variations' in the same area and its neighboring areas. Only features of neighboring areas in which the atmospheres move to the target area affect the prediction error degree estimation of the target area.

Suppose that prediction error degree estimation of a target meteorological element F_0 is performed in area $\mathbf{p}$ at time t. Furthermore, it is assumed that the prediction errors of several related meteorological elements including F_0, F_l ($l = 0, 1, \cdots, L$; $L+1$ being the number of meteorological elements used for calculating features) at time $t - s\Delta t$ ($s = 0, 1, \cdots, S$) are known, where the index s is used for referring to the current or past time steps. The proposed method calculates time average and maximum and minimum values of the prediction errors $e_l(\mathbf{p}, t)$, $x_l^{\mathrm{ave}}(\mathbf{p}, t)$, $x_l^{\max}(\mathbf{p}, t)$, $x_l^{\min}(\mathbf{p}, t)$, respectively, and the average of their time variations $x_l^{\mathrm{tv}}(\mathbf{p}, t)$ between time $t-(s+1)\Delta t$ and $t-s\Delta t$ for each meteorological element F_l in each area $\mathbf{p}$ as feature values. In this section, $e_l(\mathbf{p}, t)$ represents the prediction error of meteorological element F_l in area $\mathbf{p}$ at time t. In detail, $x_l^{\mathrm{ave}}(\mathbf{p}, t)$, $x_l^{\max}(\mathbf{p}, t)$, $x_l^{\min}(\mathbf{p}, t)$, and $x_l^{\mathrm{tv}}(\mathbf{p}, t)$ can be obtained as follows:

$$x_l^{\mathrm{ave}}(\mathbf{p}, t) = \frac{1}{S+1} \sum_{s=0}^{S} e_l(\mathbf{p}, t - s\Delta t), \tag{1}$$

$$x_l^{\max}(\mathbf{p}, t) = \max_{s=0,1,\cdots,S} e_l(\mathbf{p}, t - s\Delta t), \tag{2}$$

$$x_l^{\min}(\mathbf{p}, t) = \min_{s=0,1,\cdots,S} e_l(\mathbf{p}, t - s\Delta t), \tag{3}$$

$$x_l^{\mathrm{tv}}(\mathbf{p}, t) = \frac{1}{S} \sum_{s=0}^{S-1} x_{l,s}^{\mathrm{tv}}(\mathbf{p}, t), \tag{4}$$

where

$$x_{l,s}^{\mathrm{tv}}(\mathbf{p}, t) = e_l(\mathbf{p}, t - s\Delta t) - e_l(\mathbf{p}, t - (s+1)\Delta t). \tag{5}$$

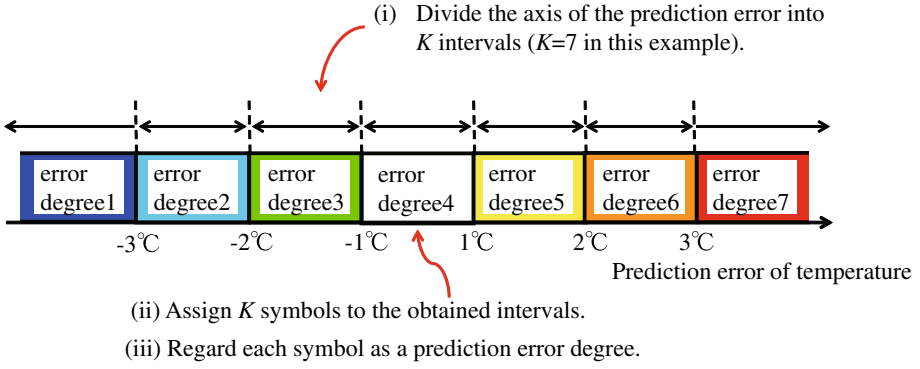

Figure 1 Example of defining the prediction error degree for meteorological element 'temperature'. (i to iii) Procedures for defining the prediction error degrees. In this figure, we set K to 7 as an example, where K is the number of classes, i.e., the number of prediction error degrees.

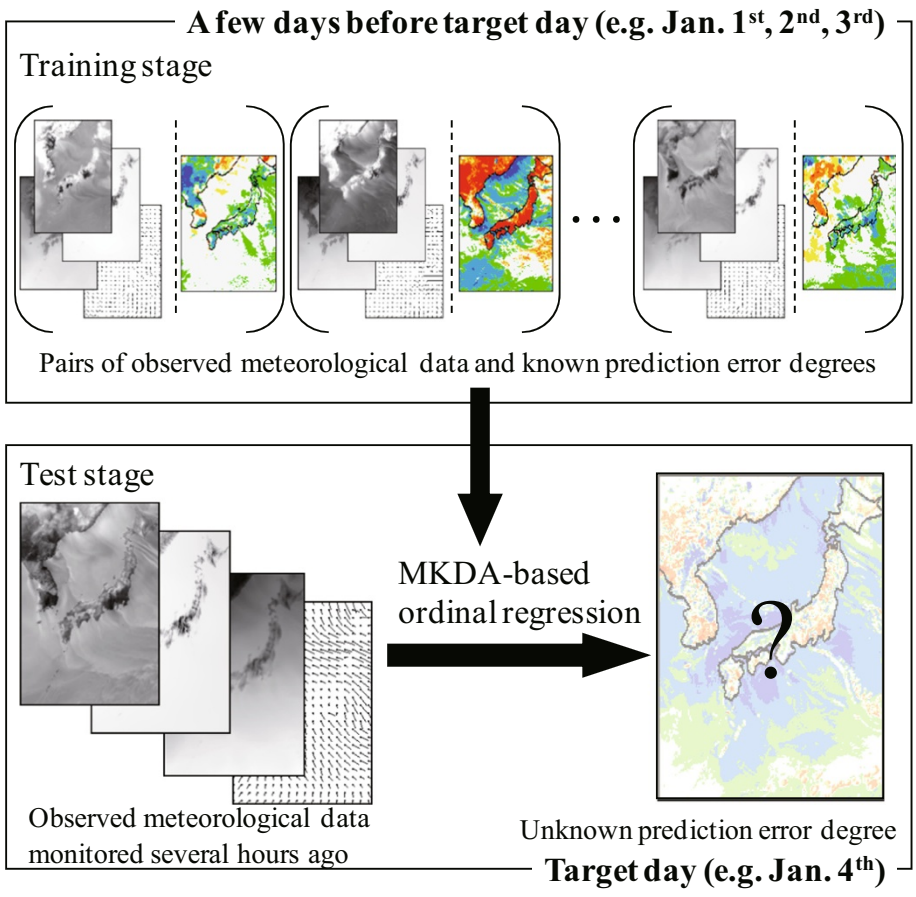

Figure 2 Overview of the proposed method. From pairs of observed meteorological data and known prediction error degrees in past observation times, e.g., a few days before the target day, training of the MKDA is performed. The prediction error degree estimation of the target day is performed from observed known meteorological data via MKDA-based ordinal regression.

For calculating the four features $x_l^{\mathrm{ave}}(\mathbf{p},t)$, $x_l^{\mathrm{max}}(\mathbf{p},t)$, $x_l^{\mathrm{min}}(\mathbf{p},t)$, and $x_l^{\mathrm{tv}}(\mathbf{p},t)$ of the target area $\mathbf{p}$ for meteorological element F_l at time t, the errors $e_l(\mathbf{p},t), e_l(\mathbf{p},t-\Delta t), e_l(\mathbf{p},t-2\Delta t), \cdots, e_l(\mathbf{p},t-S\Delta t)$ are used. This means that errors in the same area $\mathbf{p}$ at current and past times are used for the calculation. Note that in Equations 1 to 5, the index l is used for referring to the lth meteorological element F_l.

Furthermore, in order to use the prediction errors that are propagated from neighboring areas to the target area based on atmospheric movements, motion vectors representing atmospheric movements between time $t-\Delta t$ and time t are obtained by using observed wind velocities for each area. Then, the feature $x_l^{\mathrm{neighbor}}(\mathbf{p},t)$ of the prediction error for meteorological element F_l propagated from the neighboring areas to the target area $\mathbf{p}$ is obtained by the following equation:

$$x_l^{\mathrm{neighbor}}(\mathbf{p},t) = \mathrm{average}_{\mathbf{p}^*\in R_{(\mathbf{p},t)}}\left[e_l(\mathbf{p}^*,t-\Delta t)\right], \quad (6)$$

where average $[\cdot]$ is an operator calculating the average values. Furthermore, $R_{(\mathbf{p},t)}$ represents a set of areas in which atmospheres move to the target area $\mathbf{p}$ from time $t-\Delta t$ to time t. Specifically, by denoting the atmospheric movement of area $\mathbf{p}^*$ from time $t-\Delta t$ to time t as $\mathbf{v}(\mathbf{p}^*,t-\Delta t)$, $R_{(\mathbf{p},t)}$ can be represented as follows:

$$R_{(\mathbf{p},t)} = \left\{\mathbf{p}^* | \mathbf{p}^* + \mathbf{v}\left(\mathbf{p}^*, t-\Delta t\right) \approx \mathbf{p}\right\}. \quad (7)$$

By using the atmospheric movements, we can select areas in which atmospheres move to the target area $\mathbf{p}$ from time $t-\Delta t$ to time t. As shown in Equation 7, we define the neighbor of $\mathbf{p}$. In this equation, the neighbor is a set of areas $\mathbf{p}^*$ satisfying $\mathbf{p}^* + \mathbf{v}\left(\mathbf{p}^*, t-\Delta t\right) \approx \mathbf{p}$. This means $\|\mathbf{p}^* + \mathbf{v}\left(\mathbf{p}^*, t-\Delta t\right) - \mathbf{p}\|_c < \epsilon$ is satisfied for a small positive constant ϵ. Note that $\|\cdot\|_c$ represents the chessboard distance. In this study, we set ϵ to 5 km. Since the distance between the most neighboring areas is 5 km in the dataset used in the experiments, we set ϵ to 5 km. Therefore, if the distance between the most neighboring areas

changes, we should also change ϵ. Furthermore, it is well known that if the distance between the most neighboring areas becomes smaller, the performance of the numerical weather prediction also becomes better. Similarly, it can be expected that the performance of the prediction error degree estimation becomes better if the distance between the most neighboring areas becomes smaller. As shown in the example in Figure 3, the atmospheres of six areas, i.e., yellow areas, move to the target area **p**. Therefore, these six areas are regarded as neighbors. Then, the prediction errors $e_l(\mathbf{p}^*, t-\Delta t)$ of these six areas $\mathbf{p}^*$ in the previous time step $t-\Delta t$ are averaged, and the feature $x_l^{\text{neighbor}}(\mathbf{p}, t)$ is obtained. If none of the areas are moving into the target area **p**, $x_l^{\text{neighbor}}(\mathbf{p}, t)$ is set to zero in our method. By using the features in Equation 6, the influence of prediction errors propagated to the target area **p** can be considered.

From the features $x_l^{\text{ave}}(\mathbf{p},t), x_l^{\max}(\mathbf{p},t), x_l^{\min}(\mathbf{p},t), x_l^{\text{tv}}(\mathbf{p},t)$, and $x_l^{\text{neighbor}}(\mathbf{p},t)$ $(l = 0, 1, \cdots, L)$, we can define a feature vector for each area **p** at time t. Note that these five features in area **p** at time t for meteorological element F_l can be calculated from several isometric surfaces in the proposed method. Therefore, given the number of isometric surfaces as J, $5J$ features are obtained in area **p** at time t for each meteorological element F_l. Then, for each area **p** at time t, a total of $d\,(=5J\times(L+1))$ features is obtained, i.e., a d-dimensional feature vector is finally obtained.

2.2 Algorithm for estimation of prediction error degree

This subsection presents an algorithm for estimating prediction error degrees using MKDA-based ordinal regression. The proposed method estimates class labels, i.e., prediction error degrees at each area **p**, from the features obtained as described in the previous subsection.

First, we denote a set of training data $(\mathbf{x}_i, y_i) \in \mathbf{R}^d \times \mathbf{R}$ $(i = 1, 2, \cdots, M$; M being the number of training samples) as $\mathcal{T}_M$. Each $\mathbf{x}_i \in \mathbf{R}^d$ is a d-dimensional (d being the number of features shown in the previous subsection) input feature vector, and $y_i \in \{1, 2, \cdots, K\}$ is the corresponding ordered class label, where K is the number of classes. This label can be obtained by quantizing the known prediction error of target meteorological element F_0 into K ranks as shown in the example in Figure 1. In Figure 1, K is set to seven. Furthermore, the proposed method maps $\mathbf{x}_i$ into the feature space via a nonlinear map [26], and $\phi(\mathbf{x}_i) \in \mathcal{F}$ is obtained. In the proposed method, $\phi(\cdot)$ is a nonlinear map which projects an input vector to high-dimensional feature space. Furthermore, $\mathcal{F}$ represents this high-dimensional feature space. Note that its dimension depends on the definition of the corresponding kernel function of the nonlinear map $\phi(\cdot)$. Since $\phi(\mathbf{x}_i)$ is high-dimensional or infinite-dimensional, it may not be possible to calculate them directly. Fortunately, it is well known that the following computational procedures depend only on the inner products in the feature space, which can be obtained from a suitable kernel function [26]. Given two vectors $\mathbf{x}_m$ and $\mathbf{x}_n \in \mathbf{R}^d$, our method uses the following multiple kernel functions:

$$\phi(\mathbf{x}_m)'\,\phi(\mathbf{x}_n) = \sum_{l=0}^{L} \gamma_l \kappa_l(\mathbf{E}_l\mathbf{x}_m, \mathbf{E}_l\mathbf{x}_n), \tag{8}$$

where γ_l represents the weight of the lth kernel $\kappa_l(\cdot,\cdot)$ and satisfies $\gamma_l \geq 0$ and $\sum_{l=0}^{L}\gamma_l = 1$. The superscript $'$ denotes the vector/matrix transpose in this paper. Furthermore, $\mathbf{E}_l$ in Equation 8 is a diagonal matrix whose diagonal elements are zero or one, and it enables the extraction of

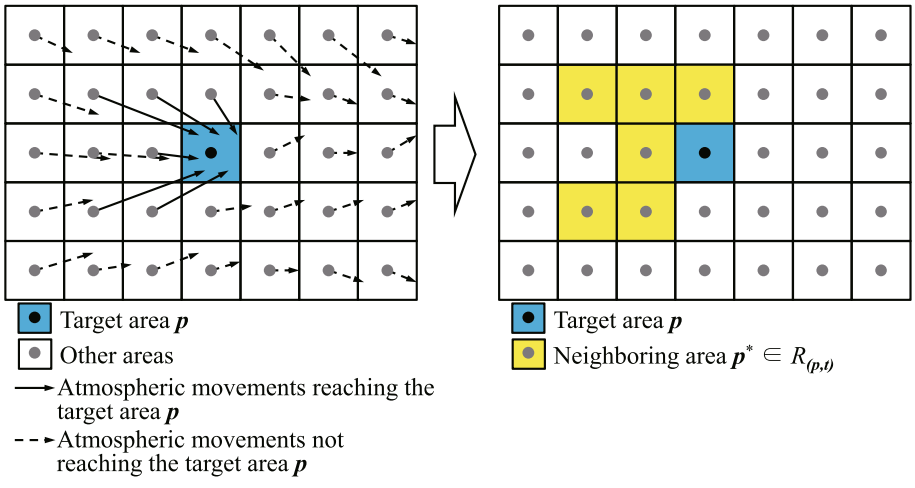

Figure 3 Example of selecting neighboring areas of target area p.

features of meteorological element F_l that are used for the lth kernel $\kappa_l(\cdot,\cdot)$. Specifically, the dimension of vectors $\mathbf{x}_m$ and $\mathbf{x}_n$ is $d(=5J\times(L+1))$ as defined in the previous subsection, and each extraction matrix $\mathbf{E}_l$ extracts $5J$ features of each meteorological element F_l. In the proposed method, the multiple kernel scheme is applied to different meteorological elements. Since we use $L+1$ kinds of meteorological elements, $L+1$ kernels are linearly combined in Equation 8. As shown in the previous subsection, various kinds of meteorological elements can be used for estimation of the prediction error degree. Thus, the proposed method extends the kernel function to a multiple kernel version as shown in Equation 8. Then, by successfully determining the parameters γ_l $(l=0,1,\cdots,L)$ in the multiple kernel function, the features can be mapped into the optimal feature space, enabling accurate ordinal regression. It is important to successfully determine the parameters γ_l, and details of their determination are shown in Section 2.2.2. Note that $\kappa_l(\cdot,\cdot)$ of each meteorological element F_l is set to the well-known Gaussian kernel in our method.

2.2.1 Sampling of training data

Note that when the kernel method is adopted, direct use of MKDA, for which computation depends only on the inner products in the feature space, becomes difficult due to the large amount of training data.

Therefore, the proposed method uses a sampling scheme.

Specifically, we regard the error data $e_l(\mathbf{p},t)$ at time t as two-dimensional signals and perform their downsampling.

Then, the new sampled training data $(\mathbf{x}_i,y_i)$ $(i=1,2,\cdots,N;N<M)$ can be obtained, where $(\mathbf{x}_i,y_i)$ is redefined, and N is the number of new training samples. In the following explanations of training of MKDA, we use these training data $(\mathbf{x}_i,y_i)$ $(i=1,2,\cdots,N)$. Also, we denote a set of these sampled training data as $\mathcal{T}_N$.

By reducing the number of training samples, the performance of the KDA-based ordinal regression tends to become worse. Note that in the proposed method, we regard the error data $e_l(\mathbf{p},t)$ as two-dimensional signals and perform their downsampling. Generally, neighboring areas in meteorological data tend to have similar features, and it seems that the distribution of training data is not drastically changed by the sampling. Thus, performance degradation tends to be avoided. Furthermore, in the proposed method, the remaining training data in $\mathcal{T}_M-\mathcal{T}_N$, which are removed by the sampling, can be used for estimating γ_l $(l=0,1,\cdots,L)$ in Equation 8. Fortunately, by using these remaining data, we can improve the performance of the error degree estimation based on the multiple kernel scheme. The details are shown in 2.2.3.

2.2.2 Derivation of MKDA

The objective of the discriminant analysis is to find a projection $\mathbf{w}$ from which different classes can be well separated. Specifically, we first define within-class and between-class scatter matrices as follows:

$$\begin{aligned}\mathbf{S}_w&=\frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{N^k}\left\{\phi\left(\mathbf{x}_j^k\right)-\mathbf{m}^k\right\}\left\{\phi(\mathbf{x}_j^k)-\mathbf{m}^k\right\}'\\&=\frac{1}{N}\sum_{k=1}^{K}\boldsymbol{\Xi}^k\mathbf{H}^k\boldsymbol{\Xi}^{k'},\end{aligned}\tag{9}$$

$$\mathbf{S}_b=\frac{1}{N}\sum_{k=1}^{K}N^k\left(\mathbf{m}^k-\mathbf{m}\right)\left(\mathbf{m}^k-\mathbf{m}\right)',\tag{10}$$

where

$$\begin{aligned}\mathbf{m}^k&=\frac{1}{N^k}\sum_{j=1}^{N^k}\phi\left(\mathbf{x}_j^k\right)\\&=\frac{1}{N^k}\boldsymbol{\Xi}^k\mathbf{1}^k.\end{aligned}\tag{11}$$

In the above equations, $\mathbf{x}_j^k$ $\left(j=1,2,\cdots,N^k\right)$ corresponds to $\mathbf{x}_i$ $(i=1,2,\cdots,N)$ belonging to the kth class (i.e., $y_i=k$), and $\mathbf{m}^k$ denotes the mean vector of $\phi(\mathbf{x}_j^k)$ belonging to the kth class as shown in Equation 11. Furthermore, $\boldsymbol{\Xi}^k=\left[\phi\left(\mathbf{x}_1^k\right),\phi\left(\mathbf{x}_2^k\right),\cdots,\phi(\mathbf{x}_{N^k}^k)\right]$, and $\mathbf{H}^k=\mathbf{I}^k-\frac{1}{N^k}\mathbf{1}^k\mathbf{1}^{k'}$ is a centering matrix satisfying $\mathbf{H}^{k'}=\mathbf{H}^k$ and $\left(\mathbf{H}^k\right)^2=\mathbf{H}^k$, where $\mathbf{I}^k$ is the $N^k\times N^k$ identity matrix, and $\mathbf{1}^k=[1,1,\cdots,1]'$ is an $N^k\times 1$ vector. The vector $\mathbf{m}=\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i)$ in Equation 10 stands for the mean vector of all samples $\mathbf{x}_i$ $(i=1,2,\cdots,N)$. Generally, the objective of the discriminant analysis can be achieved by the following equation:

$$\min J(\mathbf{w})=\frac{\mathbf{w}'\mathbf{S}_w\mathbf{w}}{\mathbf{w}'\mathbf{S}_b\mathbf{w}}.\tag{12}$$

In the proposed method, we perform ordinal regression with K ordered classes. Therefore, the goal of our method is to find the optimal projection $\mathbf{w}$ satisfying the following two points:

1. The projection $\mathbf{w}$ should minimize the within-class distance and maximize the between-class distance simultaneously.
2. The projection $\mathbf{w}$ should ensure ordinal information of different classes, i.e., the average projection of

samples from higher rank classes should be larger than that of lower rank classes.

Therefore, the formulation for the MKDA-based ordinal regression is derived from [13] as follows:

$$\begin{aligned}\min J(\mathbf{w},\rho) = \mathbf{w}'\mathbf{S}_w\mathbf{w} - C\rho \ \text{ s.t. } \ \mathbf{w}'\left(\mathbf{m}^{k+1}-\mathbf{m}^k\right)\\ \geq \rho, \ \ k=1,2,\cdots,K-1,\end{aligned} \tag{13}$$

where $C > 0$ represents a penalty coefficient. The above equation minimizes the within-class distances. Furthermore, instead of using the between-class scatter matrix directly, the above equation tries to maximize the distance between the two projected means from the closest pair of classes. Specifically, Equation 12 tries to minimize the within-class distance and maximize the between-class distance simultaneously. On the other hand, Equation 13 reformulates the problem of Equation 12. The within-class distance is minimized from $\mathbf{w}'\mathbf{S}_w\mathbf{w}$ in $J(\mathbf{w},\rho)$. Furthermore, instead of directly maximizing the between-class distance, a new constraint $\mathbf{w}'\left(\mathbf{m}^{k+1}-\mathbf{m}^k\right) \geq \rho$ $(k=1,2,\cdots,K-1)$ is introduced. In this way, our MKDA-based ordinal regression tries to estimate the projection $\mathbf{w}$ minimizing within-class distance with the constraint of ordinal information. Thus, the distribution direction can be considered by using the within-class scatter matrix $\mathbf{S}_w$ in Equation 13.

In MKDA, the projection $\mathbf{w}$ is high-dimensional or infinite-dimensional and cannot be calculated directly. Thus, the projection $\mathbf{w}$ is written as follows:

$$\begin{aligned}\mathbf{w} &= \sum_{i=1}^{N}\beta_i\phi\left(\mathbf{x}_i\right)\\ &= \Xi\boldsymbol{\beta},\end{aligned} \tag{14}$$

where $\beta_i (i=1,2,\cdots,N)$ is a linear coefficient, and $\boldsymbol{\beta} = [\beta_1,\beta_2,\cdots,\beta_N]'$. In addition, $\Xi = [\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),\cdots,\phi(\mathbf{x}_N)]$. In the proposed method, we derived Equation 14 by using the method in [13]. In several kernel methods such as KDA and KPCA, the projection is represented by a linear combination of the samples. Therefore, we adopt the above derivation. However, derivation of the representer theorem in a multiple kernel case is not so straightforward. In the proposed method, we approximately use the derivation of Equation 14. Since the theoretical analysis of this approximation cannot be shown in this paper, this will be a future work of our study.

By using Equations 9, 11, and 14 and $J(\mathbf{w},\rho)$ and $\mathbf{w}'\left(\mathbf{m}^{k+1}-\mathbf{m}^k\right)$ in Equation 13 are respectively rewritten as follows:

$$\begin{aligned}J(\mathbf{w},\rho) &= (\Xi\boldsymbol{\beta})'\left(\frac{1}{N}\sum_{k=1}^{K}\Xi^k\mathbf{H}^k\Xi^{k'}\right)\Xi\boldsymbol{\beta} - C\rho\\ &= \boldsymbol{\beta}'\left\{\frac{1}{N}\sum_{k=1}^{K}\left(\Xi'\Xi^k\right)\mathbf{H}^k\left(\Xi^{k'}\Xi\right)\right\}\boldsymbol{\beta} - C\rho\\ &= \boldsymbol{\beta}'\left\{\frac{1}{N}\sum_{k=1}^{K}\mathbf{G}^k\mathbf{H}^k\mathbf{G}^{k'}\right\}\boldsymbol{\beta} - C\rho\\ &= \boldsymbol{\beta}'\mathbf{T}\boldsymbol{\beta} - C\rho,\end{aligned} \tag{15}$$

$$\begin{aligned}\mathbf{w}'\left(\mathbf{m}^{k+1}-\mathbf{m}^k\right) &= (\Xi\boldsymbol{\beta})'\left(\frac{1}{N^{k+1}}\Xi^{k+1}\mathbf{1}^{k+1} - \frac{1}{N^k}\Xi^k\mathbf{1}^k\right)\\ &= \boldsymbol{\beta}'\left(\frac{1}{N^{k+1}}\Xi'\Xi^{k+1}\mathbf{1}^{k+1} - \frac{1}{N^k}\Xi'\Xi^k\mathbf{1}^k\right)\\ &= \boldsymbol{\beta}'\left(\frac{1}{N^{k+1}}\mathbf{G}^{k+1}\mathbf{1}^{k+1} - \frac{1}{N^k}\mathbf{G}^k\mathbf{1}^k\right)\\ &= \boldsymbol{\beta}'\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right),\end{aligned} \tag{16}$$

where

$$\mathbf{G}^k = \Xi'\Xi^k, \tag{17}$$

$$\mathbf{T} = \frac{1}{N}\sum_{k=1}^{K}\mathbf{G}^k\mathbf{H}^k\mathbf{G}^{k'}, \tag{18}$$

$$\mathbf{r}^k = \frac{1}{N^k}\mathbf{G}^k\mathbf{1}^k. \tag{19}$$

The problem of $\mathbf{w}$ in Equation 13 can be rewritten as that of $\boldsymbol{\beta}$ as follows:

$$\begin{aligned}\min J(\boldsymbol{\beta},\rho) = \boldsymbol{\beta}'\mathbf{T}\boldsymbol{\beta} - C\rho \ \text{ s.t. } \ \boldsymbol{\beta}'\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)\\ \geq \rho, \ k=1,2,\cdots,K-1.\end{aligned} \tag{20}$$

In order to solve Equation 20, we define the following Lagrangian equation:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta},\rho,\boldsymbol{\alpha}) = \boldsymbol{\beta}'\mathbf{T}\boldsymbol{\beta} - C\rho\\ - \sum_{k=1}^{K-1}\alpha^k\left\{\boldsymbol{\beta}'\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)-\rho\right\},\end{aligned} \tag{21}$$

where $\boldsymbol{\alpha} = \left[\alpha^1,\alpha^2,\cdots,\alpha^{K-1}\right]'$ represents a vector containing Lagrange multipliers. Then, by calculating $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = 0$

and $\frac{\partial \mathcal{L}}{\partial \rho} = 0$, the following equations are respectively obtained:

$$\boldsymbol{\beta} = \frac{1}{2}\mathbf{T}^{-1}\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right), \tag{22}$$

$$\sum_{k=1}^{K-1}\alpha^k = C. \tag{23}$$

From the above equations, the optimization problem in Equation 20 is turned into

$$\begin{aligned}\min(\boldsymbol{\alpha}) = &\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)'\mathbf{T}^{-1}\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)\\ &\text{s.t.}\quad \alpha^k \geq 0,\quad k=1,2,\cdots,K-1,\quad \text{and}\\ &\sum_{k=1}^{K-1}\alpha^k = C.\end{aligned} \tag{24}$$

The proposed method estimates the optimal result of $\boldsymbol{\alpha}$ by using the penalty method [27].

2.2.3 MKDA-based ordinal regression and determination of kernels' contributions

As shown in the above explanation, we can obtain the optimal projection of $\mathbf{w}$ from $\boldsymbol{\beta}$ obtained by $\boldsymbol{\alpha}$.

From the obtained optimal projection $\mathbf{w}$, the rank of an unseen input feature vector $\mathbf{x}$, i.e., the prediction error degree at each area $\mathbf{p}$, can be estimated by the following decision rule:

$$f(\mathbf{x}) = \min_{k\in\{1,2,\cdots,K\}}\left\{k : \mathbf{w}'\phi(\mathbf{x}) - b_k < 0\right\}, \tag{25}$$

where b_k is defined as

$$b_k = \frac{\boldsymbol{\beta}'\left(\mathbf{r}^{k+1}+\mathbf{r}^k\right)}{2}. \tag{26}$$

Then, from Equations 14 and 26, Equation 25 is rewritten as follows:

$$f(\mathbf{x}) = \min_{k\in\{1,2,\cdots,K\}}\left\{k : \boldsymbol{\beta}'\left(\boldsymbol{\Xi}'\phi(\mathbf{x}) - \frac{\mathbf{r}^{k+1}+\mathbf{r}^k}{2}\right) < 0\right\}. \tag{27}$$

The above equation enables ordinal regression for estimating prediction error degrees.

Note that since our method adopts a multiple kernel algorithm, we also have to determine $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \cdots, \gamma_L]'$ in Equation 8. Some methods such as simple MKL [21] have been proposed for determining $\boldsymbol{\gamma}$. However, if the simple MKL is used for estimating the parameters of the linear combination of kernels, successful performance of the error degree estimation is not possible. We guess when using the simple MKL, the result of $\boldsymbol{\gamma}$ that provides $\mathbf{w}$ optimal for Equation 13 is obtained from the sampled training data, and the generalization characteristic becomes worse, and then a phenomenon similar to overfitting occurs. This means that the fitting of $\boldsymbol{\gamma}$ tends to strongly depend on the sampled training data. Furthermore, it has been reported that the MKL approach does not always outperform a single kernel-based approach [19]. Thus, it is important to determine $\boldsymbol{\gamma}$ in such a way that the performance can be guaranteed for other new samples to keep the robustness.

Fortunately, we have the other remaining training data $(\mathbf{x}_i, y_i)$ included in $\mathcal{T}_M - \mathcal{T}_N$, which are removed by the sampling scheme. Therefore, we use them and verify the estimation performance of the error degrees, and the best result of $\boldsymbol{\gamma}$ is determined by an exhaustive search. Generally, the statistical characteristics of the test data tend to be slightly different from that of the training data. Therefore, in order to make the proposed method robust to this difference, we use the other remaining training data included in $\mathcal{T}_M - \mathcal{T}_N$. Specifically, the proposed method changes the values of γ_l $(l = 0, 1, 2, \cdots, L)$ as $0.1, 0.2, \cdots, 1.0$ with the constraints of $\gamma_l \geq 0$ $(l = 0, 1, 2, \cdots, L)$ and $\sum_{l=0}^{L}\gamma_l = 1.0$. Then, the problem in Equation 28 is optimized with $\boldsymbol{\alpha}$. Furthermore, for the data $(\mathbf{x}_i, y_i)$ included in $\mathcal{T}_M - \mathcal{T}_N$, which are the remaining training data after the sampling, we perform the estimation of prediction error degrees and calculate their mean absolute errors.

Specifically, the errors between the estimated error degree $f(\mathbf{x}_i)$ in Equation 31 and the true rank y_i are calculated. Finally, we output the optimal result of $\boldsymbol{\gamma}$ that provides the most accurate estimation results, i.e., the minimum mean absolute error. By using this exhaustive search procedure, the proposed method enables determination of the multiple kernel function. Note that in the same way as $\boldsymbol{\gamma}$, the proposed method estimates the parameter of the Gaussian kernel for each meteorological element, where the searching interval is set to 0.5.

In this way, we can perform prediction error degree estimation by using ordinal regression based on MKDA. In numerical weather prediction, various observed inputs, i.e., various input feature vectors, are obtained. Since discriminant analysis can consider the global information of the data with the distribution of the classes, the proposed method adopts it for the estimation. Furthermore, the proposed method adopts sampling of the training data and effectively uses the remaining data for estimating the optimal parameters of the multiple kernel scheme. This is the biggest difference between the proposed method and the conventional KDA-based method.

3 Experimental results

In order to verify the performance of the proposed method, this section shows results obtained by applying the proposed method to real data of numerical weather prediction. In this experiment, we used three datasets

obtained by numerical weather prediction performed in January 2010, the data being provided by the Japan Weather Association. As shown in Figure 4, each dataset contains prediction and observed data for 4 days, and the three datasets are obtained by using a sliding window including 4 days, the sliding interval of which is set to half a day. We assumed that the data at 3, 6, and 9 h after the beginning of the forecast were all known to calculate the feature vector, and we performed error degree estimation at 12 h after the beginning of the forecast by using these known data. Therefore, $S = 2$, i.e., $S + 1 = 3$ time steps were used for calculating feature vectors to estimate error degrees in these experiments. Furthermore, Δt was 3 h. Then, the training procedures were performed by using the data for the first 3 days, and verification of the prediction error degree estimation was performed for the remaining 1 day, i.e., the training and test corresponded to the first 3 days and the remaining 1 day (fourth day), respectively. The data used for the training and the test are specifically shown as follows:

Dataset 1:
Training: 'Jan. 15th 0:00-12:00', 'Jan. 16th 0:00-12:00', 'Jan. 17th 0:00-12:00'
Test: 'Jan. 18th 0:00-12:00'
Dataset 2:
Training: 'Jan. 15th 12:00-24:00', 'Jan. 16th 12:00-24:00', 'Jan. 17th 12:00-24:00'
Test: 'Jan. 18th 12:00-24:00'
Dataset 3:
Training: 'Jan. 16th 0:00-12:00', 'Jan. 17th 0:00-12:00', 'Jan. 18th 0:00-12:00'
Test: 'Jan. 19th 0:00-12:00'

Each dataset contains 52,300 areas, and the error degree estimation results for test data are obtained from those 52,300 areas. In addition, for example, the details of the training data and the test data in Dataset 1 are the following combinations (pairs of the features and the labels).
Training data ($52,300 \times 3 (= M)$ samples)
('Features extracted from 3:00, 6:00, and 9:00 on Jan. 15th', 'known error degrees at 12:00 on Jan. 15th')
('Features extracted from 3:00, 6:00, and 9:00 on Jan. 16th', 'known error degrees at 12:00 on Jan. 16th')
('Features extracted from 3:00, 6:00, and 9:00 on Jan. 17th', 'known error degrees at 12:00 on Jan. 17th')
Test data (52,300 samples)
('Features extracted from 3:00, 6:00, and 9:00 on Jan. 18th', 'unknown error degrees at 12:00 on Jan. 18th')

Furthermore, the target meteorological elements F_0, whose error degrees are estimated, and other elements used to calculate features for the estimation are shown in Table 1, where multiple isopiestic surfaces represent those of 1,000 hPa, 950 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa, and 300 hPa. As shown in Table 2, we respectively set seven ranks corresponding to the prediction error degrees for each target element in Table 1 in this experiment.

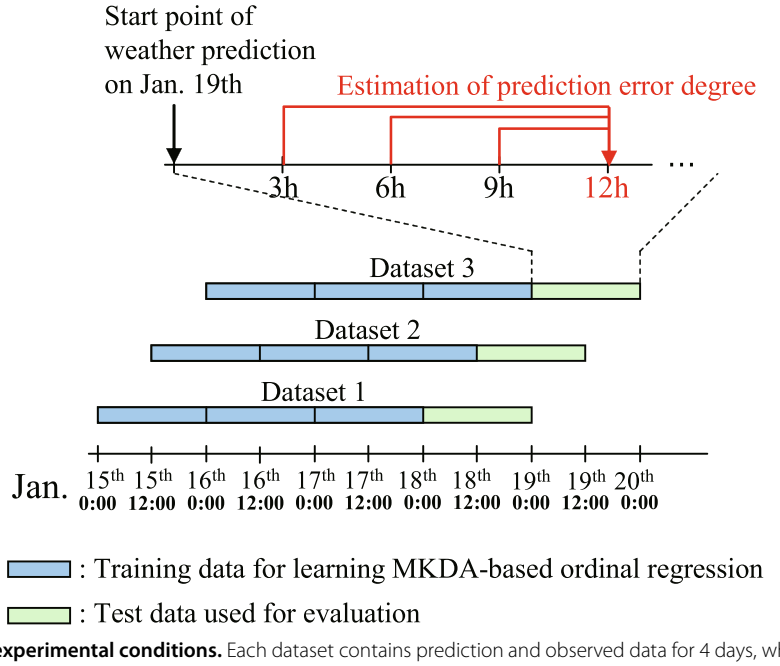

Figure 4 Overview of the experimental conditions. Each dataset contains prediction and observed data for 4 days, where the first 3 days and the the remaining 1 day respectively correspond to the training and the test. Data at 3, 6, and 9 h after the beginning of the forecast were used for error degree estimation at 12 h after the beginning of the forecast.

Table 1 Meteorological elements of estimation targets and feature extraction

Target elements	Other elements used for feature extraction
Temperature on the ground	'Temperature on the multiple isopiestic surfaces' and 'relative humidity and wind velocity on the ground and multiple isopiestic surfaces'
Relative humidity on the ground	'relative humidity on the multiple isopiestic surfaces' and 'temperature and wind velocity on the ground and multiple isopiestic surfaces'
Wind velocity on the ground	'Wind velocity on the multiple isopiestic surfaces' and 'temperature and relative humidity on the ground and multiple isopiestic surfaces'

We simply set K to 7, and there is no specific reason. This value should be determined to the optimal value for target meteorologists who monitor meteorological elements. Then, by using the features calculated from the previously known prediction errors caused in the target element F_0 and some other elements (corresponding to F_1 and F_2), unknown error degree estimation of each target element shown in the left side of Table 1 is performed.

In this experiment, the number of kernels, i.e., the number of meteorological elements, $L+1$, is 3. Since three kinds of data, 'temperature', 'relative humidity', and 'wind velocity', are used, the number of meteorological elements $L+1$ becomes 3. Note that for each meteorological element, five features shown in Section 2.1 are calculated from eight isopiestic surfaces, i.e., $J=8$, where the eight isopiestic surfaces correspond to those of the ground, 1000 hPa, 950 hPa, 925 hPa, 850 hPa, 700 hPa, 500 hPa and 300 hPa. Thus, 40 ($=5\times 8$) features are input to each kernel corresponding to each meteorological element. Then the number of all features is $d=120$ ($=40\times 3$) in this experiment. Specifically, we performed the estimation by using the proposed method and the following conventional methods: SVOR-IMC [11], SVM-EBC [12], and the simple KDA [13]. Since these conventional methods are benchmarking methods and state-of-the-art methods of ordinal regression, they are suitable for comparison with the proposed method.

First, we show the estimation performance of prediction error degrees by the proposed method and the conventional methods. Figure 5a,b,c,d,e shows the results obtained by estimating the prediction error degrees of 'temperature on the ground' in each area based on the proposed and conventional methods.

Note that the parameters used in the proposed method are the number of sampled training data N and C in Equation 13, the weights of the kernels $\gamma_l (l=0,1,2,\cdots,L)$, and the parameters of the Gaussian kernels. First, N was simply set to $\frac{M}{300}$. Next, we set the value of C to 0.1, but the final results of error degree estimation do not depend on C, the proof of which is shown in Appendix 1. As shown in 2.2.3, the weights of the kernels γ_l $(l=0,1,2,\cdots,L)$ and the parameters of the Gaussian kernels can be automatically determined by the proposed method.

Tables 3, 4, 5 respectively show the results of γ_l automatically obtained by the proposed method, and Table 6 shows the parameters of the Gaussian kernels determined by the proposed method. Note that the parameters of the conventional methods were determined on the basis of the schemes shown in their papers. In Figure 5a,b,c,d,e, black regions represent areas for which data cannot be obtained. For better subjective evaluation, zoomed portions of Figure 5a,b,c,d,e are shown in Figure 5f,g,h,i,j. Furthermore, Figure 5k,l,m,n,o shows the differences between estimated prediction error degrees and true degrees. In these results, the intensity represents the absolute difference, and if the intensity is brighter, the absolute difference is larger. Since Figure 5k does not contain any errors, the whole areas become black (zero). From the obtained results, it can be seen that the proposed method successfully estimates the prediction error degrees without suffering from large misestimation compared to the conventional methods. Figures 6 and 7 also show the results obtained by applying the proposed and conventional methods to the other elements shown in Table 1. From these results, it can be confirmed that the proposed method enables more successful estimation for various kinds of meteorological elements.

Table 2 Prediction error degrees and their error value range

Error degree	Color	Temperature	Relative humidity	Wind velocity
1	Blue	err <-3°C	err <-20%	err <-3 m/s
2	Light blue	-3°C $\le$ err <-2°C	-20% $\le$ err <-10%	-3 m/s $\le$ err <-2 m/s
3	Green	-2°C $\le$ err <-1°C	-10% $\le$ err <-5%	-2 m/s $\le$ err <-1 m/s
4	White	-1°C $\le$ err <1°C	-5% $\le$ err <5%	-1 m/s $\le$ err <1 m/s
5	Yellow	1°C $\le$ err <2°C	5% $\le$ err <10%	1 m/s $\le$ err <2 m/s
6	Orange	2°C $\le$ err <3°C	10% $\le$ err <20%	2 m/s $\le$ err <3 m/s
7	Red	3°C $\le$ err	20% $\le$ err	3 m/s $\le$ err

Temperature, relative humidity, and wind velocity respectively represent the error value of 'temperature on the ground', error value of 'relative humidity on the ground,' and error value of 'wind velocity on the ground'. Furthermore, err means error.

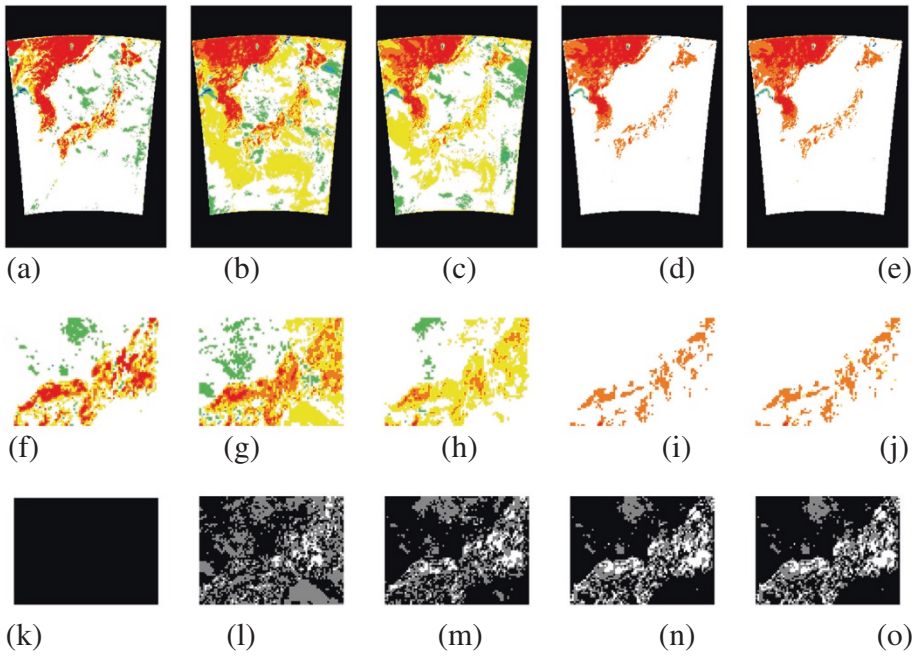

Figure 5 Results of estimated prediction error degrees for the meteorological element 'temperature on the ground'. (a) True prediction error degrees. **(b)** Results of estimation by the proposed method. **(c)** Results of estimation by the simple KDA [13]. **(d)** Results of estimation by SVM-EBC. **(e)** Results of estimation by SVOR-IMC. **(f, g, h, i, j)**. Zoomed portions of **(a, b, c, d, e)**, and **(k, l, m, n, o)** absolute differences between the true prediction error degrees and estimated results in **(f, g, h, i, j)**.

The above experiments were performed on a personal computer using Intel(R) Core(TM) i7 960 CPU 3.20 GHz with 24.0 Gbytes RAM. The proposed method was implemented by using Matlab. The average computation time for the training procedures in our method is about 4.67×10^3 s for Datasets 1 to 3, respectively. The average computation time for the test procedures is about 6.02×10^2 s for Datasets 1 to 3, i.e., 1.15×10^{-2} for each area, respectively. From the obtained results, the computation costs of the training procedures are much larger than those of the test procedures. In the proposed method, we use an exhaustive search for determining the combination parameters of multiple kernels and the parameter of each kernel. Thus, high computation costs are required. However, the exhaustive search can be simply parallelized for each combination of parameters. Therefore, by introducing a parallel search for the parameters, the above computation time for the training procedures can be reduced to about 8.96 s.

Next, we show quantitative evaluation results of the proposed method and the conventional methods. In this experiment, we adopted the following two evaluation metrics:

1. Mean absolute error (MAE): mean of absolute errors between estimated prediction error degrees and their true degrees
2. Mean square error (MSE): mean of square errors between estimated prediction error degrees and their true degrees

Table 3 Estimation results of the kernel weights for 'temperature on the ground'

Meteorological element	Dataset 1	Dataset 2	Dataset 3
γ_0 (temperature)	0.7	0.7	0.9
γ_1 (relative humidity)	0.2	0.2	0.0
γ_2 (wind velocity)	0.1	0.1	0.1

Results of γ_l ($l = 0, 1, 2$) obtained by the proposed method when estimating the prediction error degree of 'temperature on the ground'. γ_0, γ_1, and γ_2 respectively represent the weights of the kernels for temperature, relative humidity, and wind velocity.

Table 4 Estimation results of the kernel weights for 'relative humidity on the ground'

Meteorological element	Dataset 1	Dataset 2	Dataset 3
γ_0 (relative humidity)	0.2	0.7	0.4
γ_1 (temperature)	0.4	0.0	0.5
γ_2 (wind velocity)	0.4	0.3	0.1

Results of γ_l ($l = 0, 1, 2$) obtained by the proposed method when estimating the prediction error degree of 'relative humidity on the ground'. γ_0, γ_1, and γ_2 respectively represent the weights of the kernels for relative humidity, temperature, and wind velocity.

Table 5 Estimation results of the kernel weights for 'wind velocity on the ground'

Meteorological element	Dataset 1	Dataset 2	Dataset 3
γ_0 (wind velocity)	0.5	0.8	0.6
γ_1 (temperature)	0.1	0.1	0.0
γ_2 (relative humidity)	0.4	0.1	0.4

Results of $\boldsymbol{\gamma}_l$ ($l = 0, 1, 2$) obtained by the proposed method when estimating the prediction error degree of 'wind velocity on the ground'. $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_2$ respectively represent the weights of the kernels for wind velocity, temperature, and relative humidity.

We used MAE since the KDA-based ordinal regression in [13] adopted it for the evaluation. Furthermore, in order to evaluate which methods can avoid large misestimation of the prediction error degree, we used MSE in the experiments.

The error degree is defined as shown in Table 2. Therefore, the results of the prediction error degree have values ranging over $1, 2, \cdots, 6, 7$. In the experiments, we calculated MAE and MSE from these values.

Tables 7 and 8 show the results for MAE and MSE, respectively, obtained by applying our method and the conventional methods to the three datasets. Thus, in each meteorological element, the MAE and MSE were calculated from 156,900 areas. Note that in these tables, we also show the results of a method that does not consider the propagation of prediction errors based on atmospheric movements. We simply call this method 'proposed method without propagation (PM-WP)'.

From the obtained results, the proposed method enables more accurate estimation than does the PM-WP. Therefore, it can be confirmed that the calculation of features obtained by using the propagation of prediction errors based on atmospheric movements tends to be effective in our method. Furthermore, by comparing with the KDA-based method [13], we can also see that introduction of the multiple kernel algorithm into the proposed method is effective. From these tables, we can see that the proposed method tends to perform more accurate estimation than do the conventional methods. The proposed method realizes the most accurate performance when using the MSE. Thus, since the MSE becomes lower, it seems that our method can avoid large misestimation of the prediction error degree.

Table 6 Parameters of the Gaussian kernels used in the proposed method

Meteorological element	Dataset 1	Dataset 2	Dataset 3
Temperature	4.5	5.0	4.0
Relative humidity	5.0	5.0	5.0
Wind velocity	5.0	5.0	5.0

We performed a statistical test, Welch's t test, to confirm the difference between the results of the proposed method and the conventional methods in Tables 7 and 8. The results with the significance level set to 0.05 are shown. The details of the statistical test are shown in Appendix 2. In Tables 7 and 8, when the difference between the proposed method and the conventional method was confirmed by this test, the values of the corresponding conventional method are italicized. Therefore, the superiority of our method can be confirmed from this statistical test.

From the above subjective and quantitative evaluations, the effectiveness of the proposed method can be verified. As shown in the discussion of quantitative evaluation, the calculation of features considering atmospheric movements and the use of a multiple kernel scheme are suitable for estimation in the proposed method. Figure 8 shows examples of the projection results of training data and test data for the meteorological element 'temperature on the ground'. It can be seen that training and test samples are almost projected orderly. Although there are overlaps between neighboring ranks, the samples of the test data tend to be projected with maintenance of their ranks. Therefore, these results also reflect the effectiveness of our method.

In order to improve the performance of the proposed ordinal regression approach, the introduction of several state-of-the-art methods would be useful. For example, if a sufficient number of training samples cannot be obtained, the introduction of transductive ordinal regression [24] would be useful for improving the performance of the proposed method. Furthermore, as the number of the meteorological elements becomes larger, the exhaustive search for optimal parameters becomes difficult. In a such case, the introduction of several other approaches for learning kernels such as the approach in [18] may also solve this problem. This point is discussed in detail in the following section.

4 Conclusions

An MKDA-based ordinal regression method for estimating error degrees in numerical weather prediction was presented in this paper. In the proposed method, KDA-based ordinal regression is used for the estimation of prediction error degree, and the following approaches were newly adopted. Since multiple meteorological elements that have propagated from neighboring areas based on atmospheric movements are related to prediction errors in the target area, we calculated those features from observed wind velocities and merged them by using a multiple kernel algorithm. This improved the performance of error degree estimation based on KDA-based ordinal regression. Then, the proposed method enabled successful ordinal regression, i.e., successful estimation of

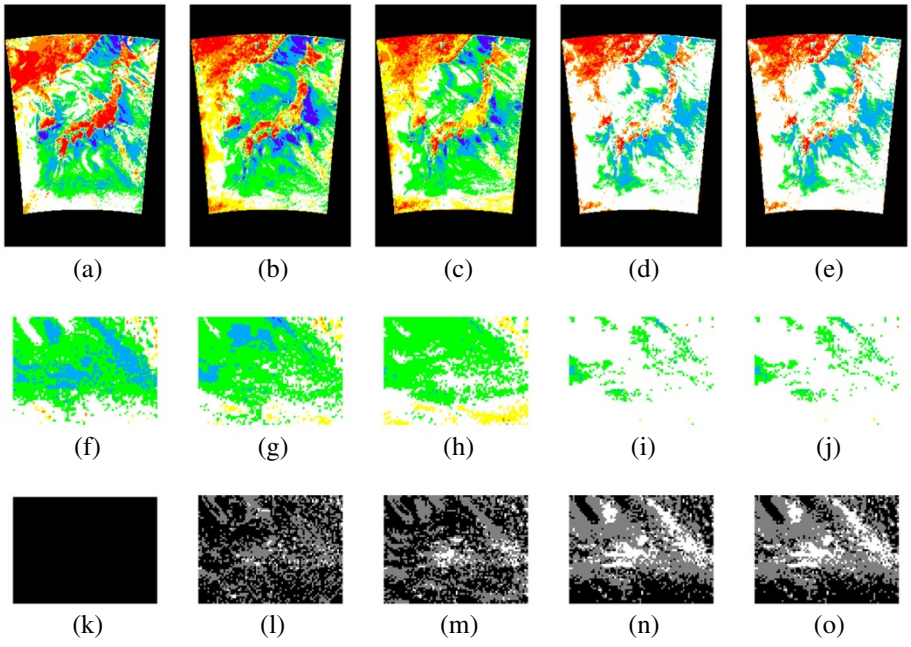

Figure 6 Results of estimated prediction error degrees for the meteorological element 'relative humidity on the ground'. (a) True prediction error degrees. **(b)** Results of estimation by the proposed method. **(c)** Results of estimation by the simple KDA [13]. **(d)** Results of estimation by SVM-EBC. **(e)** Results of estimation by SVOR-IMC. **(f, g, h, i, j)** Zoomed portions of **(a to e)**, and **(k, l, m, n, o)** absolute differences between the true prediction error degrees and estimated results in **(f, g, h, i, j)**.

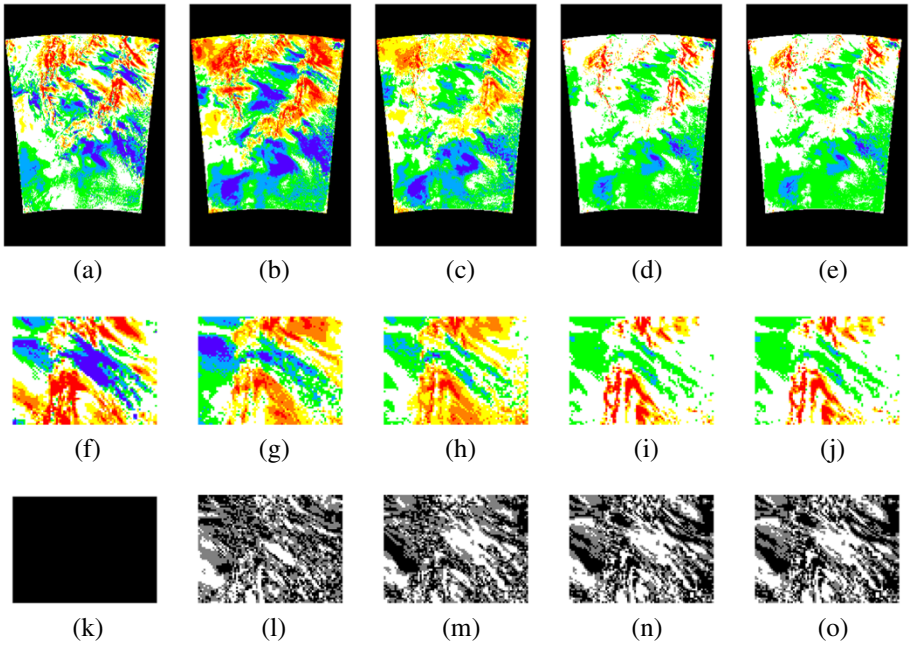

Figure 7 Results of estimated prediction error degrees for the meteorological element 'wind velocity on the ground'. (a) True prediction error degrees. **(b)** Results of estimation by the proposed method. **(c)** Results of estimation by the simple KDA [13]. **(d)** Results of estimation by SVM-EBC. **(e)** Results of estimation by SVOR-IMC. **(f, g, h, i, j)** Zoomed portions of **(a, b, c, d, e)** and **(k, l, m, n, o)** absolute differences between the true prediction error degrees and estimated results in **(f, g, h, i, j)**.

Table 7 MAE-based quantitative evaluation of the proposed method and conventional methods

Target element	Ours	PM-WP	Ref [13]	SVM-EBC	SVOR-IMC
Temperature	0.4317	*0.4458*	*0.3901*[c]	*0.3840*[b]	*0.3839*[a]
Relative humidity	0.7856[a]	*0.8017*[b]	*0.8384*[c]	*0.9030*	*0.9041*
Wind velocity	0.8654	*0.9035*	*0.8456*[a]	0.8645[c]	0.8626[b]

[a]The best method. [b]The second-best method. [c]The third-best method. PM-WP, proposed method without propagation. In addition, temperature, relative humidity, and wind velocity respectively represent temperature on the ground, relative humidity on the ground, and wind velocity on the ground. When the difference between the proposed method and the conventional method was confirmed by Welch's *t* test, the values of the corresponding conventional method are italicized.

prediction error degrees. This was also confirmed from experimental results obtained by applying our method and conventional methods to real data of numerical weather prediction.

In the proposed method, it becomes difficult to adopt the exhaustive search for determining $\boldsymbol{\gamma}$ when the number of meteorological elements becomes larger. Therefore, we will have to solve this problem by using some alternative schemes. First, it would be useful to limit the number of the meteorological elements used for estimating each target meteorological element. This approach is similar to the feature selection, i.e., we select the elements that are the most useful for estimation of the target meteorological element. It should be noted that in the experiments described in this paper, we simply selected the three elements that tend to affect each other from the view point of meteorology. Second, we can use some alternative selection algorithms of the best results of $\boldsymbol{\gamma}$ instead of the exhaustive search. Many optimization methods that can avoid an exhaustive search have been proposed, and a genetic algorithm is one of the most traditional optimization approaches. By using such approaches, we can avoid an exhaustive search and enable the use of more meteorological elements for estimating the target meteorological elements. Some other approaches for learning kernels such as the approach in [18] may also solve this problem. This provides a solution to the problem of the need to tune optimal kernel parameters. This method can be easily extended to a supervised scenario, and it has also been reported that it can outperform the conventional multiple kernel learning approaches. Therefore, the introduction of this method will improve prediction performance.

Table 8 MSE-based quantitative evaluation of the proposed method and conventional methods

Target element	Ours	PM-WP	Ref [13]	SVM-EBC	SVOR-IMC
Temperature	0.6748[a]	0.6810[b]	*0.7143*[c]	*0.7909*	*0.7903*
Relative humidity	1.1764[a]	*1.2392*[b]	*1.4394*[c]	*1.7269*	*1.7276*
Wind velocity	1.4135[a]	*1.5550*[c]	*1.5093*[b]	*1.6614*	*1.6543*

[a]The best method. [b]The second-best method. [c]The third-best method. PM-WP, proposed method without propagation. In addition, temperature, relative humidity, and wind velocity respectively represent temperature on the ground, relative humidity on the ground, and wind velocity on the ground. When the difference between the proposed method and the conventional method was confirmed by Welch's *t* test, the values of the corresponding conventional method are italicized.

Note that as shown in [28], it was reported that KDA-based ordinal regression could not outperform a straightforward approach such as that in [29]. Therefore, we should examine the relationship between the performance of methods and applied datasets from simple toy data toward a large amount of real data since we confirmed the performance of our method by using only data obtained in numerical weather prediction.

The above points should be examined in future work.

Appendices

Appendix 1

In this appendix, we show the proof that the final results of the error degree estimation do not depend on C. Specifically, given

$$\begin{aligned}\min J_2(\boldsymbol{\alpha}) = &\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)'\mathbf{T}^{-1}\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)\\ &\text{s.t.}\quad \alpha^k\geq 0,\quad k=1,2,\cdots,K-1\\ &\sum_{k=1}^{K-1}\alpha^k = C,\end{aligned}\tag{28}$$

we denote its optimal solution as $\boldsymbol{\alpha}^*$. Furthermore, by multiplying C by a positive constant λ, the following problem can be also obtained:

$$\begin{aligned}\min J_2(\boldsymbol{\alpha}_\lambda) = &\sum_{k=1}^{K-1}\alpha_\lambda^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)'\mathbf{T}^{-1}\sum_{k=1}^{K-1}\alpha_\lambda^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right)\\ &\text{s.t.}\quad \alpha_\lambda^k\geq 0,\quad k=1,2,\cdots,K-1\\ &\sum_{k=1}^{K-1}\alpha_\lambda^k = \lambda C.\end{aligned}\tag{29}$$

Then, the optimal solution of $\boldsymbol{\alpha}_\lambda = \left[\alpha_\lambda^1,\alpha_\lambda^2,\cdots,\alpha_\lambda^{K-1}\right]'$ satisfying $\sum_{k=1}^{K-1}\alpha_\lambda^k = \lambda C$ can be obtained. Next, from the above equations, $J_2(\boldsymbol{\alpha}_\lambda) = J_2(\lambda\boldsymbol{\alpha}) = \lambda^2 J_2(\boldsymbol{\alpha})$. Then, the optimal solution of Equation 29 becomes $\lambda\boldsymbol{\alpha}^*$ since $J_2(\boldsymbol{\alpha})$ becomes minimum when $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$. Therefore, if we mul-

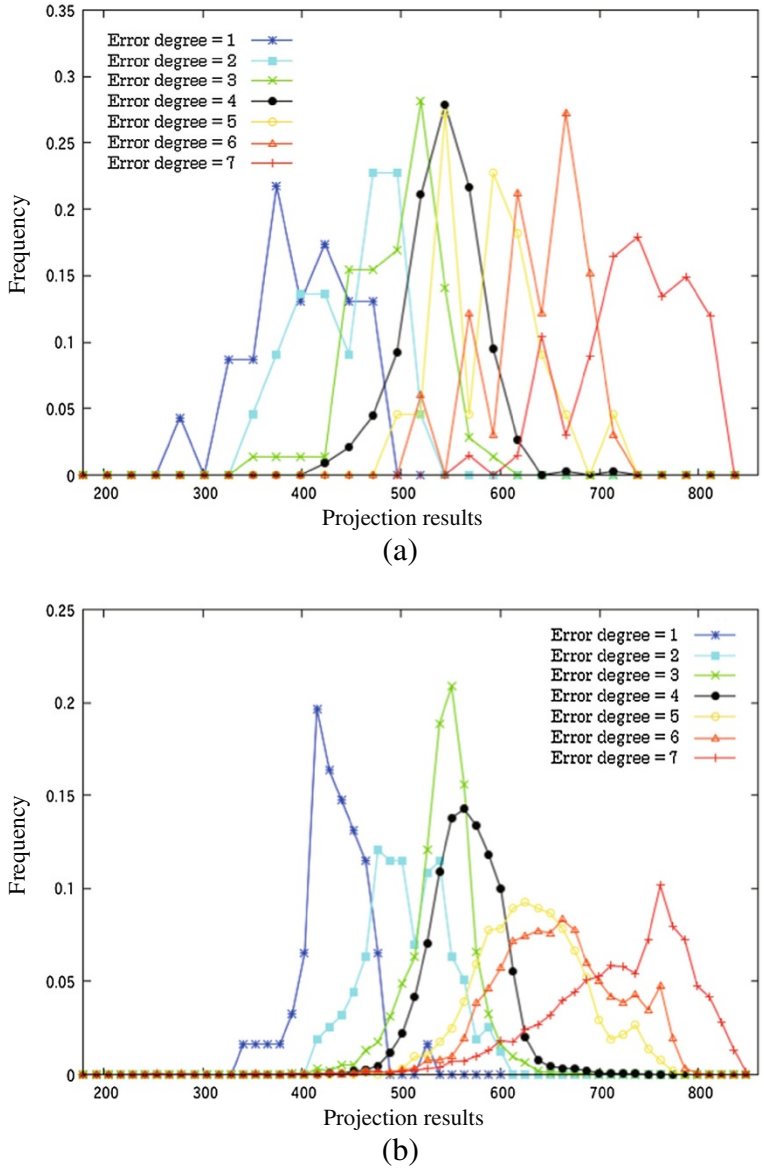

Figure 8 Projection results of training and test data. These results were obtained by applying the proposed method to the data of 'temperature on the ground'. **(a)** Histogram of projection results for the training data. **(b)** Histogram of projection results for the test data. Since the number of the test samples is much larger than that of the training samples, the graph of the test data becomes smoother than that of the training data.

tiply C by λ, the optimal solution also becomes λ times larger. Consequently, since

$$\boldsymbol{\beta} = \frac{1}{2}\mathbf{T}^{-1}\sum_{k=1}^{K-1}\alpha^k\left(\mathbf{r}^{k+1}-\mathbf{r}^k\right), \tag{30}$$

$\boldsymbol{\beta}$ also becomes λ-times larger, and

$$f(\mathbf{x}) = \min_{k\in\{1,2,\cdots,K\}}\left\{k : \boldsymbol{\beta}'\left(\boldsymbol{\Xi}'\phi(\mathbf{x}) - \frac{\mathbf{r}^{k+1}+\mathbf{r}^k}{2}\right) < 0\right\} \tag{31}$$

does not depend on λ. From the above explanation, we can see that the final results of the rank estimation do not depend on C.

Appendix 2

We show the details of the statistical test shown in the experiment section. First, we define the following evaluation value:

$$E = \begin{cases} |y-y_g| & \text{if MAE is used} \\ (y-y_g)^2 & \text{if MSE is used} \end{cases}, \tag{32}$$

where y is an estimated error degree by a method, and y_g is a correct error degree (ground truth). Given two methods, A and B, t is defined by

$$t = \frac{\bar{E}_A - \bar{E}_B}{\sqrt{\frac{S_A^2}{N_A} + \frac{S_B^2}{N_B}}}, \tag{33}$$

where $\bar{E}_A$ and $\bar{E}_B$ are the averages of E in Equation 32 for methods A and B, and thus, they correspond to MAE or MSE. Furthermore, S_A^2 and S_B^2 are the variances of E in methods A and B, respectively. The number of data ($N_A = N_B$) is the same value 156,900 ($= 52,300 \times 3$). The degree of freedom ν is calculated as

$$\nu = \frac{\left(\frac{S_A^2}{N_A} + \frac{S_B^2}{N_B}\right)^2}{\frac{S_A^4}{N_A^2(N_A-1)} + \frac{S_B^4}{N_B^2(N_B-1)}}. \tag{34}$$

We assume the null hypothesis 'For the evaluation values E obtained by the two methods A and B, the two population means, i.e., $\bar{E}_A$ and $\bar{E}_B$, are equal', and t follows a t-distribution of the degree of freedom ν. When significance level α is provided, we calculate a threshold t_0 satisfying $\Pr\{|t| \geq t_0\} = \alpha$, and if $|t| \geq t_0$, the null hypothesis is rejected. In this experiment, α is set to 0.05. In this test, we regard the proposed method and one of the conventional methods as methods A and B, respectively. Note that in this experiment, the value of the threshold t_0 becomes 1.96 for all cases.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was partly supported by Grant-in-Aid for Scientific Research (B) 25280036. In our experiments, we used meteorological data provided by the Japan Weather Association.

Received: 23 July 2013 Accepted: 11 July 2014
Published: 22 July 2014

References

1. TN Krishnamurti, Numerical weather prediction. Annu. Rev. Fluid Mech. **27**, 195–224 (1995)
2. R Kimura, Numerical weather prediction. J. Wind Eng. Ind. Aerodynamics **90**(12–15), 1403–1414 (2002)
3. E Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, 2003)
4. HL Shang, RJ Hyndman, Nonparametric time series forecasting with dynamic updating. Math. Comput. Simul. **81**(7), 1310–1324 (2011)
5. P Louka, G Galanis, N Siebert, G Kariniotakis, P Katsafados, I Pytharoulis, G Kallos, Improvements in wind speed forecasts for wind power prediction purposes using kalman filtering. J. Wind Eng. Ind. Aerodynamics **96**(12), 2348–2362 (2008)
6. MDL Alysha, JH Rob, DS Ralph, Seasonal patterns using exponential smoothing. J. Am. Stat. Assoc. **106**(496), 1513–1527 (2011)
7. PG Gould, AB Koehler, JK Ord, RD Snyder, RJ Hyndman, Vahid-F Araghi, Forecasting time series with multiple seasonal patterns. Eur. J. Oper. Res. **191**(1), 207–222 (2008)
8. C Cortes, V Vapnik, Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
9. VJ Hodge, J Austin, A survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85–126 (2004)
10. MI Petrovskiy, Outlier detection algorithms in data mining systems. Program. Comput. Softw. **29**(4), 228–237 (2003)
11. W Chu, SS Keerthi, New approaches to support vector ordinal regression, in *Proc. of the 22nd International Conf. on Machine Learning* (ACM, New York, 2005), pp. 145–152
12. L Lin, H-T Lin, Ordinal regression by extended binary classification. Adv. Neural Inform. Proces. Syst. **19**, 865–872 (2007)
13. BY Sun, J Li, DD Wu, XM Zhang, WB Li, Kernel discriminant learning for ordinal regression. IEEE Trans. Knowl. Data Eng. **22**(6), 906–910 (2010)
14. CM Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006)
15. S Mika, G Ratsch, J Weston, B Schölkopf, K-R Müller, Fischer discriminant analysis with kernels, in *Proceedings of the IXth international workshop on Neural Networks for Signal Processing* (IEEE, New York, 1999), pp. 41–48
16. G Baudat, F Anouar, Generalized discriminant analysis using a kernel approach. Neural Comput. **12**(10), 2385–2404 (2000)
17. V Roth, V Steinhage, Nonlinear discriminant analysis using kernel functions. Adv. Neural Inform. Proces. Syst. **12**, 568–574 (2000)
18. B Liu, S-X Xia, Y Zhou, Unsupervised non-parametric kernel learning algorithm. Knowl. Based Syst. **44**, 1–9 (2013). doi:10.1016/j.knosys.2012.12.008
19. M Gönen, E Alpaydn, Multiple kernel learning algorithms. J. Mach. Learn. Res. **12**, 2211–2268 (2011)
20. M Kloft, U Rückert, PL Bartlett, A unifying view of multiple kernel learning, in *Proceedings of the European Conference on Machine Learning (ECML)* (Springer, Berlin, 2010)
21. A Rakotomamonjy, FR Bach, S Canu, Y Grandvalet, Simplemkl. J. Mach. Learn. Res. **9**, 2491–2521 (2008)
22. PK Srijith, S Shevade, S Sundararajan, Validation based sparse gaussian processes for ordinal regression. **7664**, 409–416 (2012). doi:10.1007/978-3-642-34481-7_50
23. F Fernandez-Navarro, PA Gutierrez, C Hervas-Martinez, X Yao, Negative correlation ensemble learning for ordinal regression. Neural Netw. Learn. Syst. IEEE Trans. **24**(11), 1836–1849 (2013). doi:10.1109/TNNLS.2013.2268279
24. C-W Seah, IW Tsang, Y-S Ong, Transductive ordinal regression. Neural Netw. Learn. Syst. IEEE Trans. **23**(7), 1074–1086 (2012). doi:10.1109/TNNLS.2012.2198240
25. Y Liu, Y Liu, KCC Chan, J Zhang, Neighborhood preserving ordinal regression, 119–122 (2012). doi:10.1145/2382336.2382370
26. B Schölkopf, S Mika, CJC Burges, P Knirsch, K-R Müller, G Rätsch, AJ Smola, Input space versus feature space in kernel-based methods. IEEE Trans. Neural Netw. **10**(5), 1000–1017 (1999)
27. J Nocedal, SJ Wright, *Numerical Optimization, 2nd edition* (Springer, New York, 2006)
28. JS Cardoso, R Sousa, I Domingues, Ordinal data classification using kernel discriminant analysis: a comparison of three approaches, in *ICMLA, vol. 1* (IEEE, Boca Raton, 2012), pp. 473–477
29. E Frank, M Hall, A simple approach to ordinal classification, in *Proceedings of the 12th European Conference on Machine Learning.* EMCL '01 (Springer, London, 2001), pp. 145–156

doi:10.1186/1687-6180-2014-115
Cite this article as: Ogawa *et al.*: **A new method for error degree estimation in numerical weather prediction via MKDA-based ordinal regression.** *EURASIP Journal on Advances in Signal Processing* 2014 **2014**:115.