Linear and nonlinear regression prediction of surface wind components

Yiwen Mao & Adam H. Monahan

January 2018

Climate Dynamics (2018) 51:3291–3309
https://doi.org/10.1007/s00382-018-4079-5

Linear and nonlinear regression prediction of surface wind components

Yiwen Mao[1] · **Adam Monahan**[1]

Received: 3 July 2017 / Accepted: 8 January 2018 / Published online: 19 January 2018

Abstract
This study compares the statistical predictability by linear regression of surface wind components using mid-tropospheric predictors with predictability by three nonlinear regression methods: neural networks, support vector machines and random forests. The results, obtained at 2109 land stations, show that more complex nonlinear regression methods cannot substantially outperform linear regression in cross-validated statistical prediction of surface wind components. As well, predictive anisotropy (variations in statistical predictive skill in different directions) are generally similar for both linear and nonlinear regression methods. However, there is a modest trend of systematic improvement in nonlinear predictability for surface wind components with fluctuations of relatively small magnitude or large kurtosis, which suggests weak nonlinear predictive signals may exist in this situation. Although nonlinear predictability tends to be higher for stations with low linear predictability and nonlinear predictive anisotropy tends to be weaker for stations with strong linear predictive anisotropy, these differences are not substantial in most cases. Overall, we find little justification for the use of complex nonlinear regression methods in statistical prediction of surface wind components as linear regression is much less computationally expensive and results in predictions of comparable skill.

Keywords Statistical prediction · Linear regression · nonlinear regression · Predictability of surface winds

1 Introduction

Surface winds are a climatic field of interest because of the broad range of societal and economic sectors they affect, including agriculture, transport, and energy systems (Stull 2000). However, the modelling skill of surface winds using global climate models (GCMs) is limited due to coarse horizontal and vertical resolution and difficulties simulating other boundary layer processes (Holtslag et al. 2013; He et al. 2010). Driving finer-resolution dynamical models by GCM output is one approach to the simulation of surface winds. This approach has the advantage of modeling surface winds based on physics, but the drawback is that dynamical models are computationally expensive and also subject to resolution- and parameterization-dependent biases. An alternative approach is through computationally cheaper statistical prediction using well-resolved, large-scale predictors. Statistical prediction in this context refers to the prediction based on the relationship of atmospheric fields at the same time but different locations, rather than using information about the state of the atmosphere at one time to estimate the state at a later time.

Given the importance of surface winds, it is beneficial to assess the predictive skill of surface winds by statistical prediction, which is a typical application of supervised learning (Hsieh 2009). Specifically, statistical prediction requires a transfer function (TF) derived from the statistical relationship between predictors (e.g. large-scale climate fields in the free atmosphere) and predictands (local-scale surface winds) based on historical data, such that the TF can be then applied for new prediction. The skill of statistical prediction depends on how well the TF can model the relationship between predictors and predictands. Therefore, the predictability resulting from statistical predictions is strongly influenced by the characteristics of the predictor-predictand relationship, such as the functional form (linear vs nonlinear) of the predictor-predictand relationship, and the signal to noise ratio (SNR). Statistical prediction can be classified into linear or nonlinear approaches according to whether the TF used to model

✉ Yiwen Mao
ymaopanda@gmail.com

[1] School of Earth and Ocean Sciences, University of Victoria, Victoria, BC, Canada

the predictor-predictand relationship is implemented by linear or nonlinear regression methods.

This study focuses on statistical prediction of surface wind components because the direction of wind is often important in applications related to surface winds as wind is a vector quantity. For example, calculation of the transport of airborne substances requires knowledge of the vector wind. Previous studies have shown that predictability by linear TF of surface wind components projected onto different compass directions is often characterized by anisotropy, such that different projections are predicted with different levels of skill and can often achieve better predictability than predicting surface wind speed (Salameh et al. 2009; van der Kamp et al. 2012; Culver and Monahan 2013; Sun and Monahan 2013; Mao and Monahan 2017). The best predicted components may not be the conventional zonal or meridional, so to investigate predictive anisotropy, we need to consider predictability of surface wind components projected onto directions around the compass.

Mao and Monahan (2017) argued that predictive anisotropy by linear regression (LR) becomes strong when the overall SNR of predictor-predictand relationship (across all directions) is small, and directional predictability approaches isotropy when the overall SNR becomes large. The argument was based on a simple descriptive model partitioning the surface winds into two parts, perfectly correlated and uncorrelated with the large-scale atmospheric flow respectively. In this context, the "signal" was defined in terms of a linear statistical relationship.

A factor other than noise which can also lead to weak correlation between surface winds and large-scale flow (and therefore poor linear statistical predictability) is a nonlinear statistical relationship between predictands and predictors. If the predictive signal of the predictor-predictand relationship is nonlinear, the overall predictability by linear TF should be lower than using nonlinear TF. Mao and Monahan (2017) found that the directions of low linear predictability are often aligned with wind components characterized by large kurtosis. Linear predictive models are optimal when the joint distribution of the predictor and predictand are Gaussian. Therefore, it is possible that the anisotropy of linear-regression based prediction of surface wind components may result from variation of linearity of the predictor-predictand relationship in different directions. There is no a priori reason to expect that this relationship is linear, as atmospheric dynamics are themselves nonlinear.

While a nonlinear TF can result in higher predictability than linear TF because it can model a broader class of functional relationships, in practice, the cross-validated predictability by nonlinear TF may be lower than that of linear TF when the predictor-predictand relationship contains a large amount of noise. More complex algorithms are more likely to fit the noise as well as the predictive signal of the predictor-predictand relationships, which leads to the problem of overfitting. The problem of overfitting becomes more pronounced as the number of observations used in the statistical analysis is reduced.

This study aims at clarifying whether predictability of surface wind components resulting from linear TF is improved by allowing for nonlinear functional relationships, and this study focuses on prediction of conditional expectation of surface wind components given the predictors. We also assess if the predictive anisotropy found in previous studies (van der Kamp et al. 2012; Culver and Monahan 2013; Sun and Monahan 2013; Mao and Monahan 2017) is an artifact of the use of a linear-regression based TF. In other words, If the predictability resulting from nonlinear regression methods is substantially improved over all directions of projections of surface winds, we would see significantly weakened predictive anisotropy or even isotropic predictability from nonlinear regression based prediction. Specifically, we compare the characteristics of predictability (i.e. magnitude and anisotropy) of surface wind components using nonlinear TF and linear TF for 2109 land stations across a wide range of locations across the globe (the same set of stations considered in Mao and Monahan 2017).

We consider three common nonlinear machine learning methods as TFs for statistical prediction of surface wind components: neural network (NN), support vector machine (SVM) and random forest (RF), and compare the results with the statistical prediction by linear regression (LR). These three nonlinear methods have been applied to prediction of surface wind speed in a few previous studies. For example, Sailor et al. (2000) developed a methodology based on NN for downscaling GCM output to predict surface wind speed. Mohandes et al. (2004) applied SVM to predict daily averaged wind speed from Madina, Saudi Arabia, and compared the performance of SVM with NN. Their results indicate that SVM outperforms NN at this site. Davy et al. (2010) applied RF based statistical prediction to model wind variability at a coastal location in Victoria, Australia. They found that RF based statistical prediction outperforms linear regression, and the overall accuracy of RF was competitive with NN. These previous studies on statistical prediction only use a small number of stations in limited geographic regions, and the predictand in most of these studies is wind speed. The present study undertakes a systematic comparison of statistical prediction by linear and the three nonlinear methods (NN, SVM and RF) using data from meteorological stations at a wide range of locations across the world in order to get a more comprehensive picture of the statistical prediction of surface wind components.

This paper is organized as follows. Section 2 presents the data and methods used in this study. Section 3 presents the results of comparing predictability of surface wind components using statistical predictions with different TFs.

Section 4 presents a discussion comparing the skill of linear and nonlinear regression methods. The conclusions are given in Sect. 5. Brief descriptions of nonlinear regression methods are presented in Appendix A.

2 Data and methods

2.1 Predictor and predictand data

Mao and Monahan (2017) considered the statistical prediction of observed surface wind components using linear regression based TFs at a network of 2109 land stations. While these stations are distributed across the globe, most of them are concentrated in middle latitudes of the Northern Hemisphere (Fig. 1). In this study, we will consider nonlinear statistical predictions at the same stations.

2.1.1 Predictand data

The predictands (the surface wind components) at each station are derived from observational data of hourly wind speed (w) and direction (ϕ) at 10 m above the ground during a 2-min period ending at the beginning of the hour. Direction is where the wind comes from, measured clockwise from north. The time period of the data is from 1980/01/01 to 2012/12/31. These data were obtained using the WeatherData function of Mathematica 9.0 (Wolfram 2016) which includes a wide range of data sources. Chief among these sources are the National Weather Service of the National Oceanic and Atmospheric Administration (NOAA), the Unites States National Climatic Data Center, and the Citizen Weather Observer program. Only stations with fewer than 10% missing data for the period under consideration are considered. Specifically, zonal and meridional winds are derived from the original hourly wind speed and direction data as following:

$$u = -w\sin(\phi), \tag{1}$$

$$v = -w\cos(\phi). \tag{2}$$

The wind component projected onto direction θ is then calculated as

$$U(\theta) = u\sin(\theta) + v\cos(\theta), \tag{3}$$

with θ varying from 0° to 170°. In total, there are 36 surface wind components as predictands at each station. Only 18 of these are distinct, because $U(\theta) = -U(\theta + 180°)$. Averaging hourly $U(\theta)$ with daily and monthly frequency is used as daily averaged and monthly averaged surface wind components. The reason we consider 36 surface wind components projected onto compass directions is because the the best or worst predicted wind component is not always the conventional zonal or meridional component; knowledge of the predictability of u and v alone is not sufficient in general to assess the predictive anisotropy. Moreover, predictability of wind components in this context refers to projections of wind vectors onto a coordinate axis rather than wind coming from a specified direction. Prediction of the latter assumes the direction of wind is known but not the speed, which is not practical in reality.

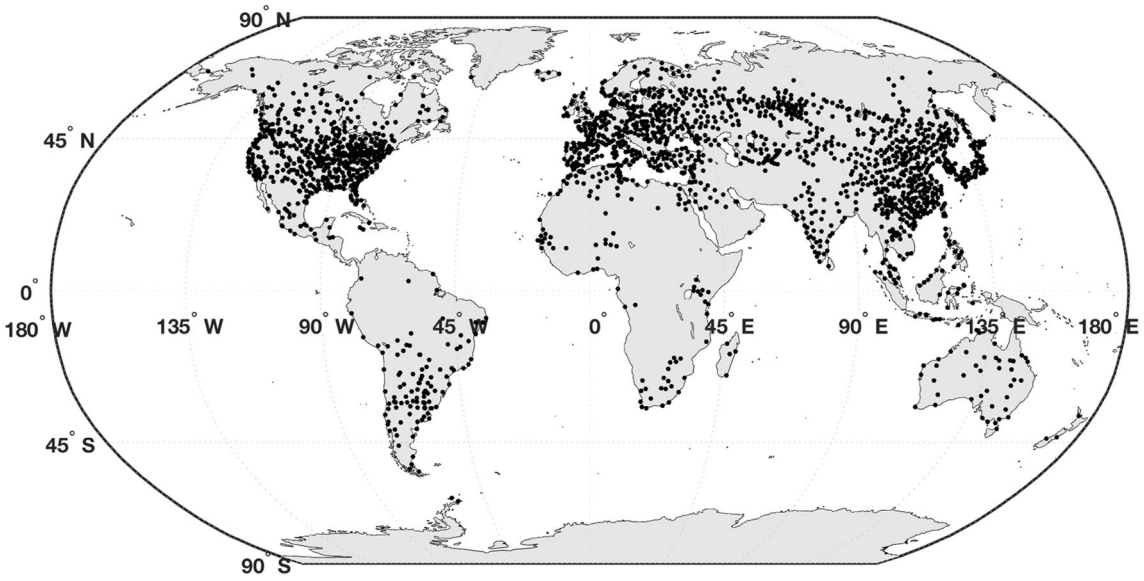

Fig. 1 The locations of the 2109 land stations used for statistical prediction of surface winds

2.1.2 Predictor data

The four mid-tropospheric meteorological fields of temperature T, geopotential height Z, zonal wind U, and meridional wind V at 500 hPa are chosen as the predictors for this study. They are obtained from NCEP-Reanalysis 2 data (Kanamitsu et al. 2002) provided by the from http://www.esrl.noaa.gov/psd/. Previous studies (Culver and Monahan (2013); Monahan (2012) have shown that the correlation structures between surface wind vectors and large-scale free tropospheric climate variables are often spread across a large area surrounding a station, such that the locations of highest predictability aloft are often not directly above the surface station. Based on the approximate size of the region of largest predictability in these previous, mid-tropospheric predictors at a given station are selected from a domain of fixed size centered at the station. In order to assess the influence of domain size on the statistical predictability of surface wind components, we compared the predictions from a $40° \times 40°$ domain with that from a smaller domain $22.5° \times 22.5°$ for a random sample of 100 stations, using all four regression methods considered (LR, NN, SVM and RF). For all regression models, the difference between the two domains is negligible for daily averaged prediction but is sometimes substantial for monthly averaged prediction. The difference is reasonable since the spatial scale of correlation between surface fields and atmospheric flow aloft is larger for longer averaging time scales. For daily averages, the correlation structures are on the synoptic scale, while for monthly averages they have the larger scale of atmospheric low-frequency variability. Based on these test results, predictors of statistical predictions are derived from the reanalysis fields within a $22.5° \times 22.5°$ grid box centered at the location of each weather station for daily averaged prediction, and a larger $40° \times 40°$ grid box is used to derive predictors for monthly averaged prediction. Since the resolution at which tropospheric variables are available from the NCEP II reanalysis is $2.5° \times 2.5°$, the prediction domain consists of 81 grid points (i.e. 9 grid points on each side) for daily averaged predictions and 256 grid points (i.e. 16 grid points on each side) for monthly averaged predictions. For each grid point the location of which is labeled as (i,j) in the domain, time series of daily and monthly averaged $(U_{ij}, V_{ij}, T_{ij}, Z_{ij})$ with a sampling frequency of 6 hours at 500 hPa are used to predict $U(\theta)$ at the station. the predictor and predictand data at the same time are considered: the statistical model considers simultaneous relationships between near-surface wind and free-tropospheric flow. Consideration of lag relationships between the predictors (not shown) indicate that the strongest statistical relationships occur with zero lag.

2.2 Measure of predictability

To minimize the influence of seasonality, the mean seasonal cycles of predictands and predictors are all subtracted before the statistical prediction models are constructed. The seasonal cycles are obtained using the harmonic fit:

$$\begin{aligned} Y_s(t) = B_0 + B_1\cos(\omega t) + B_2\sin(\omega t) + B_3\cos(2\omega t) \\ + B_4\sin(2\omega t) + B_5\cos(3\omega t) + B_6\sin(3\omega t), \end{aligned} \tag{4}$$

where $\omega = 2\pi/P$ with $P = 365$ days for daily averaged time series (after removing data of Feb 29th for convenience), and $P = 12$ months for monthly averaged time series. Including a larger number of harmonics in the seasonal cycle has essentially no effect on the resulting regression models (not shown). In order to minimize the influence of any remaining seasonality on the statistical relationship, statistical models are fit separately for the winter and summer seasons. Winter is defined as December, January, and February for Northern Hemisphere stations and June, July, August for Southern Hemisphere stations. The assignment of these months is reversed in summer. The deseasonalized time series of predictors are then scaled by their individual standard deviations in order to obtain standardized predictors.

Estimates of statistical predictability in this study are obtained using the approach of Mao and Monahan (2017). Predictions of observed surface wind components for the period 1980-2012 at each station are obtained using the modelled statistical relationship between deseasonalized predictands and predictors at each grid point (i, j) using both linear and nonlinear regression methods. Statistical predictability is assessed using leave-one-year-out cross-validation. For all four models (LR, NN, SVM, and RF), for each year of observations the prediction model is estimated using data from the other 32 years. The process is repeated 33 times to obtain time series of predicted surface wind components.

For each regression, the resulting predictability at each grid point (i,j) is measured by the squared correlation R^2_{ij},

$$R^2_{ij}(\theta) = |corr(U(\theta), \widehat{U}_{ij}(\theta))|^2 \tag{5}$$

where $\widehat{U}_{ij}(\theta)$ is the predicted time series using the four predictors (U,V,T and Z) at the grid point (i,j). A single measure of predictability across the domain is then computed for each method, denoted Π:

$$\Pi(\theta) = < R^2_{ij}(\theta) > . \tag{6}$$

At each grid point the quantity R^2_{ij} represents the predictive information carried by the predictors at the point. As we desire a measure of predictability $\Pi(\theta)$ that reflects the strongest predictive information within the domain, rather than a domain averaged predictability (which will generally include regions with very low predictive skill), the average

calculated by Eq. (6) is taken over the two (for daily prediction) and four (for monthly prediction) grid points with the top values of $R^2_{ij}(\theta)$ within the prediction domain (corresponding to 2% of the grid points in the domain). In general, estimates of predictability are not strongly sensitive to variations between 2 and 5% grids points with top $R^2_{ij}(\theta)$. This approach is used in order to have a transparent and systematic way of estimating predictability that can be applied to all stations since there is no general relationship between the location of the best predictors aloft and the location the surface stations.

For linear regression, R^2 is the natural measure of goodness of fit as it measures the proportion of total variation explained by the linear model. While this may not be exactly true for nonlinear methods if the prediction and the residual are correlated, in practice such correlations are found to be small and the R^2 between prediction and predictand remains a useful measure of fit, with the particular benefit of allowing direct comparison with linear model. In computing Eq. (6), only grid points for which $corr(U(\theta), \widehat{U}_{ij}(\theta)) > 0$ are considered. Occasionally, the cross-validated correlation between the predicted and observed wind components can become negative. Negative correlation between prediction and predictand may result from large sampling fluctuations and low intrinsic predictability (non cross-validated R^2 values are small in these cases), and negative correlation indicates that there is no predictive information carried by the predictors at the grid point. Since $\Pi(\theta)$ reflects largest potential predictive information within the domain, we exclude grid points with negative correlation from the analysis.

We also tested whether Lasso regression would be more efficient as it can automatically select predictors within the domain. To this end, Lasso regression is performed all 2109 stations in the direction of $max(R^2)$ obtained using the top 2% selection method. The number of predictors selected by Lasso is generally large, around 100 predictors for most stations for both daily and monthly predictions. The selected predictors are likely to be highly correlated with each other, which could be a source of inflated predictability. Although the results show evidence that linear predictability by Lasso regression has a somewhat larger R^2, there is no guarantee that the automatically selected predictors are truly relevant factors in explaining variability of surface wind components since Lasso may not distinguish true predictors from irrelevant variables but highly correlated with predictors with any amount of data and any amount of regularization (Zhao and Yu 2006). Moreover, fitting a nonlinear model with such a large number of predictors is even more prone to overfitting. Therefore, we focus on the top 2% method described above for each regression method in this study to represent linear and nonlinear predictability.

The directional predictability values obtained from LR, NN, SVM, and RF are denoted $\Pi_{LR}(\theta)$, $\Pi_{NN}(\theta)$, $\Pi_{SVM}(\theta)$ and $\Pi_{RF}(\theta)$ respectively. The overall predictability at a station is measured by the mean of Π over all directions, denoted $\overline{\Pi(\theta)}$, and predictive anisotropy is measured by

$$\alpha(\Pi) = \frac{min(\Pi)}{max(\Pi)}, \tag{7}$$

where $min(\Pi)$ and $max(\Pi)$ are respectively the minimum and maximum $\Pi(\theta)$ over the 36 values of θ. Values of $\alpha(\Pi)$ range between 0 to 1, such that lower $\alpha(\Pi)$ indicates a stronger degree of anisotropy. As with $\Pi(\theta)$, anisotropy values for different statistical models are indicated by subscripts.

2.3 Implementation of nonlinear models

A detailed discussion of the nonlinear regression methods considered in this study is presented in Appendix A. Various software tools exist to implement nonlinear regression analysis. The tools we used were selected on the basis of robustness of results and flexibility of implementation. In this regard, MATLAB function 'fitrsvm' in the Statistics and Machine Learning Toolbox (MathWorks 2017a) and the Python function 'RandomForestRegressor' in the Scikit-learn library of Python (Python 2016) are used for implementing SVM and RF in this study. Since we choose different training procedures for daily-averaged and monthly-averaged data according to their different data properties, the R package 'nnet' (Ripley and Venables 2016) and the MATLAB Neural Network toolbox (MathWorks 2017b) are used to implement NN for daily- and monthly-averaged data. The details will be explained later in this section.

In all regression analyses, we relate four predictor variables (X) to one predictand (Y). Each nonlinear regression method requires the specification of parameters related to model architecture. The challenge is to determine parameters which yield robust regression models for our study. To do this, we randomly select 100 stations from all available surface meteorological stations and apply the nonlinear methods with different configurations to predict daily and monthly averaged surface wind components. A controlled experiment approach is used, so that for each parameter being tested all other settings are the same. The parameters related to model architecture remain fixed for all stations considered once they have been determined by this analysis (done separately for each statistical method, averaging period, and season).

The NN model used for this study has the following structure. The architecture of the NN model is denoted $4 - N_h - 1$ (see Eq. 11), where 4 and 1 refer to the number of predictor and predictand variables respectively, and N_h is the number of hidden neurons. The number N_h is determined by controlled experiments as described above.

Since the length of the time series of daily averaged data is about 30 times longer than that of the monthly averaged data, more hidden neurons are used to model the daily averaged surface wind components. We choose $N_h = 5$ for daily-averaged data and $N_h = 2$ for monthly averaged prediction of surface wind components based on consideration of a range of values from $N_h = 2$ to $N_h = 30$. Not only the model architecture, but the training procedures for predicting daily and monthly averaged surface wind components also differ.

Early stopping is a method used to avoid overfitting the statistical model, in which a subset of the data is reserved to assess out-of-sample model performance during the training process. For early stopping, a fraction of data must be aside for validation. Early stopping is used for monthly averaged prediction by NN but not for daily averaged prediction. This choice is based on the results of the study of Amari et al. (1996). Their numerical experiments demonstrated that overfitting is not a problem when number of training samples (N observations) is larger than number of model parameters (N_p) by more than a factor of thirty ($N > 30N_p$), and that in this situation reserving a fraction of data for validation in early stopping is not better than using the whole dataset for training until convergence of the numerical optimization algorithm. However, when $N < 30N_p$, it is necessary to use measures to prevent overfitting such as early stopping. In our study, each training set uses 32 out of 33 years (i.e. 1980-2012) of winter or summer data. This means there are approximately 90×32 records of daily averaged data and 3×32 records of monthly averaged data. The number of parameters for a $4-5-1$ and $4-2-1$ NN model is 31 and 13 respectively. Daily prediction meets the threshold of $N > 30N_p$ but the monthly prediction does not.

The study of Amari et al. (1996) provided an equation to estimate the optimal fraction of data for validation,

$$f_{opt} = \frac{\sqrt{2N_p - 1} - 1}{2(N_p - 1)} \tag{8}$$

Based on Eq. (8), 16% of each segment of monthly averaged data for training (i.e. 32 out of 33 years) are randomly chosen as validation data for early stopping in each training session. Because of the different NN training procedures for daily averaged and monthly averaged data, the R package 'nnet' and the MATLAB Neural Network toolbox are used respectively for these analyses. Early stopping is automatically incorporated into the functions provided by MATLAB Neural Network toolbox. Although early stopping can be turned off in the MATLAB implementation, empirical tests in our study show that the MATLAB Neural Network toolbox is generally slower than 'nnet' in R, especially for large datasets. Therefore, we apply 'nnet' of R in predicting daily averaged data in this study.

Predictability by SVM regression is influenced by the choice of two parameters, ϵ and C, and the choice of kernel functions. The parameter ϵ sets the limit of the deviation of the regression solution from training data in the input space. However, the limit set by ϵ is not strict since observations exceeding the limit will also be "tolerated" (i.e. included in the regression process) as long as they are within the limit of deviation set by the box constraint C. The radial basis function (rbf) and linear function are chosen as the kernel types for daily and monthly averaged predictions, respectively, based on empirical tests as described before.

Modeling by RF requires the specification of the maximum number of features used for data partition when looking for the best split in tree regressions and the number of tree regressions used for ensemble averaging. In this study, the maximum number of features is taken to be the number of variables of predictors (i.e. 4), and the number of tree regressions is chosen to be 100. Empirical tests show that there is no evident improvement in predictability of both daily and monthly averaged surface wind components when the number of tree regressions in RF is larger than 100.

3 Results

In this study, we contrast two aspects of the predictability of surface wind components $\Pi(\theta)$ resulting from the linear regression model and the three nonlinear regression models: the overall magnitude and the directional variability of predictability respectively quantified by $\overline{\Pi(\theta)}$ and $\alpha(\Pi)$ (Eq. 7). Fig. 2 shows the spatial distribution of the regression method which gives the highest $\overline{\Pi(\theta)}$ out of the four methods considered (i.e. LR, NN, SVM and RF) at each station. Nonlinear methods (primarily SVM for daily-averaged and NN for monthly-averaged data), outperform LR in terms of $\overline{\Pi(\theta)}$ at most stations (Fig. 2). The same method generally gives the highest $\overline{\Pi(\theta)}$ for both summer and winter prediction at most stations, as is evident in the minimal seasonal differences of the the spatial patterns in Fig. 2. In contrast, there are distinct differences between the results of monthly and daily averaged predictions. For daily (monthly) averaged data, SVM (NN) outperforms other methods at the majority of stations. Those stations for which LR is the best method of predictability are generally distributed in the subtropics for daily timescales, and in the midlatitudes for monthly timescales.

Fig. 3 compares the values of $\overline{\Pi(\theta)}$ resulting from LR with those resulting from the three nonlinear methods: NN, SVM and RF. The difference between $\overline{\Pi(\theta)}$ resulting from LR and any of the three nonlinear regression methods is not large for most stations: the scatter of points falls close to the 1:1 line. In general, regions of relatively low predictability by linear regression remain poorly predicted using nonlinear

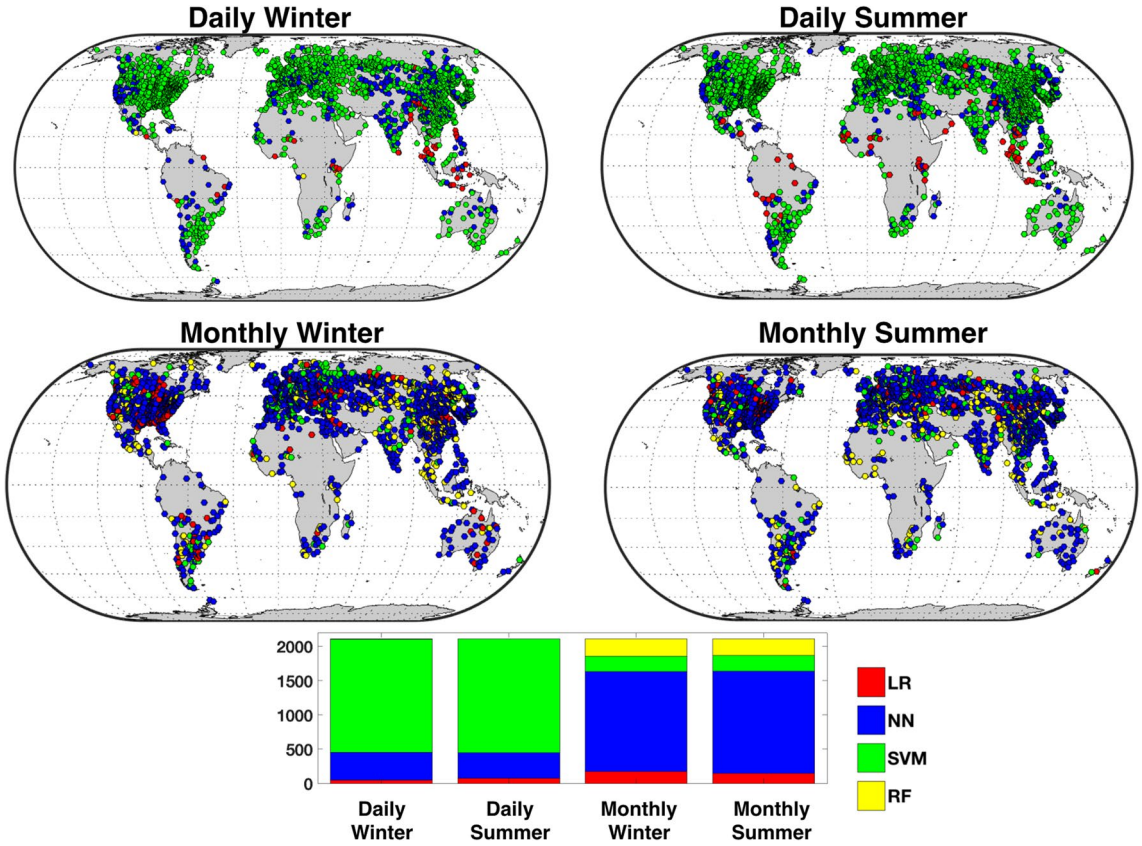

Fig. 2 Spatial distribution of the regression method which gives highest $\Pi(\theta)$ at each station for four cases of prediction: predicting daily and monthly averaged surface wind components in winter and summer. The number of stations belonging to each regression method that gives the highest $\Pi(\theta)$ are shown for each class of prediction

regression methods. The similarity between linear and nonlinear regression methods is most pronounced for NN and SVM; prediction skills from RF show substantially more scatter. Nonlinear regression methods differ from each other by the complexity of the algorithm and may be more or less advantageous in detecting an underlying nonlinear signal in different situations. For the data we consider, the predictability by RF is generally lower than those of NN and SVM (Figs. 2, 3). Also, consistent with the results of Monahan (2012), there is no systematic increase skill from monthly-averages of daily predictions relative to direct prediction of monthly-averaged quantities.

Fig. 4 compares the anisotropy of $\Pi(\theta)$ resulting from LR and the three nonlinear regression methods. The predictive anisotropy values generally fall around the 1:1 but with a larger scatter than the values of $\overline{\Pi(\theta)}$ in Fig. 3. Recall that substantially weakened predictive anisotropy resulting from nonlinear regression based prediction would indicate that the corresponding linear predictive anisotropy is the result of inadequate modeling of the predictand-predictor relationship along some directions of projection by LR, in other words, an artifact of LR. Broadly speaking, regions of strong(weak) predictive anisotropy resulting from linear regression are generally regions of strong(weak) anisotropy from nonlinear regression methods. However, a majority of stations have slightly weaker nonlinear predictive anisotropy resulting from at least one of the nonlinear regression methods (as indicated by the larger number of points falling above the 1:1 line than below it). Among the various nonlinear regression methods, the variation between linear and nonlinear $\alpha(\Pi)$ is largest for RF.

Our results show that nonlinear regression methods (NN, SVM and RF) do not substantially improve predictability of surface wind components relative to LR. Although at many stations linear predictive anisotropy is stronger than nonlinear predictive anisotropy resulting from at least one of the nonlinear regression, the differences are small. The results suggest that model complexity is not a major factor in determining the overall magnitude and anisotropy of statistical predictability of surface wind components.

We will now consider the spatial distribution of differences in predictive skill between linear and nonlinear regression models by comparing linear predictability, $\Pi(\theta)_{LR}$, with the highest predictability among the three

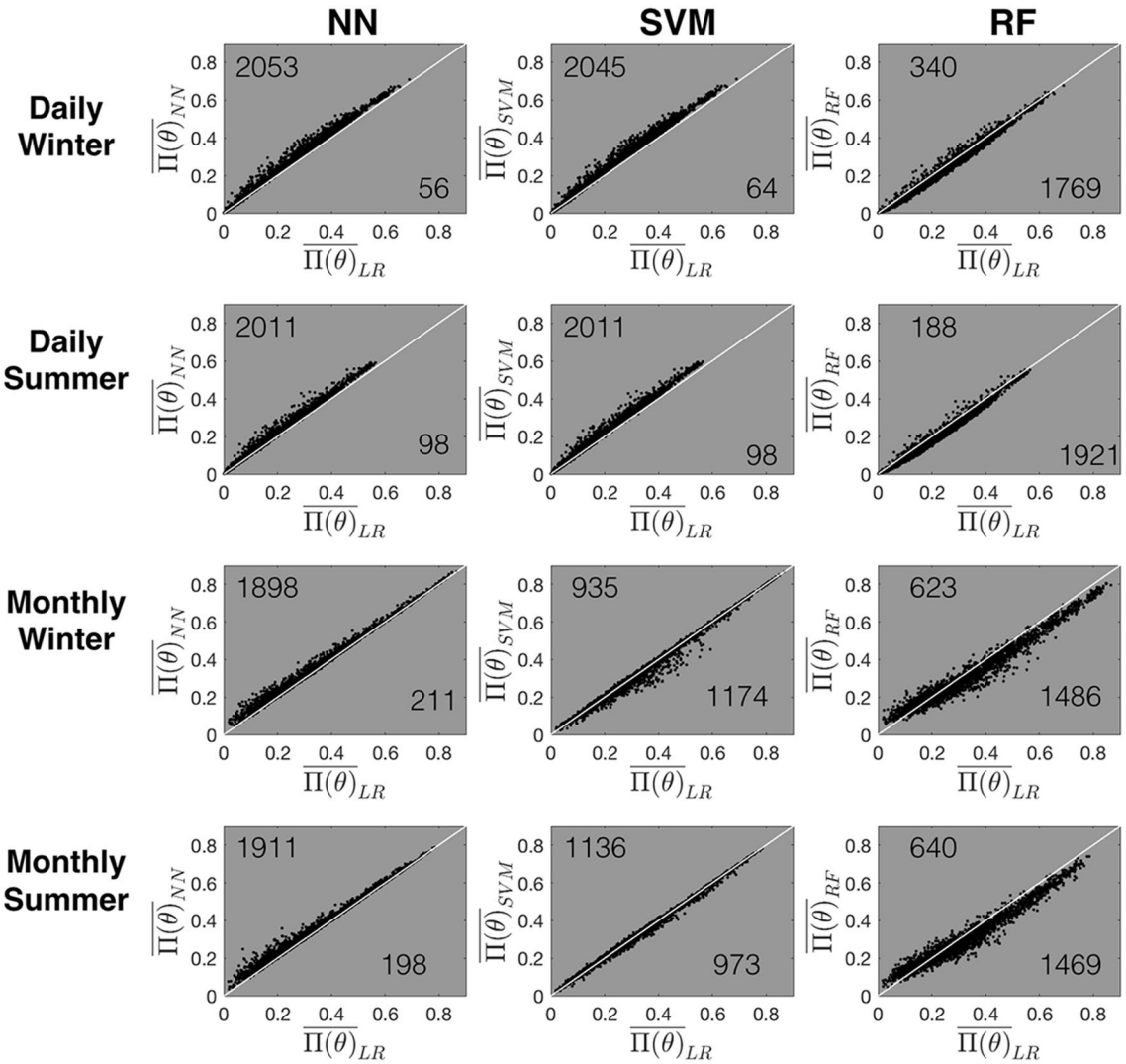

Fig. 3 Scatter plots of $\overline{\Pi(\theta)}$ of predictability of surface wind components resulting from LR against those resulted from NN, SVM and RF respectively at the 2109 surface meteorological stations are shown with the 1:1 line. The number of stations with larger $\overline{\Pi(\theta)}$ by linear regression and nonlinear regression is labeled at the lower right and upper left respectively

nonlinear regression models along each direction of projection, denoted $\Pi(\theta)_{NL}$. Specifically, we will consider the spatial distribution of $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ and $\frac{\alpha(\Pi)_{NL}}{\alpha(\Pi)_{LR}}$. Only wintertime results are shown as differences between seasons are relatively small. From a broad view, stations with large difference between linear and nonlinear prediction are usually located in regions characterized by surface heterogeneity, such as major mountain ranges or land/sea contrast (Figs. 5, 6). The pattern varies with time scales and is different for $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ and $\frac{\alpha(\Pi)_{NL}}{\alpha(\Pi)_{LR}}$. Most stations with large values of $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ for daily prediction are located in the mountainous regions of North America and Asia, with a few stations in coastal regions. For monthly prediction, large values of $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ are nearly absent in the North American Cordillera but are commonly found in the mountainous regions of Asia and South America as well as along the coast of South America, Africa and South East Asia. The association of large values of $\frac{\alpha(\Pi)_{NL}}{\alpha(\Pi)_{LR}}$ with mountainous and coastal regions is more obvious than that of $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ for both daily and monthly predictions. The locations with large values of $\frac{\overline{\Pi(\theta)_{NL}}}{\overline{\Pi(\theta)_{LR}}}$ and $\frac{\alpha(\Pi)_{NL}}{\alpha(\Pi)_{LR}}$ tend to correspond to locations of low linear predictability and strong linear predictive anisotropy. That is, the nonlinear models tend to outperform LR in areas of relatively low statistical predictability.

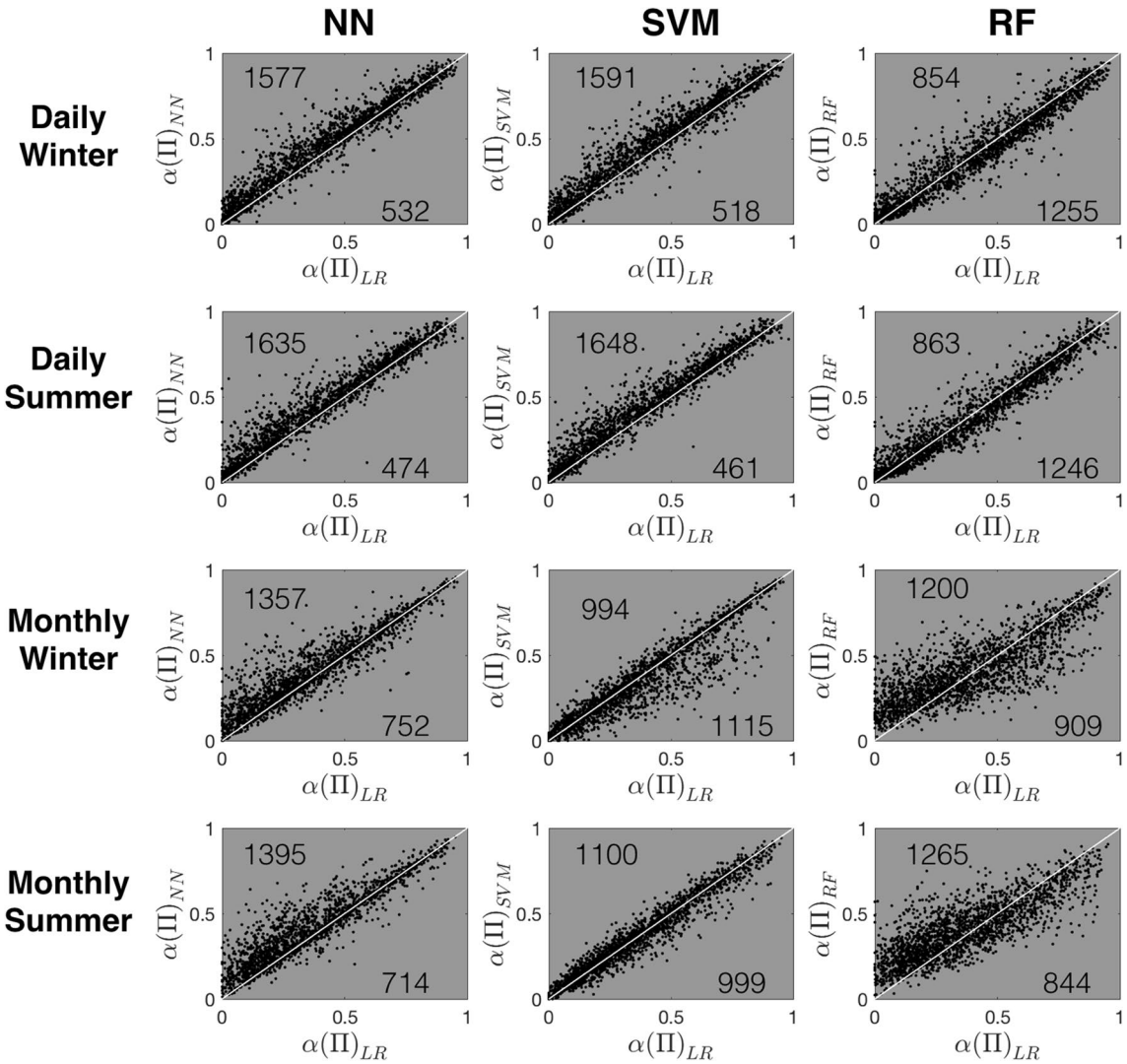

Fig. 4 As in Fig. 3 for $\alpha(\Pi)$

Representative stations are selected to illustrate typical locations where relatively large differences between $\Pi(\theta)_{LR}$ and $\Pi(\theta)_{NL}$ are observed (Fig. 7). Station information is listed in Table 1. Stations 1,2 and 5 are in mountainous regions, and stations 3 and 4 are characterized by land/sea contrast. At all stations considered, nonlinear predictability exceeds that of LR in all directions. For stations 2 and 3, predictive anisotropy resulting from linear regression is weakened greatly to almost isotropic predictability by nonlinear regression; in contrast, predictive anisotropy by nonlinear regression becomes stronger than that by linear regression at station 1. For other stations, the small increase in predictability by nonlinear regression does not substantially change the predictive anisotropy.

To test whether the difference between linear and nonlinear predictability is statistically significant at the 95% confidence level for a station, the quantities, $r_{LR} = \sqrt{\overline{\Pi(\theta)_{LR}}}$ and $r_{NL} = \sqrt{\overline{\Pi(\theta)_{NL}}}$ respectively representing mean directional correlation coefficients between the observed and predicted $U(\theta)$ are transformed to the normally distributed variables z_{LR} and z_{NL} by Fisher's transformation

$$z = 0.5(ln(1 + r) - ln(1 - r)). \tag{9}$$

It follows that the statistics $z_{NL} - z_{LR}$ to be normally distributed with the standard error $\sqrt{\frac{1}{N-3} + \frac{1}{N-3}}$, where N is the degrees of freedom at a station. Since the time series of daily averaged data is autocorrelated, N is taken as half of the total number of data points in time series of daily averaged data, whereas N is number of data points in time series of monthly averaged data.

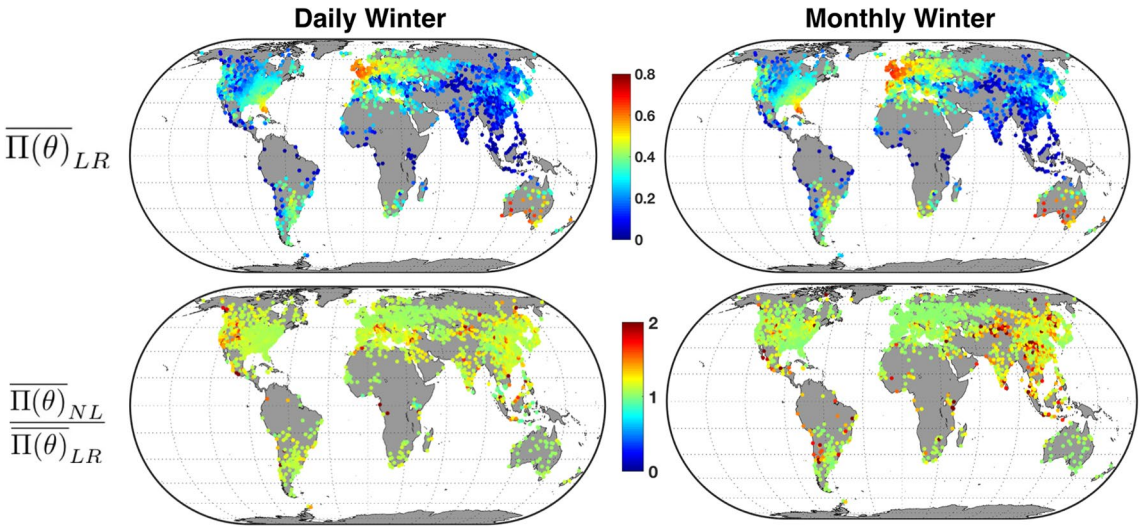

Fig. 5 Spatial distribution of $\overline{\Pi(\theta)}_{LR}$ and $\frac{\overline{\Pi(\theta)}_{NL}}{\overline{\Pi(\theta)}_{LR}}$ for daily and monthly averaged prediction of winter surface wind components

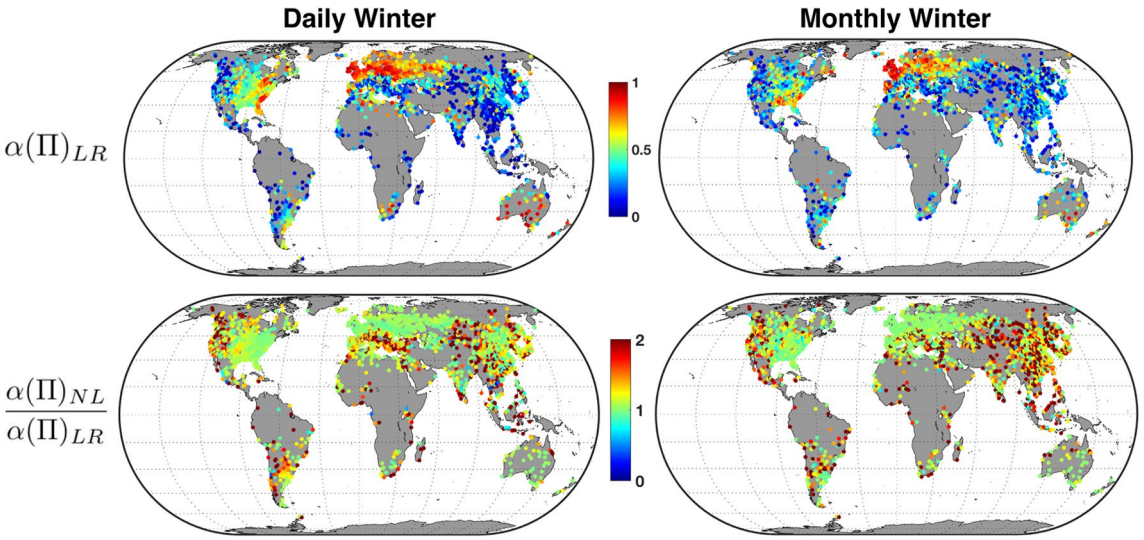

Fig. 6 Spatial distribution of $\alpha(\Pi)_{LR}$ and $\frac{\alpha(\Pi)_{NL}}{\alpha(\Pi)_{LR}}$ for daily and monthly averaged prediction of winter surface wind components

Table 1 Representative meteorological stations shown in Fig. 7. The name refers to the international ID of the stations

Station	Name	Location	Latitude (°)	Longitude (°)
1	LTAT	Malatya Erhac Airport, Turkey	38.4400	38.0900
2	LOWS	Salzburg Airport, Austria	47.7930	13.0040
3	LFKB	Bastia, France	42.5530	9.4840
4	KSFO	San Francisco Intl Airport	37.6190	− 122.3750
5	KNID	China Lake, Calfornia, USA	35.6860	− 117.6920

The results of one-tail z-test show that for predictions of daily averaged data, there are 244 (≈ 10%) stations in summer and 416 (≈ 20%) stations in winter with z_{NL} significantly larger than z_{LR} at 0.05 significance level; there is no station with $z_{NL} > z_{LR}$ at 0.05 significance level for prediction of monthly averaged data. There is no station with z_{LR} significantly larger than z_{NL} at 0.05 significance level for both daily and monthly prediction in both

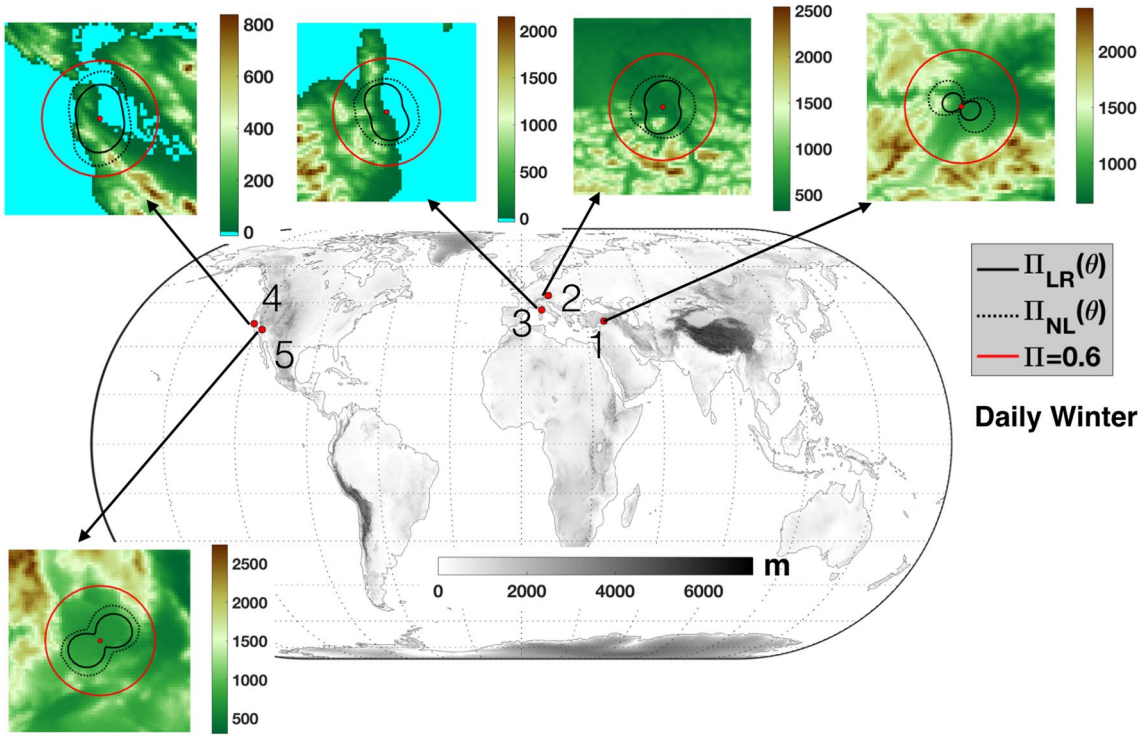

Fig. 7 Comparison of $\Pi(\theta)_{LR}$ with $\Pi(\theta)_{NL}$ resulting from the best predictability among NN, SVM, RF in each direction of projection of surface winds at five representative stations shown with topography for daily averaged prediction in winter

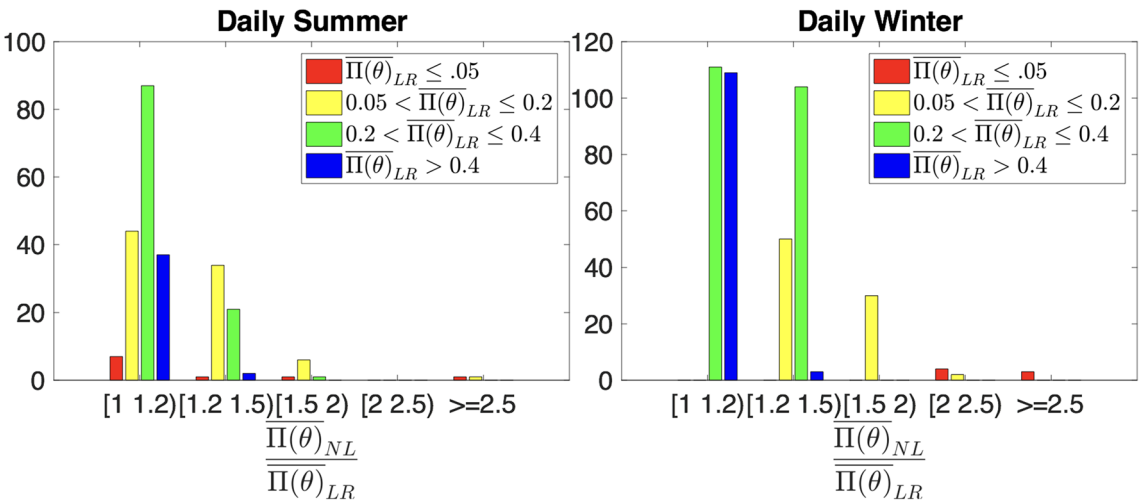

Fig. 8 Number of stations with statistically significant $z_{NL} > z_{LR}$ at the 0.05 significance level for daily predictions, sorted by relative improvement of nonlinear to linear prediction $\frac{\overline{\Pi(\theta)}_{NL}}{\overline{\Pi(\theta)}_{LR}}$ and by the magnitude of linear predictability

seasons. For predictions of daily averages, those situations in which z_{NL} is significantly larger than z_{LR} are generally associated with a small improvement in predictability by the nonlinear model (Fig. 8). In general, only stations with very small linear predictability (i.e. $\overline{\Pi(\theta)}_{LR} < 0.05$) correspond to large values of $\frac{\overline{\Pi(\theta)}_{NL}}{\overline{\Pi(\theta)}_{LR}}$. From a practical point of view, such improvement may not be very meaningful as there is not much utility in replacing a poor predictability (e.g. $R^2 = 0.01$) by linear regression with a better yet still poor predictability (e.g. $R^2 = 0.1$) by nonlinear regression

methods, particularly considering the relative computational expense of nonlinear regression methods.

Mao and Monahan (2017) showed that the strength and anisotropy of LR-based prediction of surface wind components are related to the magnitude and non-Gaussianity of their fluctuations (as measured respectively by standard deviation, σ and kurtosis, *kurt*). Specifically, the best-predicted surface wind components tended to be along the direction of the most variable fluctuations closest to having a Gaussian distribution. We now investigate if there are relationships between the directional structure of statistical properties of surface wind components (i.e. standard deviation and kurtosis) and differences in predictability resulting from linear and nonlinear regression models. In this analysis, we exclude stations of very small predictability (those for which both $\overline{\Pi(\theta)_{NL}}$ and $\overline{\Pi(\theta)_{LR}}$ are smaller than 0.1). Small differences in predictability between linear and nonlinear methods at such stations are not meaningful. First, we define $\beta(\theta) = \frac{\Pi_{NL}(\theta)}{\Pi_{LR}(\theta)}$ along each direction of projection. Fig. 9 shows the directional relationship of $\beta(\theta)$ with $\sigma(\theta)$ and $kurt(\theta)$ across all valid stations as measured by the histogram of the rank correlation coefficients $\rho(\sigma, \beta)$ and $\rho(kurt, \beta)$. Values of the rank correlation coefficient near 1 indicate that maxima of these directional quantities are aligned, as are minima, while values near -1 indicate that maxima in one quantity are oriented with minima in the other. Correlations near zero correspond to no common orientations of directional structures.

In general, $\rho(\sigma, \beta)$ and $\rho(kurt, \beta)$ are dominated by strong negative values and strong positive values respectively. This pattern is stronger for daily predictions and weaker for

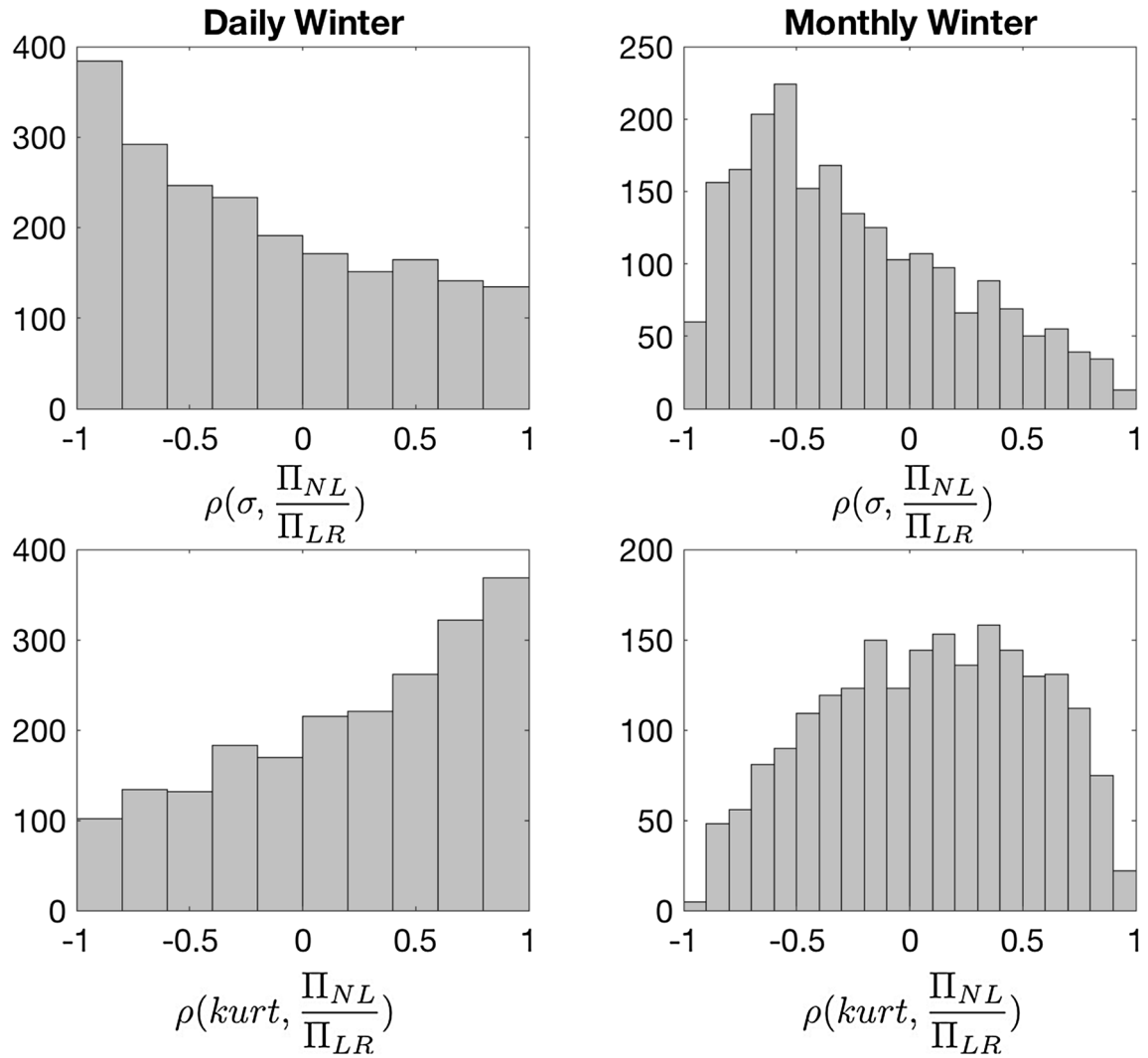

Fig. 9 Histograms of rank correlation coefficients between directional variation of $\beta(\theta) = \frac{\Pi(\theta)_{NL}}{\Pi(\theta)_{LR}}$ and the statistical properties of surface wind components: directional standard deviation $\sigma(\theta)$ and kurtosis $kurt(\theta)$ of surface wind components

monthly predictions. The result implies that the improvement of nonlinear predictability over linear predictability tends to be larger in the direction of surface wind components with small fluctuations and non-Gaussian distribution with heavier tails and sharp central peaks as indicated by smaller standard deviation and larger kurtosis. On the other hand, nonlinear predictability tends to be equal to or smaller than linear predictability in the direction of surface wind components with larger fluctuations and smaller kurtosis. As kurtosis values less than 3 and β values less than 1 are uncommon in our data, we see that LR tends to have comparable skill to nonlinear regression in the direction of surface wind components characterized by larger fluctuations and data distribution closer to Gaussian.

To assess how enhancement of predictive skill by nonlinear regression models can be related to the strength of fluctuations and non-Gaussianity of surface wind components, we plot the maximum of $\beta(\theta)$ over all directions as functions of the directional maxima of σ and kurtosis (Fig. 10). This figure also shows 25th, 50th, and 75th percentiles of $max(\beta)$ estimated from even-sized bins of $max(\sigma)$ and $max(kurt)$. Despite the presence of considerable scatter, the figure shows that relatively large increases of predictability by using nonlinear predictability (i.e. large values of max(β)) are more likely to be found at large values of max($kurt$) and small values of max(σ). We emphasize that relatively large values of β are generally associated with low overall predictability (not many stations with relatively large $max(\beta)$ correspond to high $\overline{\Pi}_{NL}$). For the majority of stations with $max(kurt)$ close to 3, Π_{NL} differs little from Π_{LR}, and stations characterized by large fluctuations of surface wind components (i.e. large values of $max(\sigma)$) also tend to have similar linear and nonlinear predictability.

4 Discussion

The general form of a regression model is:

$$y = f(x;\gamma) + \text{ noise}\,, \tag{10}$$

where $f(x;\gamma)$ is the model functional relationship between predictors and predictands, and γ is a vector of parameters. For a particular regression model applied to a particular data

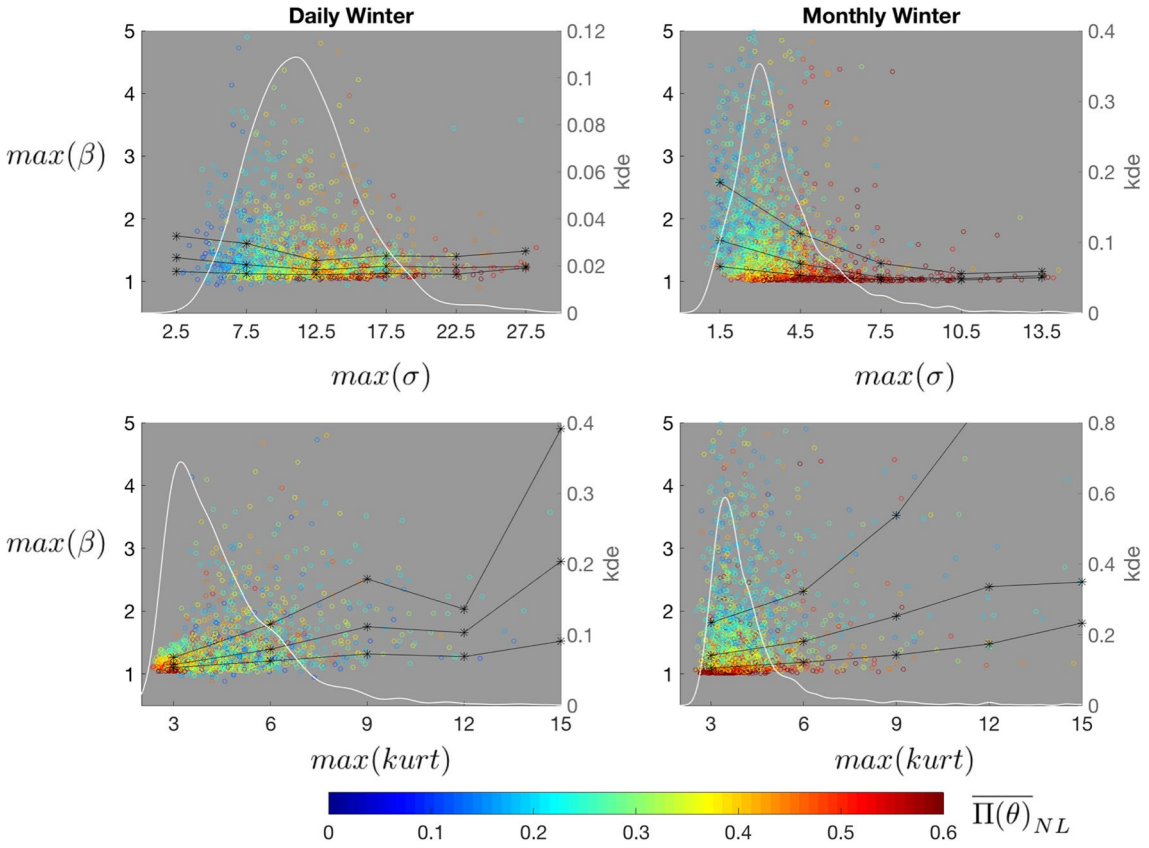

Fig. 10 Scatter plots of $max(\beta)$ vs $max(\sigma)$ and max($kurt$) of surface wind components colored by $\overline{\Pi(\theta)}_{NL}$. The white curve indicates kernel density estimation (kde) of the statistics: $max(\sigma)$ and $max(kurt)$. The black lines show the 25th, 50th and 75th percentile of $max(\beta)$ estimated from even-sized bins of statistical values $max(\sigma)$ and $max(kurt)$ centered at the values corresponding to each plot mark

set, the "noise" includes contributions from variations in y not dependent on x (the intrinsic noise), and systematic biases between the class of functional forms that can be taken by $f(x;\gamma)$ and the true functional relationship between predictor and predictand. Poor predictability of y by model $f(x;\gamma)$ can therefore result from : (1) low SNR of the relationship between predictors and predictands, resulting from large variability of the intrinsic noise or a weak predictor-predictand relationship, or (2) the inadequate representation of the predictor-predictand relationship by the regression model $f(x,\gamma)$. By considering a collection of linear and nonlinear regression models with the same predictors and predictands, we include a broad range of possible representations of the relationship between free-tropospheric fields and surface wind components.

If the predictor-predictand relationship is characterized by strong noise (i.e. small values of SNR), the resulting statistical predictability will not be high regardless of the type of regression. With large intrinsic noise, the cross-validated predictability resulting from linear regression may exceed that of a nonlinear regression model because nonlinear regression models generally have a larger number of parameters and are more prone to fitting noise. In contrast, if the predictor-predictand relationship is characterized by relatively weak intrinsic noise, the predictability may depend systematically on the type of regression models used for statistical prediction. If the underlying predictive signal is strong (with a large SNR), both linear regression and nonlinear regression models will result in high predictability when the underlying relationship is close to linear, so long as the nonlinear regression considered can reduce to linear regression (as is the case for all of the nonlinear models we consider). In contrast, when the true relationship is nonlinear, prediction by LR should be inferior to that by the other regression methods.

Following the above discussion, there are two possible reasons that may explain why nonlinear regression does not substantially outperform linear regression. First, there is no evident nonlinear predictive signal, and second, the intrinsic noise associated with the predictor-predictand relationship is strong. The information we have at hand cannot unambiguously indicate which of these two possible reasons is more important in explaining why nonlinear regression cannot greatly outperform linear regression at most stations. However, the relationships between the statistical properties of surface wind components and β can provide us some clues on this issue. Mao and Monahan (2017) show that high linear predictability tends to be associated with surface wind components characterized by relatively large fluctuations with near-Gaussian distribution, and poor linear predictability is generally associated with wind components characterized by small fluctuations and non-Gaussian distribution. The association of high linear predictability and $\beta \approx 1$ with surface wind components characterized by large fluctuations and Gaussian distribution may suggest that the nonlinearity of predictive signals at these stations is weak.

On the other hand, our results show a systematic improvement in predictive skill resulting from nonlinear regression for surface wind components characterized by small fluctuations and non-Gaussian distribution with heavy tails and sharp central peaks. Although the trend of systematic improvement is not strong, it suggests that the nonlinear predictive signal is stronger than linear predictive signal at these stations. We emphasize that any such nonlinear predictive signals in general are weak at these stations as indicated by low predictability for both linear and nonlinear models. In this situation, the predictor-predictand relationships are dominated by strong noise which renders both nonlinear and linear regression methods of limited utility (i.e. with large modeling errors). Finally, we cannot rule out the possibility that overfitting may also play a role in causing $\beta > 1$ for stations characterized by low predictability resulting from both linear and nonlinear regressions. Although cross-validation should reduce the chance of overfitting, it cannot completely eliminate overfitting. In particular, some aspects of TFs are not cross-validated, such as model architecture and predictor selection; therefore, we cannot rule out the possibility of a slight inflation of skill due to factors such as these, especially for nonlinear regression models which are more prone to overfitting than linear regression, hence, $\beta > 1$.

5 Conclusion

We have evaluated the predictability of surface wind components by statistical prediction using linear and a range of nonlinear regression methods as transfer functions at 2109 land stations across a wide range of locations, and the results in this study are based on predictability of conditional expectation of surface wind components given predictors. The results show that linear predictability of surface wind components at most stations is lower than predictability obtained by at least one of the three nonlinear regression methods. Except for a small number of stations, the difference in predictability of surface wind components by linear regression and the best of the three nonlinear regression methods is generally not substantial and significant at the 0.05 level. Where there are improvements, these are generally found at stations where the overall level of predictability is low for all methods. Moreover, the anisotropy of predictability of surface wind components projected along different directions is a common feature for both linear and nonlinear regression methods, although stations with strong linear predictive anisotropy tend to have slightly weaker nonlinear predictive anisotropy. We found that many of the stations with relatively large changes in the magnitude and degree of anisotropy of predictability are in regions characterized by

surface heterogeneity (e.g. major mountain ranges, land-water contrast). This statement cannot be reversed: not all stations in regions of heterogeneous landscape show relatively large differences in skill between linear and nonlinear regression predictions. Moreover, although the difference between linear and nonlinear predictability is generally small, where they do differ the nonlinear predictability tends to be higher than linear predictability in the direction of wind components with smaller fluctuations and non-Gaussian distributions characterized by heavier tails or higher peaks. This pattern is move obvious for daily averaged prediction than monthly averaged prediction. Averaging time series generally causes them to become more Gaussian and the statistical relationship between variables to be more linear (Yuval and Hsieh 2002), which may explain the fact that improvements in nonlinear prediction relative to linear prediction are associated with surface wind components with more non-Gaussian distributions. However, the connection between non-Gaussianity and nonlinear predictability is not strong for these data.

Alternative methods exist for selecting predictors for the prediction of surface wind components. For example, there is evidence to suggest that linear predictability can be improved by Lasso regression which automatically selects grid predictors in the domain, although one cannot rule out the possibility of overprediction in this case. The investigation of different robust methods of predictor selection is a useful direction of future study.

The relatively small change in skill of prediction by linear and nonlinear regression methods indicates that statistical predictability of surface wind components is generally not substantially influenced by the functional form but limited by the intrinsic noise. Future study is needed to assess what physical factors determine the magnitude of the intrinsic noise in the predictor-predictand relationship, which following Mao and Monahan (2017) we interpret as being associated with local variability of atmospheric circulations on meso- and smaller scales. Our results suggest that since nonlinear predictive models generally do not substantially improve predictability of surface wind components using the predictors in this study, there is no advantage building the transfer function of statistical prediction using nonlinear regression methods which are more computationally expensive to apply than linear regression.

Acknowledgements The authors gratefully acknowledge helpful comments and suggestion from two anonymous reviewers. This research was supported by the Discovery Grants program of the Natural Sciences and Engineering Research Council of Canada.

Appendix: Nonlinear regression methods

There are three nonlinear regression methods used in this study: neural network (NN), support vector machines (SVM) and random forests (RF). These three methods are supervised learning methods in which a function is inferred from a set of training data consisting of pairs of predictors and predictands in a process referred to as training. After training, the estimated function can be used for new predictions. Supervised learning methods differ from each other by the algorithms they use to infer the regression function from training data. A brief introduction to the NN, SVM and RF algorithms will now be presented.

Neural network

The method of NN in nonlinear regression is inspired from the structure of neurons in biology. There are many types of neural network. In this study we use feed-forward neural networks, which is the most widely-applied NN. In feed-forward NN, the predictand is related to the predictor by a sequence of linked computational elements known as hidden layers (Hsieh 2009). Figure 11 demonstrates the structure of a feed-forward NN used to model a regression with P predictors and K predictands by a single layer of M hidden neurons. This structure is similar to that of a two-stage regression model (Hastie et al. 2009).

In the first stage of the NN regression, the values of the hidden neurons $Z = [Z_1, \ldots, Z_m, \ldots, Z_M]$ are computed as linear combinations of the inputs $X = [X_1, \ldots X_p, \ldots, X_P]$,

$$Z_m = s(\alpha_{0m} + \alpha_m^T X), \quad m = 1, \ldots, M, \tag{11}$$

where α_{0m} are the offset scalars and α_m are the weight vectors of predictors X. The function $s(v)$ is usually chosen to be the sigmoid function $s(v) = 1/(1 + e^{-v})$ which asymptotically saturates for large positive and negative values of v: $s(v) \to 0$ as $v \to -\infty$ and $s(v) \to 1$ as $v \to +\infty$. In the second stage of the regression models, each predictand Y_k is computed as the linear combination of hidden neurons $Z = [Z_1, \ldots, Z_m, \ldots, Z_M]$,

$$\widehat{Y}_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \ldots, K, \tag{12}$$

where $\widehat{Y}_k$ is the modeled value of target Y_k, β_{0k} are the offset scalars, and β_k are the weight vectors. Together, the parameters α_{0m}, α_m and β_{0m}, β_k are also called the weights of the NN model.

Training of the NN is done using an optimization algorithm to seek weights of the model which minimize the objective function $J = \sum_{k=1}^{K} \sum_{i=1}^{N} (Y_{ik} - \widehat{Y}_{ik})^2$ where N is the number of observations. Parameters of the initial state can be chosen randomly. The minimization of the objective function J is done iteratively until some convergence criterion is met. The primary issue we need to be mindful of in training neural network is overfitting. Measures to prevent overfitting should be considered in both model architecture (i.e. number of hidden neurons) and model training.

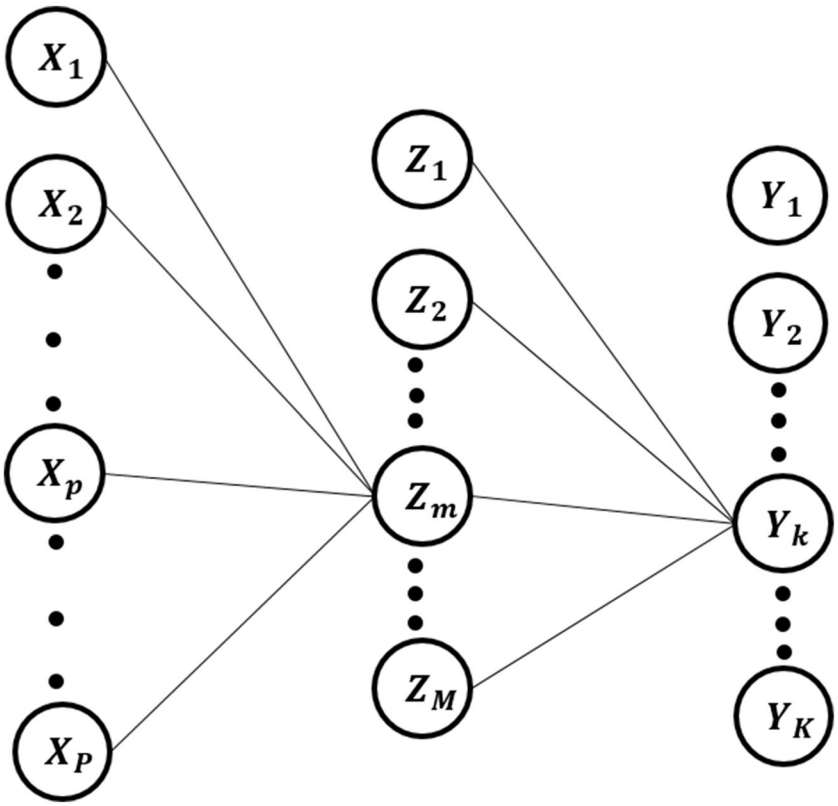

Fig. 11 Schematic of a single hidden layer, feed-forward neural network, where X_i and Y_i are predictors and predictands, and Z_i are the hidden neurons

Hidden neurons The complexity of a neural network is increased by increasing the number of hidden neurons. A neural network with one hidden layer of a finite (but sufficiently large) number neurons can approximate any continuous function to arbitrary accuracy (Csáji 2001). The modeling challenge is choosing the right number of hidden neurons. Models with too few hidden neurons might not have the flexibility to capture the nonlinear signal of the data. Models with too many hidden neurons might be so flexible that they will fit the noise of the data. The appropriate number of hidden neurons can be chosen from empirical testing (as described in Sect. 2).

Training methods Early stopping is a common method of preventing overfitting, in which the training process stops well before J reaches the global minimum. In this method, the training data are divided into two subsets. All data in the first subset undergo the training process as described above to update the weights of the model. Each iteration in the training is referred to as an epoch. The training process is repeated for many epochs. The second dataset is used to evaluate the objective function at each epoch. As the number of training epochs increase, the value of J over the validation data generally decreases at first before increasing. Training beyond the point at which J starts to increase will only contribute to model overfitting and is therefore stopped. The model parameters are those obtained at the minimum of J over the validation data. Note that the separation into training and validation sets is done as part of the parameter estimation process, and is distinct from the data subsetting associated with cross-validation.

Support vector machine

Support vector machines are characterized by the use of kernel functions which represent the mapping of observations in the input space into a high-dimensional feature space where linear regression can be used. Therefore, a nonlinear model in the input space can be learned from a subset of observations (support vectors) by linear regression in the high-dimensional feature space. The following mathematical formulation is based on the documentation of the Statistics and Machine Learning Toolbox in MATLAB (MathWorks 2017c).

Suppose x_n represents one case of training data in the input space with observed response value y_n, and $g(x_n)$ represents a mapping function which maps x_n in the input space to the feature space. The function, $f(x_n)$, which is used to model y_n, can be constructed as

$$f(x_n) = g(x_n)^T \beta + b \tag{13}$$

The goal of SVM regression is to find a function $f(x)$ that deviates from y by a value no larger than ϵ for each observation of training data, and at the same time is as flat as possible. Flatness requires that the weight vector of $f(x)$ should

be as small as possible. The problem is formulated as minimizing the objective function

$$J(\beta) = \frac{1}{2}\beta^T\beta, \tag{14}$$

subject to the constraint that all residuals less than ϵ: that is for all n, $|y_n - (g(x_n)'\beta + b)| \le \epsilon$. The set of x values satisfying $|y - f(x)| \le \epsilon$ is known as the ϵ-tube. However, for a given ϵ, not all observations may satisfy the constraint of falling in the ϵ-tube; therefore, slack variables ζ_n and ζ_n^* are used to allow the regression models to tolerate errors up to the value of $\epsilon + \zeta_n$ or $\epsilon + \zeta_n^*$, where ζ_n and ζ_n^* represent the upper and lower limit of the extension of the ϵ-tube respectively (Fig. 12).

By including slack variables, the minimization of the objective function becomes

$$J(\beta) = \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{N}(\zeta + \zeta^*), \tag{15}$$

subject to: for all n

$y_n - (x_n'\beta + b) \le \epsilon + \zeta_n$,
$(x_n'\beta + b) - y_n \le \epsilon + \zeta_n^*$,
$\zeta_n \ge 0$
$\zeta_n^* \ge 0$.

The constant $C > 0$ determines the largest deviations from ϵ which can be tolerated. The formulation of $J(\beta)$ in Eq. (15) is also known as the primal formula (Vapnik 2013).

Estimating the weights of $f(x)$ of SVM regression in Eq.(15) corresponds to minimizing the ϵ-insensitive loss function defined as

$$L_\epsilon = \begin{cases} 0 & \text{if } |y - f(x)| \le \epsilon \\ |y - f(x)| - \epsilon & \text{if } |y - f(x)| > \epsilon \end{cases} \tag{16}$$

for which the errors associated with observations within the ϵ-tube are ignored (Vapnik 2013). This optimization problem of minimizing Eq. (16) is computationally simpler to solve in its Lagrange dual formulation (MathWorks 2017c). Constructing a Lagrangian function for Eq. (16) requires nonnegative multipliers α_n and α_n^* for each observation x_n. The dual formulation of minimizing Eq. (16) involves minimizing

$$\begin{aligned} L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) < g(x_i), g(x_j) > \\ + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \epsilon\sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*), \end{aligned} \tag{17}$$

where $< g(x_i), g(x_j) >$ is the inner product of the predictors after mapping, and Eq. (17) is subject to

$\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0$,
and that for all n:
$0 \le \alpha_n \le C$,
$0 \le \alpha_n^* \le C$,
$\alpha_n(\epsilon + \zeta_n - y_n + f(x_n)) = 0$,
$\alpha_n^*(\epsilon + \zeta_n^* - y_n + f(x_n)) = 0$,
$\zeta_n(C - \alpha_n) = 0$,
$\zeta_n^*(C - \alpha_n^*) = 0$.

These conditions indicate that Lagrange multipliers $\alpha_n = 0$ and $\alpha_n^* = 0$ when observations are inside the ϵ-tube. The dual formulation Eq. (17) is solved by using quadratic programming techniques the details of which are beyond the scope of this paper but can be found in Platt (1998). The solution has the form:

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*) < g(x_n), g(x) > + b. \tag{18}$$

In other words, the solution $f(x)$ only depends on those x_n which satisfy $(\alpha_n - \alpha_n^*) \ne 0$, and this subset of x_n from training data is denoted support vectors which fall within a distance ζ or ζ^* from the boundary of ϵ-tube as shown in Fig. 12. Since by construction of SVM regression models, most cases of training data are inside the ϵ tube, the number of support

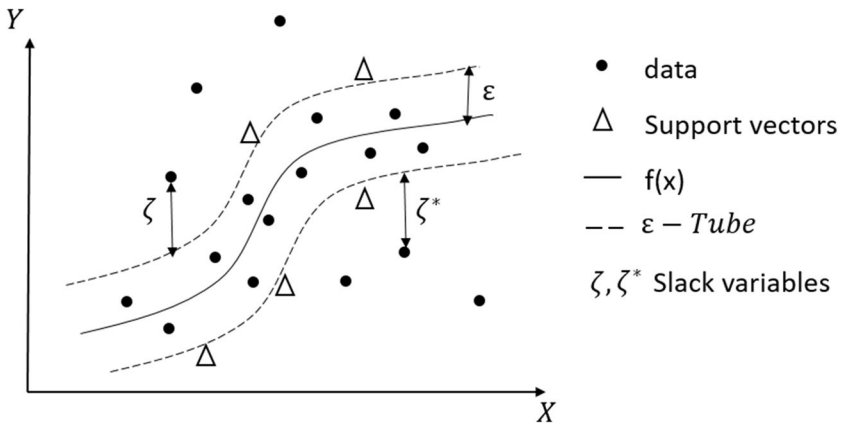

Fig. 12 Schematic of regression by support vector machine

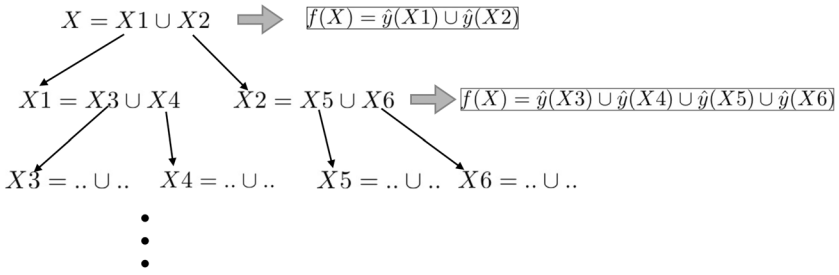

Fig. 13 Schematic of a tree regression by binary recursive partitioning. X represents the training data in the input space. $\hat{y}$ refers to a simple model used to fit the training data in each sub-space after the split. f stands for the resultant tree model at each level of partitioning

vectors is small comparing to the number of observations in training data. The transformation function $g(x)$ that maps each x_n of training data in the input space to the feature space is unknown, but $< g(x_n), g(x) >$ in Eq. (19) can be approximated by kernel functions $K(x_n, x) =< g(x_n), g(x) >$. There are many admissible kernel functions, and we list some common kernel types in the following:

- linear kernel: $K(x_n, x) =< x_n, x >$
- polynomial kernel: $K(x_n, x) = (1 + x'_n x)^p$, where $p = 2, 3, 4 \ldots$
- radial basis function kernel: $K(x_n, x) = \exp(-\gamma \|x_n - x\|^2)$, $\gamma > 0$
- sigmoid kernel: $K(x_n, x) = \tanh(\gamma < x_n, x > +\tau), \gamma > 0$.

The class of functions that can be approximated are determined by the chosen kernel form.

Generally, there are three factors that can influence the accuracy of SVM regression: the values of C, ϵ and the chosen kernel form. The parameter C determines the trade-off between model complexity (i.e. the flatness of $f(X)$) and the degree to which deviations larger than ϵ are tolerated in the optimization of the loss function. Larger C indicates that more data are used in the process of training but the resulting model can be complex and overfit the data; on the other hand, smaller C might make the regression model prone to underfitting because the training process might not have enough training data to characterize the underlying structure. By controlling the width of the ϵ-tube in the training data, the parameter ϵ influences the number of support vectors used in the training process. Finally, depending on the properties of training data, some kernel forms may work better than others.

Random forests

Random forests, first proposed by Breiman (2001), belong to the category of ensemble learning methods that generate many regression models and aggregate their results (Hastie et al. 2009). As the name, random forest, suggests, the individual model is a tree-based regression model. The basic idea of a tree-based regression is partitioning the input space into a set of subspaces, and then fitting a simple model (usually a constant) in each subspace. The partition is generally done by binary recursive partition in which the input space is first split into two regions, and each sub-region is fit with a simple model. This splitting process is repeated in the resulting sub-regions until some stopping rule is applied as illustrated in Fig. 13. The following formulation is based on Hastie et al. (2009). Suppose there are p input variables and one response for each of N observations in a dataset: that is, (x_i, y_i) for $i = 1, 2, \ldots N$ with $x_i = (X_{i1}, X_{i2}, \ldots X_{ip})$. Starting from the entire input space, the two planes after each split can be expressed as $R_1(j, s) = X|X_j \leq s$ and $R_2(j, s) = X|X_j > s$, where X_j and s represent the splitting variable and splitting point respectively. The best fit is achieved by seeking the variable X_j and split point s which solve

$$\min \left[\min \sum_{x \in R_1(j,s)} (y_i - c_1)^2 + \min \sum_{x \in R_2(j,s)} (y_i - c_2)^2 \right], \quad (19)$$

where c_1 and c_2 are the constants used to model values in R_1 and R_2 respectively. The split variable X_j and split point s which lead to best fit can be determined by scanning through input variables x_i, and depending on the objective of a regression problem, it is not necessary to use all p input variables in determining the best fit. The response y_i can be modeled by the tree regression

$$T(x) = \sum_{m=1}^{M} c_m I(x \in R_m), \quad (20)$$

where R_1, R_2, ..., R_M are regions resulting from partition in the tree regression, and c_1, c_2, ..., c_m are the constants used to model responses in the corresponding region.

Tree regression is generally prone to overfitting as the training processes use all training data including noise. The essential idea is to average many noisy but approximately unbiased models, thereby reducing the variance (Hastie et al. 2009). Averaging many regression trees constructed randomly from bootstrapped samples, the resulting model can be expressed as

$$f(X) = \frac{1}{B} \sum_{b=1}^{B} T(X_b), \quad (21)$$

where B is the total number of trees, and b characterizes the bth tree-based regression $T(X)$ from a bootstrap sampling. In general, there is no specific rule to determine the fraction of data used for bootstrap model construction. The Python function 'RandomForestRegressor', which we used in this study, adopts the strategy of building each tree from drawing a sample of equal size of the input training data with replacement. The estimate of error is obtained by prediction over data not in the bootstrap sample (out-of-bag or OOB data)

One of the biggest merits of random forest analysis is its simplicity. There are only two parameters for RF, and the solution is not very sensitive to their values (Liaw and Wiener 2002). The first parameter is the number of input variables needed to consider when seeking the best split during tree regression. The second parameter is the number B of individual tree regressions used in RF.

References

Amari SI, Murata N, Müller KR, Finke M, Yang HH (1996) Statistical theory of overtraining-is cross-validation asymptotically effective? In: Advances in neural information processing systems, pp 176–182

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Csáji BC (2001) Approximation with artificial neural networks. Faculty of Sciences, Eötvös Loránd University, Hungary 24:48

Culver AM, Monahan AH (2013) The statistical predictability of surface winds over western and central Canada. J Clim 26(21):8305–8322

Davy RJ, Woods MJ, Russell CJ, Coppin PA (2010) Statistical downscaling of wind variability from meteorological fields. Boundary-layer Meteorol 135(1):161–175

Hastie TJ, Tibshirani RJ, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

He Y, Monahan A, Jones C, Dai A, Biner S, Caya D, Winger K (2010) Land surface wind speed probability distributions in North America: observations, theory, and regional climate model simulations. J Geophys Res 115(D04):103

Holtslag A, Svensson G, Baas P, Basu S, Beare B, Beljaars A, Bosveld F, Cuxart J, Lindvall J, Steeneveld G et al (2013) Stable atmospheric boundary layers and diurnal cycles: challenges for weather and climate models. Bull Am Meteorol Soc 94(11):1691–1706

Hsieh WW (2009) Machine learning methods in the environmental sciences: neural networks and kernels. Cambridge University Press, Cambridge

van der Kamp D, Curry CL, Monahan AH (2012) Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. II. Predicting wind components. Clim Dyn 38(7–8):1301–1311

Kanamitsu M, Ebisuzaki W, Woollen J, Shi-Keng Y et al (2002) NCEP-DOE AMIP-II reanalysis (r-2). Bull Am Meteorol Soc 83(11):1631

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22

Mao Y, Monahan A (2017) Predictive anisotropy of surface winds by linear statistical prediction. J Clim 30(16):6183–6201

MathWorks (2017a) fitrsvm. https://www.mathworks.com/help/stats/fitrsvm.html

MathWorks (2017b) Getting Started with Neural Network Toolbox. https://www.mathworks.com/help/nnet/getting-started-with-neural-network-toolbox.html

MathWorks (2017c) Understanding Support Vector Machine Regression. https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

Mohandes M, Halawani T, Rehman S, Hussain AA (2004) Support vector machines for wind speed prediction. Renew Energy 29(6):939–947

Monahan AH (2012) Can we see the wind? Statistical downscaling of historical sea surface winds in the subarctic northeast Pacific. J Clim 25(5):1511–1528

Platt J (1998) Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep, Microsoft Research

Python (2016) RandomForestRegressor-scikit-learn 0.18.1 documentation. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Ripley B, Venables W (2016) Package 'nnet'. https://cran.r-project.org/web/packages/nnet/nnet.pdf

Sailor D, Hu T, Li X, Rosen J (2000) A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change. Renew Energy 19(3):359–378

Salameh T, Drobinski P, Vrac M, Naveau P (2009) Statistical downscaling of near-surface wind over complex terrain in southern France. Meteorol Atmos Phys 103(1–4):253–265

Stull RB (2000) Meteorology for scientists and engineers: a technical companion book with Ahrens' Meteorology Today. Brooks/Cole

Sun C, Monahan A (2013) Statistical downscaling prediction of sea surface winds over the global ocean. Journal of Climate 7938–7956

Vapnik V (2013) The nature of statistical learning theory. Springer Science & Business Media, New York

Wolfram (2016) WeatherData source information. http://reference.wolfram.com/language/note/WeatherDataSourceInformation.html. Accessed 1 Jan 2016

Yuval Hsieh W (2002) The impact of time-averaging on the detectability of nonlinear empirical relations. Quarterly Journal of the Royal Meteorological Society 583(1609–1622)

Zhao P, Yu B (2006) On model selection consistency of Lasso. Journal of Machine learning research 7(Nov):2541–2563