Statistical Prediction of Winter Haze Days in the North China Plain Using the Generalized Additive Model

ZHICONG YIN

Key Laboratory of Meteorological Disaster, Ministry of Education, and Joint International Research Laboratory of Climate and Environment Change, and Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing, and Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

HUIJUN WANG

Key Laboratory of Meteorological Disaster, Ministry of Education, and Joint International Research Laboratory of Climate and Environment Change, and Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing, and Nansen-Zhu International Research Centre, Institute of Atmospheric Physics, and Climate Change Research Center, Chinese Academy of Sciences, Beijing, China

(Manuscript received 19 January 2017, in final form 11 June 2017)

ABSTRACT

Winter (December–February) haze days in the North China Plain (WHDNCP) have recently dramatically increased. In addition to human activities, climate change and variability also contributed to the severe situation and supported the possibility of seasonal predictions. In this study, using the generalized additive model (GAM), the sea surface temperature around the Alaska Gulf and sea ice area of the Beaufort Sea were selected as the predictors to establish a statistical prediction model (SPM). The difference between the current and previous year of WHDNCP (WDY) was predicted first and was then added to the observation of the previous year to obtain the final predicted WHDNCP. For WDY prediction, the root-mean-square error of the SPM using GAM was 3.01 days. In addition to the annual variation, the tropospheric biennial oscillation features and the dramatically increasing trend after 2010 were both captured successfully. Furthermore, for the final predicted WHDNCP anomalies, the long-term trend and turning points were simulated well, and the percentage of the same mathematical sign was 91.7%. Independent prediction tests were performed for 2014 and 2015, and the forecast bias was 0.86 and 0.19 days, respectively. To assess the predictive ability, recycling independent tests (including real-time hindcasts for the period 2005–15) were also applied, and the percentage of the same sign was 100%.

1. Introduction

In winter (October–February), severe haze is a macroscopic manifestation of air pollution that affects society, the climate system (Liao et al. 2015), and human health (Chen et al. 2013). The number of haze days has recently increased markedly in the economically developed regions of eastern China (Ding and Liu 2014; Wang and Chen 2016). For example, in January of 2013, because of a persistent and heavy haze event (Zhang et al. 2014), Beijing, China, released its first orange (the second-highest level) haze warning in its meteorological history. Since then, the orange haze warning has been released several times each year in Beijing, and the first red (the highest level) air-pollution warning occurred in December of 2015. As represented by Beijing, the cities in the North China Plain (NCP) suffered from severe-haze events that underwent similar trends (Yin and Wang 2016a). Therefore, this study aims to develop a statistical prediction model (SPM) of winter haze days in the NCP (WHDNCP).

Early studies have documented that human activities, such as rapid urbanization and fossil-fired economies, play a great role in this long-term trend of haze days in eastern China (Wang et al. 2013). In addition, some recent studies documented that climate factors are also efficient contributors to the variation in haze days (Yang et al. 2016; Wang and Chen 2016). The occurrence of severe haze in the NCP was correlated with specific atmospheric circulations, such as weakened northerly...
winds near the surface, an enhanced temperature inversion in the boundary layer, a northward East Asian jet stream (EAJS) in the upper atmosphere (Chen and Wang 2015), and a weakened East Asian winter monsoon (Yin et al. 2015). Sea surface temperature (SST) anomalies over the subtropical western Pacific Ocean during the preceding autumn (September–November) induced the Pacific–East Asia teleconnection and influenced WHD_{NCP} (Yin and Wang 2016a; Yin et al. 2017). The Atlantic SST also exhibited a significant inphase relationship with WHD_{NCP} on both decadal and interannually time scales (Xiao et al. 2015). A decline in the Arctic sea ice (SI) during the preceding autumn also showed a close relationship with an increase of haze pollution in eastern China (Wang et al. 2015). Although studies on the influences of climate anomalies on haze are still sparse, these new findings suggest a more feasible means for seasonal prediction of WHD_{NCP}.

As mentioned above, the external forcings (e.g., SST and Arctic SI) could modulate the incidence of winter haze in the NCP through atmospheric circulations, and we speculate that these processes may likely include some nonlinear processes. The generalized additive model (GAM) allows the data to determine the shape of a smooth function (Yee and Mitchell 1991) and displayed good performance at handling complex nonlinear and nonmonotonic relationships (Hastie and Tibshirani 1990). In a study by Yin and Wang (2016b), predictions of WHD_{NCP} by GAM and multiple linear regression (MLR) showed that both methods performed well and analogously, which seemingly indicated that a linear relationship dominated the prediction of the difference between the current and previous year (DY) of WHD_{NCP} after including seven predictors. A question raised here was whether nonlinear relationships existed and were important for the predictions of WHD_{NCP}. We speculated that nonlinear processes should exist but were being counteracted by internal interactions among the multiple predictors, that is, the multicollinearity problem. To address the above issue, we chose only two predictors that were also employed by Yin and Wang (2016b), that is, Pacific SST and Arctic SI, and used the GAM approach to train the SPM. This prediction of WHD_{NCP} on the basis of fewer predictors is more convenient for real-time operations, represents the general conditions of winter air pollution, and is beneficial for policies to mitigate air pollution in China.

The data and methods employed are introduced in section 2. Section 3 describes the results, including the predictors, the prediction model, and its validation. The main conclusions and some necessary discussion are in section 4.

2. Datasets and methods

The monthly geopotential heights at 500 (Z500) and 200 (Z200) hPa, wind at 850 hPa, and surface data (horizontal resolution: 2.5° × 2.5°) from 1979 to 2016 were obtained from the National Centers for Environmental Prediction–National Center for Atmospheric Research reanalysis (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html; Kalnay et al. 1996). The Extended Reconstruction SST dataset with a horizontal resolution of 2° × 2° was supported by the National Oceanic and Atmospheric Administration (Smith et al. 2008). The SI concentrations were obtained from the Met Office Hadley Centre with a horizontal resolution of 1° × 1° (Rayner et al. 2003). The China ground observations were collected by the National Meteorological Information Center of China and are those that were used to reconstruct the climatic WHD_{NCP} data to which Yin et al. (2015) refer. Then, the DY of the WHD_{NCP} was calculated.

This interannual increment approach treated the WHD_{NCP} DY (WDY) as the predictand to utilize the characteristics of tropospheric biennial oscillation (TBO) in East Asia and the observed information from the previous year (Wang et al. 2000, 2012). Relative to the original variables, the variances of the predictors and predictand were both amplified—that is, the predictive signals were more significant. In addition, the long-term linear trend was removed when calculating the DY (Fan and Tian 2013). After adding the predicted WDY to the observed WHD_{NCP} from the year before, the decadal components were contained in the final prediction. The DY approach has been widely used in China and successfully predicted the summer rainfall in the middle and lower reaches of the Yangtze River and winter temperature in northern China in real-time predictions in 2016 (Fan et al. 2016).

The GAM approach has been used previously to evaluate the influence of local meteorological conditions on tropospheric ozone (Davis et al. 1998) and air quality (Pearce et al. 2011). When modeling, the MLR method fitted the original inputs (i.e., predictand and predictors) and output a linear model. In a different way, as a data-driven method, the GAM approach used a smoothing function s determined by the predictors x_i themselves to transform the expressions (Yee and Mitchell 1991). Furthermore, the dependent variable y in GAM could address different probability distributions by the link function g. With coefficients β_k, intercept β_0, and residual ε, the GAM equation can be described as follows:

\[ g(\hat{y}) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + \epsilon. \]

The models were trained from the normalized datasets from 1979 to 2013. During cross validation, the
root-mean-square error (RMSE), mean absolute error (MAE), linear correlation coefficient (LCC), and explained variance (EV) were calculated. For the independent tests of 2014 and 2015, the forecast biases (i.e., the predicted value minus the measurement) were other calculated parameters. The percentage of the same sign (PSS; same sign means the mathematical sign of the fitted and observed WDY was the same) was also assessed for the WDY and WHD_{NCP}. Because of limitations on the haze observations (Yin et al. 2017), the length of qualified WHD_{NCP} was only from 1979 to 2016, which resulted in a lack of independent tests. Thus, the recycling independent tests (RIT) were designed to fix this issue. In the RIT predictions, the SPM was trained by the samples from 1980 to an expiration year of training data, and the WHD_{NCP} anomalies from the next year to 2015 were independently predicted. The expiration year of training data moved forward from 2004 to 2014 so that there were 11 modeling processes and 66 independent predicted results. For example, the SPM trained by the data from 1980 to 2010 could produce independent prediction from 2011 to 2015.

3. Results

WHD_{NCP} exhibited interdecadal and interannual variation (Fig. 1). Before 1992, the WHD_{NCP} showed clear interannual variation, but the trend of WHD_{NCP} was not obvious. Since then, the trend became more significant than the interannual variation but decreased from 1993 to 2010 and increased after 2010. The change in the long-term trend enhanced the difficulty of seasonal prediction. One of the advantages of the DY approach was that the linear trend could be removed. After subtracting the previous year’s observation, the WDY displayed remarkable TBO features with no obvious trend, illustrating that the DY approach was suitable for its prediction. From 2011, the current year’s WHD_{NCP} was almost larger than the previous year. In particular, the WDY in 2013 was 17 days and was thus an outlier that was hard to be simulated.

As mentioned above, although studies on the impacts of climate anomalies on haze pollutions were still sparse and insufficient, Yin and Wang (2016b) tried to focus on the relationships rather than physical mechanisms and to develop statistical models of the WHD_{NCP} with seven predictors. The WHD_{NCP} was significantly correlated with the Pacific SST (Yin and Wang 2016a) and Arctic SI (Wang et al. 2015) in the preceding autumn, whose physical mechanisms were studied clearly and respectively. Significant correlations also existed in the DY fields (figure not shown). Thus, the means of the variables for the most significantly correlated areas were selected as predictors here. The first predictor \( x_1 \) was the area-averaged SST DY around the Alaska Gulf (36°–56°N, 130°–170°W), whose LCC with WDY was 0.47, above the 99% confidence level (Figs. 2a,b). The area-averaged SI-area DY of the Beaufort Sea (73°–78°N, 130°–165°W) was selected as the second predictor \( x_2 \), and its LCC with WDY was 0.37, above the 95% confidence level (Fig. 2c). Furthermore, the LCC between \( x_1 \) and \( x_2 \) was 0.06, indicating that they were relatively independent predictors.

The positive phases of the eastern Atlantic/western Russia, the western Pacific (Barnston and Livezey 1987), and the Eurasia (Wallace and Gutzler 1981) teleconnection patterns could influence the anomalous anticyclone over northern China and then lead to weak ventilation conditions that resulted in a larger number of WHD_{NCP} (Yin et al. 2017). The correlated DY atmospheric circulations were identified, as shown in Figs. 3 and 4. In Fig. 3a, teleconnections such as the eastern Atlantic/western Russia pattern, the western Pacific pattern, and the Eurasia pattern could be recognized clearly and influence WDY through modulating the local anticyclone and meteorological conditions over the NCP. The positive Z500 anomalies over the NCP and Japan were the key contributors that could confine the vertical dispersion of particles and block the activity of cold air. On the surface, the anomalous southeaster from the western Pacific to the NCP area was significant in that it not only reduced the surface wind speed but also transported more moisture to produce favorable conditions for more WHD_{NCP} than in the previous year (Fig. 3a). To reveal the mechanism, the LCC between the predictor \( x_1 \) and the circulations (i.e., Z500 and surface wind) was also calculated. The correlated circulations were similarly distributed as with WDY, such as positive Z500 anomalies and the southeaster on the surface (Fig. 3b). In East Asia, the Rossby wave train was always propagated along the EAWS, and wave activity showed a positive correlation with the intensity of

---

**Fig. 1.** WHD_{NCP} anomalies (black) from 1979 to 2015 and WDY (red) from 1980 to 2015 (units: days).
the EAJS (Yang et al. 2002). The positive SI-area DY around the Beaufort Sea (i.e., $x_2$) could stimulate positive and negative 200-hPa geopotential height ($Z_{200}$) anomalies in the north and south of the NCP, respectively. The induced pressure gradient would generate easterly anomalies, weakening the EAJS, which also existed at 850 hPa and indicated weak meridional circulation anomalies (Fig. 4). Under such circulation, the cold activities around NCP were obviously weaker than in the previous year, and more haze events occurred (Chen and Wang 2015). To summarize, the associated mechanism was that the correlated global teleconnections modulated the local circulation (e.g., the anticyclone anomalies over the NCP area) and then the meteorological conditions, such as weak surface wind speed, high relative humidity, and confined vertical motion, weaken the horizontal and vertical dispersion of aerosols (Yin et al. 2017).

Although the linear correlations between the predictors and predictand were significant and could be interpreted physically, note that nonlinearity also needed to be taken into account (Figs. 2b,c). Thus, the GAM approach was used to establish the SPM (SPM_{GAM}) of WDY. The SPM_{GAM} was as follows:

$$WDY \times 10 = 3.42 \times s(x_1) + 3.56 \times s(x_2).$$

The simulated and independently tested performance is shown in Fig. 5 and Table 1. The LCC between the observed and fitted WDY was 0.79, exceeding the 99.99% confidence level. The trend, especially the dramatically increasing trend after 2010, was captured well. Furthermore, this DY model using the GAM approach could simulate the TBO features and account for 62.4% of the total variance. In addition, the RMSE and MAE were 3.01 and 2.25 days, respectively. The fitted maximum also occurred in 2013 and was close to the observed value. The PSS was 75.7%. As independent prediction
tests, the forecast biases of 2014 and 2015 were 0.86 and 0.19 days, respectively, illustrating good performance. For comparison, the MLR method was used to build a linear SPM (SPMMLR) whose RMSE and MAE (i.e., 3.98 and 3.12 days, respectively) were larger than the SPMGAM. The expression for SPMMLR was

\[
\text{WDY} \times 10 = 4.47x_1 + 3.38x_2.
\]

Only 32.5% variance of WDY could be explained, and the obvious increasing trend and the extrema cannot be reflected (figure not shown). The biases of the independent tests were 0.07 and 2.12 days (Table 1). We speculate that the reason for these shortcomings was that the nonlinear relationship could not be included in the SPMMLR. In the scatterplots (Figs. 2b,c), a portion of the points were not close to the linear regression line, indicating that nonlinearity possibly existed. Meanwhile, the WDY in 2013 was an outlier that was difficult to fit to the linear model. During GAM modeling, the predictors were first transformed and then were “weight added” together to gain the predicted results. Different from the MLR, the smoothing terms were more significantly correlated with WDY (Fig. 6). The LCC between smoothed \( x_1 \) and measured WDY (i.e., 0.56) was more significant than before smoothing, and the predictor \( x_1 \)

![Fig. 4. The LCC between Z200 DY (shades)/850-hPa wind DY (arrows) in winter and \( x_2 \) from 1980 to 2013. The dots indicate that LCC (wind) exceeds the 90% confidence level, and the green box shows the NCP area.](http://journals.ametsoc.org/doi/pdf/10.1175/JAMC-D-17-0013.1)

![Fig. 5. The temporal variation of the observed (black line) and SPMGAM fitted (red line) WDY from 1980 to 2013. The observed and predicted WDY results for 2014 and 2015 are shown by the black squares and red open circles, respectively.](http://journals.ametsoc.org/doi/pdf/10.1175/JAMC-D-17-0013.1)
played a more important role from 1980 to 2010 than after 2011. The smoothed $x_2$ (LCC with measured WDY was 0.62) primarily contributed to simulating the maximum and the increasing trend after 2010. The advantages of the GAM approach to address the nonlinear relationship were significant and helpful to the predictions. As shown in Fig. 6, the first predictor seemed to be the primary contributor and significantly more important; therefore, another GAM model was rebuilt with only $x_1$ (SPMGAM2) to test this speculation. The RMSE (4.02), MAE (2.7), and EV (32.5%) of SPMGAM2 were all worse than those of SPMGAM (Table 1). Thus, among these three trained models, the SPMGAM showed the best performance.

In general, the final targets of seasonal prediction were the climatic anomalies. To achieve the predicted WHD_{NCP}, the observed WHD_{NCP} in the previous year was added to the predicted WDY. For example, the predicted WDY in 2015 and the observed WHD_{NCP} in 2014 were added together to calculate the predicted WHD_{NCP} in 2015. In the leave-one-out cross validation, the LCC between the observations and simulations was 0.83, passing the 99.99% confidence test and indicating that the long-term trend was introduced successfully. After detrending, the LCC was 0.80 and also exceeded the 99.99% confidence level, illustrating that the SPMGAM also reflected the annual variation well. The PSS of WHD_{NCP} was 91.7%, better than that of WDY, indicating the importance of the observation in the previous year (Fig. 7). In addition to 2010, other trend-turning years (i.e., 1989, 1995, and 2005) were identical. Although the minimum occurred 1 yr later, the maximum and the remarkable increasing trend after 2010 were captured successfully. Furthermore, in the leave-five-out cross validation (figure not shown), the LCC between the observed and predicted WHD_{NCP} (detrending) was 0.77 (0.57), passing the 99% confidence test. The PSS of WHD_{NCP} anomalies was 91.2%, showing good robustness.

More independent tests were executed in the RIT predictions to validate the performances. The WHD_{NCP} anomaly was independently predicted 11 times for 2015, 10 times for 2014, and so on. The PSS of WHD_{NCP} anomalies was 100%, showing good accuracy. The predicted values for each year did not vary much (Fig. 8), indicating good reliability and robustness. For example, when the SPM was trained by the samples only from 1980 to 2004, the predicted WHD_{NCP} anomalies for 2014 and 2015 were also close to the results predicted in Fig. 5 (i.e., the SPM was trained by the data from 1980 to 2013) and the measurements. It was notable that, although the mathematical signs of anomalies were captured correctly, the biases in 2010, 2011, and 2013 were larger than in the other years; the reason remains to be investigated. Using only two predictors, the SPM established here was appropriate for real-time seasonal climate prediction. During the real-time operation, the measurements from 1980 to the year before the forecast year could be used to build or update the model. The hindcasts for the period of 2005–15 were executed. As is shown in Fig. 9, the models performed well to forecast WHD_{NCP} anomalies, accounting for 85% of the total variance. The LCC between the predicted and measured anomalies (DY) of WHD_{NCP} is 0.92 (0.83), and the predicted anomalies were closely related to the observations (Fig. 9b). The PSS of WHD_{NCP} anomalies was also calculated (i.e., 100%) from 2005 to 2015.

### 4. Conclusions and discussion

In this study, an SPM of WHD_{NCP} was established using the GAM approach. To take advantage of the annual increment, the WDY was treated as the predictand, and the preceding-autumn SST DY around the Alaska Gulf and SI-area DY of the Beaufort Sea were selected as the predictors. In physical terms, these two predictors could influence the horizontal and vertical dispersion of particulate matter by modulating the global and local anomalous atmospheric circulations.
The SPM$_{GAM}$ performed better than the SPM$_{MLR}$. For SPM$_{GAM}$, the RMSE and EV were 3.01 days and 62.4%, respectively. The TBO features and the dramatically increasing trend after 2010 were captured successfully. The PSS was 75.7%, and the forecast biases of 2014 and 2015 were 0.86 and 0.19 days, respectively. After adding the predicted WDY (leave-one-out cross validation) to the observed information of the year before, the long-term trend was introduced and the annual variations were simulated. The PSS of the final prediction was 91.7%, which was better than that of WDY. Furthermore, the turning points of the trend, the extrema, and the remarkable increasing trend after 2010 were captured successfully. Furthermore, the PSS for RIT predicted anomalies was 100%, and the predicted values for each year were relatively stable and robust. In the real-time hindcasts for the period 2005–15, the LCC between the predicted and observed anomalies (DY) of WHD$_{NCP}$ passed 99% confidence level, and the PSS of WHD$_{NCP}$ anomalies was 100%.

Excluding human activities, the prediction model was trained with only two climatic predictors, which was suitable for real-time operations. In this study, we presumed that the pollutant emissions between the current and previous year were similar. Thus, when the WDY was calculated, the socioeconomic component of WHD$_{NCP}$ was eliminated (Yin and Wang 2016b), and the contribution of the meteorological factors was magnified. The LCC between WDY and the DY of total energy consumption was calculated and was not significant (figure not shown). Certainly, this hypothesis was reasonable but primitive, so human activities, as the fundamental driver, should be taken into account in the future, depending on the length and quality of data. Furthermore, the GAM approach addressed non-linearity implicitly by calling the “gam” library of the R software program (Kabacoff 2010). In this paper, the nonlinear relationships between the predictand and predictors were not analyzed sufficiently and physically. Future studies should pay close attention to this issue to improve the understanding of the variation of wintertime

**FIG. 7.** The temporal variation of the observed (black) and SPM$_{GAM}$ leave-one-out simulated (1980–2013) and independently predicted (2014 and 2015) (red) WHD$_{NCP}$ anomalies.

**FIG. 8.** The WHD$_{NCP}$ anomalies from RIT predictions (bars) and observations (dashed lines). The expiration year of training data means that the SPM was trained by the datasets from 1980 to this year, and the WHD$_{NCP}$ anomalies from the next year to 2015 were independently predicted. For example, the results plotted with yellow bars (existing from 2011) were independently predicted by the SPM trained by the data from 1980 to 2010.
haze pollution. Furthermore, when there were multiple predictors, the advantage of GAM was not significant (Yin and Wang 2016b); that is, the nonlinear processes were possibly counteracted by the internal interactions among the multiple predictors. In contrast, the GAM approach performed better when the number of predictors was relatively small.

Acknowledgments. This research was supported by National Key Research and Development Plan (2016YFA0600703), National Natural Science Foundation of China (41421004), CAS–PKU Partnership Program, Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (20172007), and KLME Open Foundation (KLME1607). The reanalysis data were obtained online by National Key Research and Development Plan (2016YFA0600703), and KLME Open Foundation.

REFERENCES


