JOURNAL OF ADVANCED TRANSPORTATION
J. Adv. Transp. 2016; **50**:1754–1769
Published online 18 November 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/atr.1427

Integrated model for traffic flow forecasting under rainy conditions

Han Qiu[1,3], Ruimin Li[1]* and Hao Liu[2]

[1]*Department of Civil Engineering, Tsinghua University, Beijing, China*
[2]*Beijing Transportation Information Center, Beijing, China*
[3]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA USA*

SUMMARY

Inclement weather, such as heavy rain, significantly affects road traffic flow operation, which may cause severe congestion in road networks in cities. This study investigates the effect of inclement weather, such as rain events, on traffic flow and proposes an integrated model for traffic flow parameter forecasting during such events. First, an analysis of historical observation data indicates that the forecasting error of traffic flow volume has a significant linear correlation with mean precipitation, and thus, forecasting accuracy can be considerably improved by applying this linear correlation to correct forecasting values. An integrated online precipitation-correction model was proposed for traffic flow volume forecasting based on these findings. We preprocessed precipitation data transformation and used outlier detection techniques to improve the efficiency of the model. Finally, an integrated forecasting model was designed through data fusion methods based on the four basic forecasting models and the proposed online precipitation-correction model. Results of the model validation with the field data set show that the designed model is better than the other models in terms of overall accuracy throughout the day and under precipitation. However, the designed model is not always ideal under heavy rain conditions. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS: traffic forecasting; online precipitation-correction model; two-level integrated model; data fusion

1. INTRODUCTION

Inclement weather, such as heavy rain, snow, and fog, is a major cause of serious traffic accidents and severe traffic congestion in urban areas or freeway networks. Thus, traffic is sometimes congested throughout the road networks of an entire city. According to [1], such weather is responsible for approximately 15% of traffic congestion on roadways in the United States. Severe weather conditions can also lead to traffic accidents that generate thousands of casualties each year. Moreover, inclement weather significantly affects traffic flow conditions, such as road capacity, traffic flow volume, and traveling speed [2]. Thus, investigations on improving the accuracy of traffic condition forecasting under inclement weather should be conducted to provide reliable travel information to drivers and to minimize traveling risk during inclement weather conditions.

Traffic forecasting has been extensively researched over the past three decades [3–5]. Inclement weather is another popular topic that has seen many relevant breakthroughs [6, 7]. Nonetheless, only a few researchers have attempted to investigate traffic condition forecasting methods under inclement weather conditions because of data scarcity or other reasons. At present, the online collection of integrated traffic flow data and weather information is a viable approach to address this issue because of the large-scale deployment of various traffic flow and weather detectors. These high-quality data

*Correspondence to: Ruimin Li, Department of Civil Engineering, Tsinghua University, Beijing 10084, China. E-mail: lrmin@tsinghua.edu.cn

provide researchers with an opportunity to initiate research on traffic forecasting under inclement weather conditions.

In the current study, we propose a two-level integrated model for traffic flow forecasting under inclement weather conditions. At present, we consider only rainfall variables that are restricted by the detection parameters of weather detectors. In the proposed model, four basic models are initially adopted to obtain four basic forecasting values. These results are then refined based on the fitted regression relationships between forecasting error and precipitation value for each basic model. We also introduce a correction model to reflect the influence of precipitation. Finally, data fusion methods are applied to combine the refined results into a final forecasting value. The subsequent validation performed using historical data indicates that the proposed model outperforms the single models in terms of forecasting accuracy under rainy conditions.

The remainder of this paper is organized as follows. Section 2 reviews previous studies conducted on traffic forecasting models under inclement weather conditions. Section 3 introduces the overall model structure and discusses several concerns regarding its construction. Section 4 describes the stepwise development of the proposed model, including the selection of the four basic forecasting models, the assumption of precipitation correction, the formulation of an online implementation strategy, and the creation of a method for fusing values from individual models. Section 5 briefly introduces the research data and then highlights the model validation process and the performance result. Section 6 presents the conclusion drawn and suggests future research directions.

2. LITERATURE REVIEW

Traffic forecasting is the basis of real-time traveler information systems and other intelligent transportation systems. Traffic forecasting methods have been researched extensively in the past decades [3–5]. At present, the forecasting methods used in this field can be categorized into two classes: single-step forecasting models, which forecast for only one time slice at a time, and multiple-step forecasting models, which provide forecasting values for an entire continuous period at a time. Single-step forecasting models [3] mainly include the autoregressive integrated moving average (ARIMA) [8], local linear regression (LLR) [9], artificial neural network (ANN) [10–12], and support vector regression (SVR) [13–18] models. Multiple-step forecasting models mainly include LLR, SVR, nonparametric regression [19], and spectral analysis [20]. Another classification method for forecasting models is based on the different mechanisms of models. Some models are mainly based on statistical concepts, whereas others are based on machine learning [21]. Among the aforementioned models, ARIMA, LLR, and nonparametric regression are statistical forecasting models, whereas ANN and SVR are machine learning-based forecasting models.

As a recent topic in the field of traffic forecasting, integrated forecasting models, which combine forecasting results from various single forecasting models through data fusion methods, have recently gained considerable attention [21–23]. Some works [24–26] found that integrated models were not significantly superior to single models, and other studies [21, 27] confirmed this finding in the field of traffic forecasting. However, some studies [28–31] claimed that integrated models were more stable and robust than single models toward detector failures. Tselentis, Vlahogianni, and Karlaftis [21] determined that when input data were partly missing or contaminated, integrated models could generate accurate results at low risk in the event of forecasting failure.

Various inclement weather conditions, such as rain or snow, will influence traffic flow characteristics [6, 7]. Recently many studies on this subject have focused on the influence of weather on traffic demand or supply parameters, such as traffic flow volume and speed. Although the results are slightly inconsistent in terms of accuracy, all these studies indicate that inclement weather can significantly influence traffic flow volume and traveling speed.

In the past years, studies have been conducted to investigate traffic forecasting methods under inclement weather conditions. These methods may be grouped into two main classes. Under the first class, several studies have implicitly included weather information in their models and have investigated the observed difference as a result. For example, Min and Wynter [32] constructed various scenarios in which a different set of parameters could be applied to their multivariate, spatiotemporal

DOI: 10.1002/atr

autoregressive model. Weather information is an important factor in determining which data entry belongs to which scenario. Similarly, Qiao, Haghani, and Hamedi [33] classified data into different categories and then applied the *k*-nearest neighbor algorithm to predict each category. Both traffic condition and weather information were considered during the classification process. By contrast, Fei, Lu, and Liu [34] regarded weather influence as an unquantifiable factor and assumed that such influence would implicitly appear as random traffic flow noise. They designed an adaptive control framework that enabled the system to learn and adjust automatically to such noise.

Under the second class, weather information has also been explicitly included as part of the models developed in some works. Nookala [9] proposed an LLR model with linear correction for weather information. He demonstrated that a model with linear correction could limit forecasting errors by comparing the forecasting results of this model with those of the original LLR model.

Tsirigotis, Vlahogianni, and Karlaftis [35] designed the ARIMA model with eXternal input (ARIMAX) for inclement weather conditions and compared this model with the traditional ARIMA model. On average, the forecasting accuracy of ARIMAX is better than that of the traditional ARIMA model; however, the improvement is not uniform. That is, the improvement is significant in some locations, whereas forecasting accuracy decreases substantially in other areas.

Florio and Mussone [36] first attempted to include neurons in the neural network model for weather information. However, they did not examine the efficiency of this inclusion. Li and Rose [37] applied the same inclusion method but did not regard weather information as a control variable nor investigate the difference. Subsequently, Huang and Ran [38] adapted the same method and concluded through model validation that the use of weather information increased forecasting accuracy. Li and Chen [39] recently adopted different combinations of rainfall variables as inputs in their neural network formulation. They also attempted to select the best combination by analyzing the differences among the results; however, they did not consider comparing their formulation with the model that was formulated without any rainfall variable.

Some of the aforementioned works have improved forecasting accuracy by including weather information in their models. However, these works neither tested their models on sufficiently large data sets nor conducted rigorous statistical tests to examine the corresponding performance. Consequently, explanation power was reduced; whether "improvement" was an actual enhancement or a variation caused by randomness could not be determined.

In the current study, we successfully investigate the correlation between precipitation and traffic flow volume based on a larger data set with certain statistical significance. The obtained knowledge is then applied to develop a forecasting model that can maintain its computing speed with reduced forecasting errors.

3. MODEL STRUCTURE

This study proposed a two-level integrated model that initially used several basic forecasting models (i.e., first-level models) to obtain multiple forecasting values for traffic flow. The influence of precipitation was determined by using a correction model (i.e., a second-level model) with each forecasting value. Several fusion methods were adopted to obtain the final forecasting value $\hat{f}_{n+1}$. The model structure is illustrated in Figure 1.

In Figure 1, f_n is the true detection value of traffic flow volume for time interval n, w_n is the detected precipitation value for time interval n, $\hat{f}_{n+1,m}$ is the traffic flow volume forecasted using method m for time interval $n+1$, and $\widehat{f'}_{n+1,m}$ is the corrected value of $\hat{f}_{n+1,m}$.

This correction model exhibits several advantages over models integrated with weather information directly. First, given that precipitation data are sporadic, the integrated model may become overfitted and the overall forecasting accuracy may be reduced. The correction model is at least as accurate as these models without weather information. Second, the simple structure of the correction model allows a researcher to perform statistical analysis and determine the significance of the effect of precipitation. Third, the correction model can be easily implemented and extended to include other basic forecasting models, thereby making this model suitable for various engineering applications.

DOI: 10.1002/atr

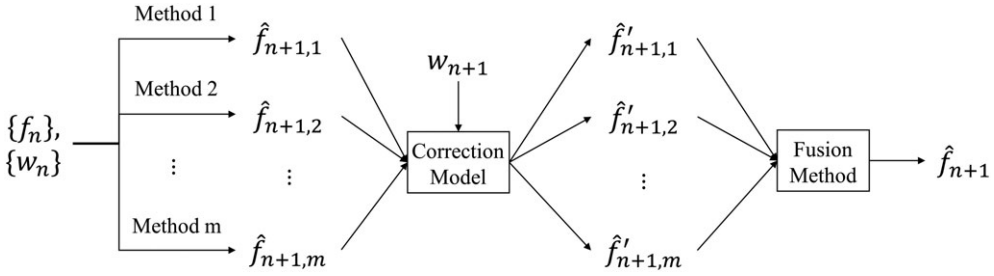


Figure 1. Model structure.

4. MODEL DEVELOPMENT

4.1. Individual basic models

For the proposed two-level model, several basic models should initially be selected at the first level. Any effective forecasting model can be used as basic models, such as the models mentioned in literature review. For simplicity, the following four methods were used in this study: LLR [9], neural network with external input (NNX) [36–39], principal component analysis and SVR (PCA–SVR) [40], and PCA and spectral analysis (PCA–spec) [19]. Appendix briefly introduces these models and their parameter selections. These models, which include traditional statistical models and machine learning models, can capture both the local trends with single-step forecasting models and global patterns with multiple-step forecasting models.

For simplicity, these individual models are calibrated off-line and are invariant within the overall model. The calibrated parameters are also introduced in the appendix for reference. In practice, the online calibration of these models will converge and can be regarded as invariant after a certain period; therefore, we can consider this model as a favorable approximation.

4.2. Correction method

The next step in designing the proposed two-level model is formulating assumptions for secondary correction. In this study, the following correction assumption was used:

$$\begin{cases} \left(\hat{f}_n - f_n\right)/\hat{f}_n = cu_n + \delta_n, \delta_n \tilde{} N\left(0, \sigma^2\right), i.i.d. \\ u_n = 1 - e^{-\lambda w_n} \end{cases} \tag{1}$$

where $\hat{f}_n$ is the forecasted value of traffic flow volume, δ_n is a random error, and c and λ are the model parameters. Moreover, we assume that λ can be determined solely by sequence $\{w_n\}$.

The preceding assumption was selected for several reasons. First, given that the correction must be applied in a relative manner, we choose the percentage difference $\left(\hat{f}_n - f_n\right)/\hat{f}_n$ instead of the difference $\hat{f}_n - f_n$ as the dependent variable. Second, to fit a simple linear regression model with an intuitive bound into the dependent variable (± 1), precipitation w_n must be transformed to bind the independent variable. The logistic transformation $e^{-\lambda w_n}$ is among the simple choices that we have considered. As another advantage of using this nonlinear transformation, the extreme values of w_n are smoothened to alleviate the heteroscedasticity of the estimated error term $\hat{\delta}_n$. We assume that the positive precipitation value w_n^+ follows exponential distribution $Exp(\lambda)$ and uses the average as an estimate of $1/\lambda$. Under this assumption, the logistic transformation will give u_n a uniform distribution on [0,1].

DOI: 10.1002/atr

Assumption (1) provides the following simple means of correction:

$$\hat{f}_n^{'} = (1 - \hat{c}\, u_n)\hat{f}_n\,, \tag{2}$$

in which $\hat{c}$ is the least square estimate of c.

Given that $\hat{c}$ always changes in actual situations, the calculation of $\hat{c}$ must be updated online in the proposed secondary model. This process only becomes easy when dealing with the least square estimate $\hat{c}$, but the validity of this approach depends on the assumption of error term δ_n, which may be inaccurate. In this study, the data set that includes δ_n may follow a heavy tail distribution instead of a normal distribution. Thus, we considered excluding potential outliers in the data set to keep using least square estimate $\hat{c}$. Two outlier detection techniques were adopted [41]. The first technique is a studentized residual, i.e.,

$$t_i = \frac{r_i}{s(i)\sqrt{1 - u_i^2/\sum_j u_j^2}} \sim t(N-2), \tag{3}$$

in which

$$s^2(i) = \frac{\sum_j r_j^2 - \frac{r_i^2}{1-u_i^2/\sum_j u_j^2}}{N-2} \tag{4}$$

and r_i is the least square fit residual for data point u_i. N is the size of the entire data set.

We use Cook's d-estimator as the second technique, which can be expressed as follows:

$$D_i = \frac{u_i^2\left(u_i^2/\sum_j u_j^2\right)}{s^2\left(1 - u_i^2/\sum_j u_j^2\right)^2}, \tag{5}$$

in which

$$s^2 = \frac{\sum_j r_j^2}{N-1}. \tag{6}$$

From an engineering perspective, outlier detection at each time slice is not time-efficient. Therefore, we adopted a two-stage approach. At midnight (24:00), all data included in the aforementioned outlier detection were used to obtain the initial data set and calculate the $\hat{c}$ value for the day (with complexity $O(N^2)$). Subsequently, $\hat{c}$ was directly updated with the standard form of the least square estimator (with complexity $O(N)$) within each day for each new piece of data entered.

4.3. Fusion method

One cannot easily determine if a basic method will outperform others before experimenting with the data. In general, each method has its advantages in dealing with a specific data pattern. Model fusion and automatic selection techniques can be used to combine these strengths and improve accuracy.

To combine the forecasting values $\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_m$ into a final $\hat{f}$, the following basic model is considered:

$$\hat{f} = \sum_i p_i \hat{f}_i\,, \tag{7}$$

in which parameter $p_i, i = 1, \cdots, m$ is non-negative and normalized as

DOI: 10.1002/atr

$$\sum p_i = 1. \tag{8}$$

At this point, we consider estimating the $\overrightarrow{p_n}$ of weight $\overrightarrow{p}$ for each time step n based on the estimates obtained from different methods $f_{n,i}(i=1,\cdots,m)$ and the true value f_n.

Let

$$\Delta_{n,i} = f_{n,i} - f_n,\ i = 1,2,\cdots,m, \tag{9}$$

where $\Delta_{n,i}$ is the forecasted error. If all $\Delta_{n,i}$ values have the same sign, then we simply set the $p_{n,i}$ that corresponds to the smallest $|\Delta_{n,i}|$ as 1 and all the others as 0. Otherwise, let

$$p_{n,i} = \frac{|\Delta_{n,i}|}{\sum_i |\Delta_{n,i}|},\ i = 1,2,\cdots,m. \tag{10}$$

Finally, p_i is estimated based on the average of all $p_{n,i}$ values as follows:

$$p_i = \frac{1}{N}\sum_n p_{n,i}. \tag{11}$$

We assume that all the estimates, despite not having the same accuracy, are valid. However, if some models, such as model i, fail in the process, then the overall model may be biased because p_i is affected by historical estimate $p_{n,i}$ and does not instantly drop to 0. In several extreme cases, the integrated model performs worse than the best single basic model. Appendix presents an example. Thus, a strategy for detecting and discarding models that are prone to failure must be devised to avoid contamination.

We propose a simple detection method based on the logic that if a single basic model performs significantly worse than the best one over a recent short period, then this model "fails." We use the mean absolute percentage error (MAPE) to evaluate the performance of each model and impose performance threshold Δ to determine whether the performance of a certain model is "significantly worse." MAPE was calculated as follows:

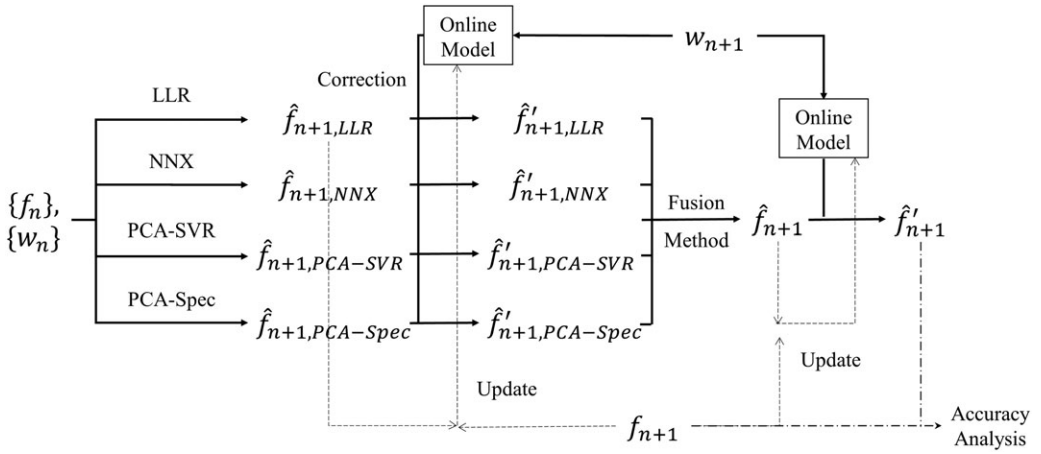


Figure 2. Specific model structure.

DOI: 10.1002/atr

Table I. Relationship between weather detectors and traffic flow detectors.

Rain gauge location name	Ancient observatory			Longtan Lake	Xiannongtan	Daguanyuan	Guanyuan
Corresponding traffic flow detector code	2010	2011	2012	2023	2030	2033	2052
Rain gauge location name	Lizeqiao			Shibalidian	Olympic Sports Center		Chaoyang Park
Corresponding traffic flow detector code	3034	3035	4004	4005	4050	4051	5062

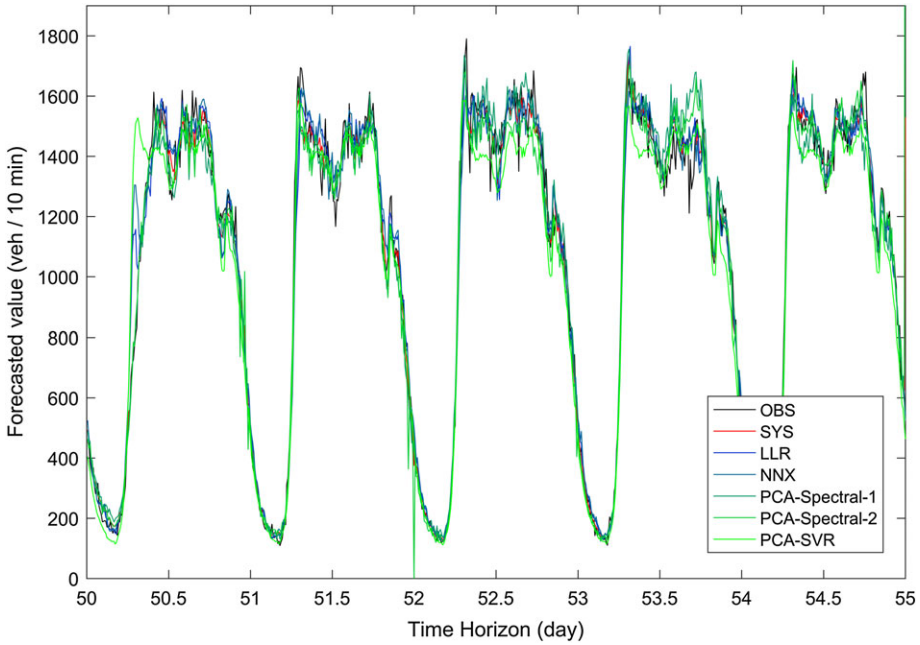


Figure 3. Traffic flow forecasting (corrected for weather), day 51–55.

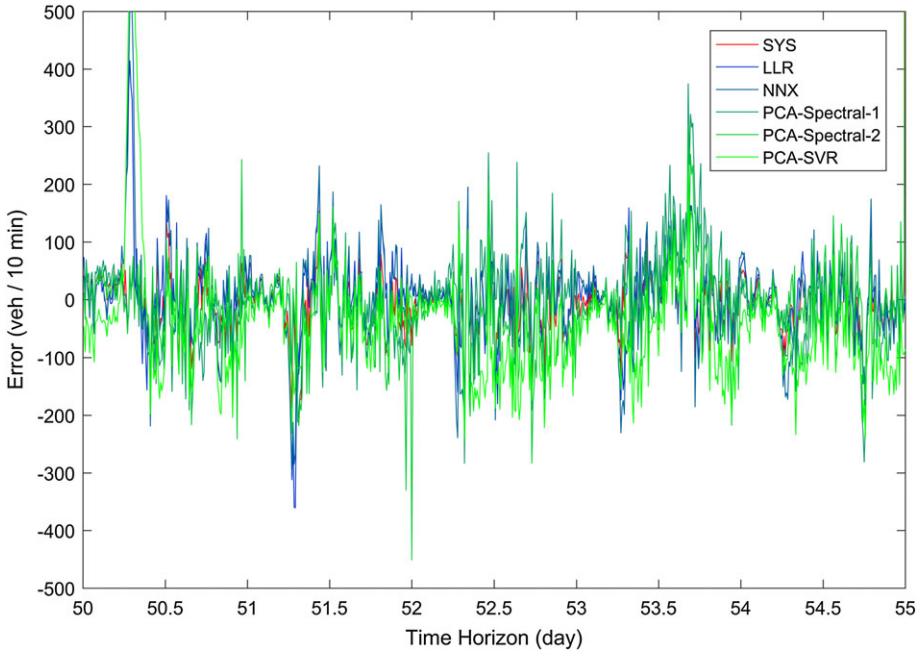


Figure 4. Forecasting error (corrected for weather), day 51–55.

DOI: 10.1002/atr

Table II. Model validation result.

Traffic road code	Model	MAPE (%)								
		10 minutes ahead			1 hour ahead			1 day ahead		
		ALL	$w>0$	$w>1$	ALL	$w>0$	$w>1$	ALL	$w>0$	$w>1$
2010	OPT_UC	4.93	5.62	—	6.64	8.28	—	6.70	10.67	—
	OPT_C	4.90	5.33	—	6.45	6.37	—	6.45	6.37	—
	SYS_UC	4.82	5.47	—	5.64	6.67	—	5.62	6.83	—
	SYS_C	4.82	5.41	—	5.64	6.67	—	5.61	6.68	—
2011	OPT_UC	6.78	5.92	—	8.60	7.49	—	8.60	7.49	—
	OPT_C	6.77	5.49	—	8.53	6.44	—	8.53	6.44	—
	SYS_UC	6.71	6.00	—	7.84	6.77	—	7.78	6.85	—
	SYS_C	6.71	6.00	—	7.84	6.75	—	7.79	6.88	—
2012	OPT_UC	7.90	8.70	—	11.63	18.64	—	11.85	25.43	—
	OPT_C	7.83	6.15	—	11.23	9.74	—	11.27	9.74	—
	SYS_UC	7.84	5.59	—	9.73	9.99	—	9.93	10.77	—
	SYS_C	7.84	5.53	—	9.74	10.42	—	9.92	10.40	—
2023	OPT_UC	3.95	4.49	11.02	5.56	6.61	11.57	5.56	6.61	11.22
	OPT_C	3.94	4.27	6.89	5.49	6.44	10.59	5.49	6.44	10.59
	SYS_UC	3.87	4.36	10.35	4.48	5.81	7.12	4.60	6.03	9.21
	SYS_C	3.87	4.26	6.47	4.48	5.82	8.70	4.60	5.96	10.48
2030	OPT_UC	3.73	3.86	4.17	5.09	4.23	6.15	5.18	4.23	6.15
	OPT_C	3.72	3.72	4.76	5.06	4.25	4.94	5.13	4.25	4.94
	SYS_UC	3.47	3.64	3.59	4.04	4.36	5.61	4.11	4.43	4.84
	SYS_C	3.47	3.64	3.59	4.03	4.23	5.04	4.10	4.36	4.74
2033	OPT_UC	3.46	3.39	3.33	4.53	3.41	3.33	4.60	3.41	3.33
	OPT_C	3.44	3.32	4.56	4.51	3.38	4.56	4.57	3.38	4.56
	SYS_UC	3.29	3.17	4.16	3.69	3.61	4.93	3.64	3.72	5.31
	SYS_C	3.29	3.12	3.66	3.69	3.51	4.93	3.63	3.54	4.11
2052	OPT_UC	5.74	5.77	7.59	7.90	8.54	6.89	7.90	8.54	8.46
	OPT_C	5.73	5.70	7.30	7.64	7.78	7.30	7.64	7.78	7.30
	SYS_UC	5.70	5.84	8.45	7.06	7.19	8.97	7.04	7.63	9.99
	SYS_C	5.70	5.84	8.45	7.06	7.09	8.97	7.03	7.51	9.99
3034	OPT_UC	4.39	5.01	4.84	6.00	7.70	2.99	6.00	8.97	2.23
	OPT_C	4.39	4.93	4.84	5.99	7.56	4.14	5.99	8.70	2.23
	SYS_UC	4.24	5.09	9.22	5.06	7.56	6.83	5.18	7.46	4.65
	SYS_C	4.24	4.91	9.22	5.04	7.00	6.36	5.16	6.86	5.10
3035	OPT_UC	4.72	6.50	5.44	5.60	8.26	3.92	5.60	9.00	3.17
	OPT_C	4.71	6.21	3.59	5.60	7.96	8.82	5.60	8.95	9.21
	SYS_UC	4.60	6.41	6.27	5.07	8.10	13.23	5.06	8.73	13.33
	SYS_C	4.60	6.33	6.27	5.06	7.75	11.26	5.06	8.73	13.33
4004	OPT_UC	3.75	5.25	7.54	6.07	8.74	9.91	6.85	9.70	8.90
	OPT_C	3.74	5.04	7.49	6.00	8.13	7.82	6.73	8.26	7.82
	SYS_UC	3.66	5.14	7.97	4.97	7.24	7.91	5.27	7.95	6.59
	SYS_C	3.66	5.07	7.97	4.97	7.32	7.91	5.23	7.37	6.63
4005	OPT_UC	3.74	4.23	5.00	4.91	5.76	5.00	4.91	5.76	5.00
	OPT_C	3.74	4.31	4.69	4.82	5.41	4.69	4.82	5.41	4.69
	SYS_UC	3.50	4.08	4.95	4.06	4.68	4.66	4.14	5.21	4.83
	SYS_C	3.49	3.91	4.90	4.04	4.40	4.63	4.11	4.79	5.32
4050	OPT_UC	12.12	3.96	1.37	18.38	3.96	1.37	18.38	3.96	1.37
	OPT_C	12.10	3.91	1.37	18.37	3.91	1.37	18.37	3.91	1.37
	SYS_UC	8.50	3.53	1.97	14.29	4.25	1.68	14.53	3.96	1.66
	SYS_C	8.50	3.53	1.97	14.29	4.25	1.68	14.53	3.96	1.66
4051	OPT_UC	6.72	6.08	1.54	8.00	6.08	1.54	8.00	6.08	1.54
	OPT_C	6.72	6.01	1.32	8.00	6.01	1.32	8.00	6.01	1.32
	SYS_UC	3.72	2.99	1.95	5.02	3.10	1.97	6.85	6.42	1.97
	SYS_C	3.72	2.97	1.75	5.02	3.09	1.86	6.85	6.34	1.97
5062	OPT_UC	5.13	5.60	0.92	7.64	9.77	1.24	9.17	10.39	1.63
	OPT_C	5.13	5.60	0.92	7.62	9.29	0.97	9.01	9.54	0.97
	SYS_UC	5.13	5.67	1.21	6.81	8.74	2.31	7.32	9.14	7.60
	SYS_C	5.13	5.67	1.21	6.80	8.66	2.31	7.31	9.05	7.60

Notes:
— indicates that no heavy rain record exists for this road site.
OPT represents the optimal MAPE from the best single basic model among the four chosen models. The value for ALL or "$w>0$" is not necessarily taken from the same basic model.
SYS represents the MAPE given by our two-level integrated model.
UC denotes the MAPE result without the correction model, and C denotes the MAPE result with the correction model.
IMP represents the relative improvement of our model on the optimal single basic model.
"ALL" considers the result from all data points and represents the efficiency in the overall scenario.
"$w>0$" considers the results from data with $w>0$ and represents the efficiency in a rainy day scenario (including the scenario where $w>1$).
"$w>1$" considers the results from data with $w>1$ and represents the efficiency in a heavy rain scenario.

DOI: 10.1002/atr

$$MAPE = \frac{1}{N}\sum_{n=1}^{N} |\hat{f}_n - f_n|/f_n. \tag{12}$$

The detection method is detailed as follows.

At the beginning of each fusion step, all the temporary MAPEs (in the past M time steps) of each single basic model are compared. If model i has a temporary MAPE that is larger than the best temporary MAPE plus performance threshold Δ, then this model is disregarded in this fusion step. The corresponding historical estimate $p_{n,i}$ in $\overrightarrow{p_n}$ is set to 0. For the remaining single basic models, the fusion task (Equations (10) and (11)) is performed.

4.4. Summary

For clarity, the aforementioned steps are summarized in Figure 2. In this figure, $f, \hat{f}$, and $\hat{f}^{'}$ represent the true detection value, the forecasted value, and the corrected forecasted value, respectively.

5. MODEL VALIDATION

In this study, we validate our model using historical recordings of traffic flow and precipitation detection data from several ring-shaped expressways in Beijing. The data were collected from June 1, 2013 to July 31, 2013, which represented the rainy season in Beijing. To generate a specific "site–road" pair, our data consist of a time series of traffic flow volume from a traffic flow detector (corresponding to a road) and a time series of precipitation from a rain gauge detector (corresponding to a site). Each traffic flow data set recorded the traffic flow volume of the road section every 2 minutes, whereas each precipitation data set recorded the precipitation amount on-site every 5 minutes (measured by a rain gauge, and the minimum resolution was 0.1 mm). All data were processed in 10-minute intervals to unify the forecasting period. The proposed model was validated using the data from the 14 site–road pairs shown in Table I.

In the validation, data for the first month (4320 time steps) were used as the basic training set, and the training of the online correction model from the 4321st time step was initiated. The estimate of the final 3600 time steps was applied in the accuracy analysis. MAPE was used as the major statistic for describing forecasting accuracy.

The value of f_n significantly influences MAPE, and a small f_n contributes substantially to this measure. Considering the high traffic flow volume, the forecasting performance of the model during daytime may be concealed by MAPE at night under the same random variation. Given that our major concern is forecasting performance during daytime, we focus only on the MAPE from 6:00 to 22:00. The left side of the correction model, which is shown as Equation (1), adopts a form that is similar to that of MAPE. Thus, we consider only the data entered from 6:00 to 22:00 in the training process for the correction model to avoid the unnecessary influence of nighttime estimations.

Figures 3 and 4 show the forecasting value and error on days 51 to 55 for the data from detector 2030. Our forecasting system achieves favorable accuracy.

To determine the performance of the proposed model, we analyze the error statistics according to the following classifications: (i) the different forecasting periods are the traffic flow volumes for 10 minutes ahead, 1 hour ahead, and 1 day ahead; (ii) The different weather conditions are the traffic flow volumes for all weather conditions (ALL), for all rainy conditions ($w > 0$), and for heavy rain conditions ($w > 1$).

The results are listed in Table II.

The following conclusions can be drawn from the results.

DOI: 10.1002/atr

I) The proposed two-level integrated forecasting model performed significantly better than the best basic single basic model in a rainy scenario. The improvement is higher than 10% for most site–road pairs and forecasting steps ahead.
II) The comparison with each single basic model indicates that the proposed two-level integrated forecasting model may not be influenced by the length of the forecasting steps ahead. Thus, our model can efficiently improve long-term forecasting accuracy.
III) Under a heavy rain scenario, the proposed two-level integrated forecasting model may generate significant forecasting errors. Thus, the transformation of w_n inefficiently explains the situation with a large w_n. In this study, the data set obtained under heavy rain is inadequate, which may have destabilized the forecasting result. Therefore, the influence of heavy rain on traffic flow condition should be studied with a large data set.

6. CONCLUSION

In this study, a two-level integrated traffic flow volume forecasting model was proposed for rainy conditions. This model is simple to calculate and manage because it uses four basic models as the first-level model. Moreover, a quasi-linear relation assumption is applied to construct the second-level correction model. Improving the design of the online updating process and the fusion method also ensures the feasibility of the model in real-world applications. The validation of the model with the given historical data indicates that the proposed model is more efficient than other basic models. This model does not only significantly improve forecasting accuracy during rainy scenarios but also improve long-term forecasting accuracy. Thus, this model is competitive in the future design of traffic forecasting systems. However, the proposed model does not perform better than the best single forecasting model for over half of the heavy rain scenarios. Thus, this issue must be investigated in detail with extensive data support in the future.

In this study, only the interaction between precipitation and traffic flow volume in traffic flow volume forecasting was examined. In the future, the performance of this two-level integrated model with regard to the interaction among other meteorological factors, such as wind speed and snowfall, and other traffic flow parameters, such as speed and occupancy, must be investigated if appropriate data can be detected. This two-level model can be expanded into a complex multilevel model that may simultaneously consider the influences of multiple factors.

ACKNOWLEDGEMENTS

The authors are grateful to the National Natural Science Foundation of China (71361130015), the Beijing Natural Science Foundation (8162024) and the Collaborative Innovation Center for Capital World City's Smooth Traffic Construction for sponsoring and supporting this study.

REFERENCES

1. Systematics C. Traffic congestion and reliability: linking solutions to problems: executive summary. Federal Highway Administration, 2004.
2. Lam WH, Tam ML, Cao X, *et al*. Modeling the effects of rainfall intensity on traffic speed, flow, and density relationships for urban roads. *Journal of Transportation Engineering* 2013; **139**(7): 758–770. DOI:10.1061/(ASCE)TE.1943-5436.0000544.
3. Vlahogianni EI, Karlaftis MG Golias JC. Short-term traffic forecasting: where we are and where we're going. *Transportation Research Part C: Emerging Technologies* 2014; **43**: 3–19. DOI:10.1016/j.trc.2014.01.005.
4. Mori U, Mendiburu A, Álvarez M, *et al*. A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science* 2015; **11**(2): 119–157. DOI:10.1080/23249935.2014.932469.
5. Vlahogianni EI, Golias JC Karlaftis MG. Short-term traffic forecasting: overview of objectives and methods. *Transport Reviews* 2004; **24**(5): 533–557. DOI:10.1080/0144164042000195072.
6. Rakha H, Farzaneh M, Arafeh M, et al. *Empirical Studies on Traffic Flow in Inclement Weather*: Virginia Tech Transportation Institute, Blacksburg, VA 2007.

7. Rakha HA, Zohdy I, Park S, *et al*. Microscopic analysis of traffic flow in inclement weather—part 2, 2010.
8. Williams BM, Hoel LA. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. *Journal of Transportation Engineering* 2003; **129**(6): 664–672. DOI:10.1061/(ASCE)0733-947X(2003)129:6(664).
9. Nookala LS. *Weather Impact on Traffic Conditions and Travel Time Prediction*University of Minnesota: Duluth, 2006.
10. Vlahogianni EI, Karlaftis MG Golias JC. Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. *Transportation Research Part C: Emerging Technologies* 2005; **13**(3): 211–234. DOI:10.1016/j.trc.2005.04.007.
11. Yin H, Wong SC, Xu J, et al. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies* 2002; **10**(2): 85–98. DOI:10.1016/S0968-090X(01)00004-3.
12. Zheng W, Lee D-H Shi Q. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. *Journal of Transportation Engineering* 2006; **132**(2): 114–121. DOI:10.1061/(ASCE)0733-947X(2006)132:2(114).
13. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and Computing* 2004; **14**(3): 199–222. DOI:10.1023/B:STCO.0000035301.49549.88.
14. Wu C-H, Ho J-M Lee D-T. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems* 2004; **5**(4): 276–281. DOI:10.1109/TITS.2004.837813.
15. Castro-Neto M, Jeong Y-S, Jeong M-K, et al. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications* 2009; **36**(3): 6164–6173. DOI:10.1016/j.eswa.2008.07.069.
16. Ma J, Theiler J Perkins S. Accurate on-line support vector regression. *Neural Computation* 2003; **15**(11): 2683–2703. DOI:10.1162/089976603322385117.
17. Yang H, Huang K, King I, et al. Localized support vector regression for time series prediction. *Neurocomputing* 2009; **72**(10): 2659–2669. DOI:10.1016/j.neucom.2008.09.014.
18. Lippi M, Bertini M Frasconi P. Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems* 2013; **14**(2): 871–882. DOI:10.1109/TITS.2013.2247040.
19. Smith BL, Williams BM Oswald RK. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies* 2002; **10**(4): 303–321. DOI:10.1016/S0968-090X(02)00009-8.
20. Tchrakian TT, Basu B O'mahony M. Real-time traffic flow forecasting using spectral analysis. *IEEE Transactions on Intelligent Transportation Systems* 2012; **13**(2): 519–526. DOI:10.1109/TITS.2011.2174634.
21. Tselentis DI, Vlahogianni EI Karlaftis MG. Improving short-term traffic forecasts: to combine models or not to combine? *IET Intelligent Transport Systems* 2014; **9**(2): 193–201. DOI:10.1049/iet-its.2013.0191.
22. Jiang X, Zhang L Chen X. Short-term forecasting of high-speed rail demand: a hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China. *Transportation Research Part C* 2014; **44**: 110–127. DOI:10.1016/j.trc.2014.03.016.
23. Li L, Chen X Zhang L. Multimodel ensemble for freeway traffic state estimations. *IEEE Transactions on Intelligent Transportation Systems* 2014; **15**(3): 1323–1336. DOI:10.1109/TITS.2014.2299542.
24. Rapach DE, Strauss JK. Bagging or combining (or both)? An analysis based on forecasting US employment growth. *Econometric Reviews* 2010; **29**(5-6): 511–533. DOI:10.1080/07474938.2010.481550.
25. Wei X, Yang Y. Robust forecast combinations. *Journal of Econometrics* 2012; **166**(2): 224–236. DOI:10.1016/j.jeconom.2011.09.035.
26. Hibon M, Evgeniou T. To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting* 2005; **21**(1): 15–24. DOI:10.1016/j.ijforecast.2004.05.002.
27. Wong KK, Song H, Witt SF, *et al*. Tourism forecasting: to combine or not to combine? *Tourism Management* 2007; **28**(4): 1068–1078. DOI:10.1016/j.tourman.2006.08.003.
28. Sun S, Zhang C Yu G. A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems* 2006; **7**(1): 124–132. DOI:10.1109/TITS.2006.869623.
29. Stathopoulos A, Dimitriou L Tsekeris T. Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-Aided Civil and Infrastructure Engineering* 2008; **23**(7): 521–535. DOI:10.1111/j.1467-8667.2008.00558.x.
30. Zhang Y. Hourly traffic forecasts using interacting multiple model (IMM) predictor. *IEEE Signal Processing Letters* 2011; **18**(10): 607–610. DOI:10.1109/LSP.2011.2165537.
31. Djuric N, Radosavljevic V, Coric V, et al. Travel speed forecasting by means of continuous conditional random fields. *Transportation Research Record: Journal of the Transportation Research Board* 2011; **2263**: 131–139. DOI:10.3141/2263-15.
32. Min W, Wynter L. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies* 2011; **19**(4): 606–616. DOI:10.1016/j.trc.2010.10.002.
33. Qiao W, Haghani A Hamedi M. Short-term travel time prediction considering the effects of weather. *Transportation Research Record: Journal of the Transportation Research Board* 2012; **2308**: 61–72. DOI:10.3141/2308-07.
34. Fei X, Lu C-C Liu K. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies* 2011; **19**(6): 1306–1318.

35. Tsirigotis L, Vlahogianni EI Karlaftis MG. Does information on weather affect the performance of short-term traffic forecasting models? *International Journal of Intelligent Transportation Systems Research* 2012; **10**(1): 1–10. DOI:10.1016/j.trc.2010.10.005.
36. Florio L, Mussone L. Neural-network models for classification and forecasting of freeway traffic flow stability. *Control Engineering Practice* 1996; **4**(2): 153–164. DOI:10.1016/0967-0661(95)00221-9.
37. Li R, Rose G. Incorporating uncertainty into short-term travel time predictions. *Transportation Research Part C: Emerging Technologies* 2011; **19**(6): 1006–1018. DOI:10.1016/j.trc.2011.05.014.
38. Huang S-H, Ran B. *An Application of Neural Network on Traffic Speed Prediction Under Adverse Weather Condition*University of Wisconsin: Madison, 2003.
39. Li C-S, Chen M-C. Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks. *Neural Computing and Applications* 2013; **23**(6): 1611–1629. DOI:10.1007/s00521-012-1114-z.
40. Jin X, Zhang Y, Yao D. Simultaneously prediction of network traffic flow based on PCA-SVR, *Advances in Neural Networks*–ISNN 2007: Springer, 2007; 1022-1031.
41. Draper NR, Smith H Pownell E. *Applied Regression Analysis*Wiley: New York, 1966.
42. Beale MH, Hagan MT, Demuth HB. Neural Network Toolbox Reference (R2016a). Retrieved June 22, 2016 from http://www.mathworks.com/help/pdf_doc/nnet/nnet_ref.pdf

APPENDIX A.1: BASIC FORECASTING MODELS

A.1.1 LOCAL LINEAR REGRESSION (LLR)

For n, consider the following autoregressive model for forecasting:

$$x_n = a_n^T X_n = a_n^T (x_{n-1}, \cdots, x_{n-d}).$$

To obtain parameter a_n, we consider the following weighted linear regression:

$$a_n = \arg\min_{a_n} \sum_{i \leq m} w_i \cdot \left(x_i - a_n^T \overrightarrow{X_i} \right)^2,$$

where weight w_i, which measures the proximity between current input $\overrightarrow{X_n}$ and history input $\overrightarrow{X_i}$, is computed as follows:

$$w_i = w\left(\overrightarrow{X_i}, \overrightarrow{X_n} \right).$$

Therefore, the history data that are similar to the current input will be given more weight.
The forecasted value is computed as follows:

$$\hat{x}_n = a_n^T X_n.$$

In this work, We consider the training set size $m = 30$, $day = 4320$, delay $d = 3$, and the following distance function:

$$w(X, Y) = \exp\left(-\frac{|X - Y|_2^2}{4\sigma_x^2} \right),$$

where σ_x^2 is the variance of input series $\{x_n\}$.

A.1.2 NEURAL NETWORK WITH EXTERNAL INPUT (NNX)

Consider the following nonlinear mapping:

$$x_n = f(x_{n-1}, \cdots, x_{n-d}, y_n, y_{n-1}, \cdots, y_{n-d}),$$

where $\{y_n\}$ is the external input (precipitation $\{w_n\}$). We attempt to fit this function into a two-layer feedforward neural network that includes a hidden layer and an output layer.

We use the neural network toolbox of MATLAB [42] to implement this model. In particular, we adopt the nonlinear autoregressive with external input (NARX) structure with delay $d = 3$ and 10 hidden neurons. Additional details can be found in MATLAB online documentation.

A.1.3 PRINCIPAL COMPONENT ANALYSIS AND SUPPORT VECTOR REGRESSION (PCA–SVR)

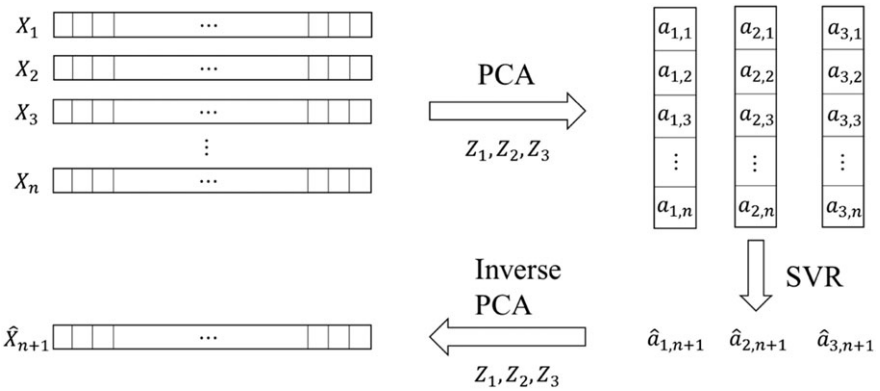
We view the data from certain period c ($c = 1$ *day*) as the following vector:

$$X_n = \{x_{c(n-1)+1}, \cdots, x_{cn}\}.$$

To reduce dimension, we perform PCA (with 28 training days), and then project each X_i on the first three components $Z_1, Z_2,$ and Z_3, following the advice of Jin *et al*. [40]. Afterward, we apply SVR on projections a_{1i}, a_{2i}, and a_{3i} to forecast $a_{1(n+1)}$, $a_{2(n+1)}$, and $a_{3(n+1)}$, and reconstruct the following forecasted vector:

$$\hat{X}_{n+1} = a_{1(n+1)}Z_1 + a_{2(n+1)}Z_2 + a_{3(n+1)}Z_3.$$

The preceding process is illustrated as follows:



In this model, only the parameters in SVR should be determined. We use the radial basis function kernel and let $d = 3$ (we use the past three values to forecast). Then, for each $j \in \{1, 2, 3\}$, we take $C = 100$, $\varepsilon = \min_i |Z_{ji}|$, and $\lambda = 0.5\sum_i (Z_{ji} - \overline{Z}_j)^2$. These parameters are tuned to achieve the best performance across all detectors.

PRINCIPAL COMPONENT ANALYSIS AND SPECTRAL ANALYSIS (PCA–SPEC)

We view the data from certain period c ($c = 1$ *week*) as the following vector:

$$X_n = \{x_{c(n-1)+1}, \cdots, x_{cn}\},$$

DOI: 10.1002/atr

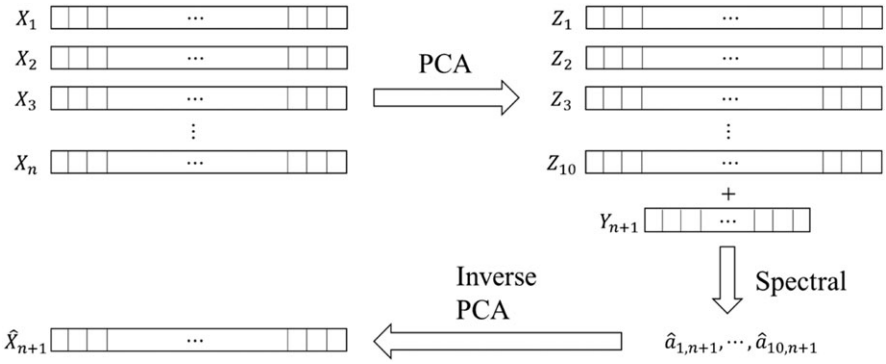
apply PCA (with four training weeks), and project each X_i on the first 10 components $Z_1, Z_2, \cdots, Z_{10}$. Instead of direct forecasting (e.g. PCA–SVR), we use a partial observation in the new cycle $Y_{n+1} = \{x_{cn+1}, \cdots, x_{cn+k}\}$ ($k = 1$ day *or* 6 days), and consider the following vector:

$$W_i = \{z_{i1}, \cdots, z_{ik}\}, i = 1, \cdots, 10.$$

We regress Y on W_i to obtain the length estimates $a_{1(n+1)}, \cdots, a_{10(n+1)}$ for the entire observation of the new cycle X_{n+1}, and then reconstruct the forecasted vector as follows:

$$\hat{X}_{n+1} = \sum_{i=1}^{10} a_{i(n+1)} Z_i.$$

This process is illustrated as follows:



This model does not require any parameter. We only forecast the value of the next day and then rerun the model each day (although a complete cycle lasts 1 week).

EXAMPLE OF FUSION FAILURE

The PCA–spec model significantly outperforms all the other models (corrected for precipitation) in conducting 1 hour-ahead forecasting for detector 4050. If we apply the original fusion method (10–12), then the integrated results are significantly worse than those of PCA–spec as shown in Table III.

Table III. One hour-ahead forecasting with detector 4050.

	MAPE (%)		
Methods	ALL	$w > 0$	$w > 1$
LLR	24.5803	11.4026	47.7189
NNX	42.4104	44.2189	100.5907
PCA–Spec	18.3740	3.9968	2.3102
PCA–SVR	41.3012	51.8825	123.7445
Integrated model	20.20	10.84	23.38

Notes:
"ALL" denotes the results from all the data points and represents efficiency in the overall scenario.
"$w > 0$" considers the results from data with $w > 0$ and represents efficiency in a rainy day scenario (including the scenario where $w > 1$).
"$w > 1$" considers the results from data with $w > 1$ and represents efficiency in a heavy rain scenario.

DOI: 10.1002/atr

However, when applying the revised fusion method (with $M = 10$ and $\Delta = 2\,\%$) in the same case, the accuracy of the integrated model increases to 14.13%, 4.02%, and 1.71% for the overall, rainy, and heavy rain scenarios, respectively. Therefore, the revised fusion method is highly robust toward individual failure.

VALIDATION OF CORRECTION ASSUMPTION

We check if the correlation between the forecasting error $(\hat{f}_n - f_n)/\hat{f}_n$ and precipitation proxy u_n is linearly significant; if yes, then the linear correction will improve forecasting accuracy. Using the same data set, we adopt the data from the first month and the last 25 days as the training and testing sets for our four basic forecasting models, respectively. Table IV shows that the correction factor $\hat{c}$ is significant (with a significance level of $\alpha = 0.05$) in 63.54% of the scenarios, thereby validating our assumption.

COMPARISON AMONG BASIC MODELS

Table V shows the efficiencies of the individual basic models (corrected for precipitation) with lag = 1 day. None of these models can outperform the other models in each scenario. In the general scenario, the LLR and PCA–spec models exhibited better efficiency than the other two models, whereas in the rainy day scenario, the NNX and PCA–SVR models were considerably more accurate than the other two models. Therefore, we assume that forecasting efficiency can be improved by integrating the results from different models.

Table IV. Assumption examination result.

Traffic detector	General		Lag = 10 minutes		Lag = 1 hour		Lag = 1 day	
	Model	Sign	Model	Sign	Model	Sign	Model	Sign
2010	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	1	NNX	1	NNX	1	NNX	1
2011	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	1	NNX	0	NNX	0	NNX	0
2012	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	1	NNX	1	NNX	1	NNX	1
2023	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	1	NNX	0	NNX	0	NNX	1
2030	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	0	NNX	0	NNX	0	NNX	1
2033	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	0	NNX	0	NNX	0	NNX	1
2052	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	0	NNX	0	NNX	0	NNX	1
3034	PCA–spec	0	LLR	0	LLR	0	LLR	0
	PCA–SVR	0	NNX	0	NNX	1	NNX	0
3035	PCA–spec	1	LLR	1	LLR	0	LLR	0
	PCA–SVR	1	NNX	0	NNX	1	NNX	0
4004	PCA–spec	1	LLR	1	LLR	1	LLR	1
	PCA–SVR	1	NNX	0	NNX	1	NNX	0
4005	PCA–spec	1	LLR	0	LLR	1	LLR	0
	PCA–SVR	1	NNX	0	NNX	1	NNX	0
5062	PCA–spec	1	LLR	0	LLR	1	LLR	1
	PCA–SVR	1	NNX	0	NNX	0	NNX	0

Note: "Sign" indicates model significance, which is equivalent to 1 if $\hat{c}$ is considerably different from 0.

DOI: 10.1002/atr

Table V. VEfficiency comparison among selected models.

Traffic detector	Model	General MAPE % (corrected)	Rainy MAPE % (corrected)	Model	General MAPE % (corrected)	Rainy MAPE % (corrected)
2010	PCA–spec	7.41	8.07	LLR	7.19	9.01
	PCA–SVR	9.76	6.81	NNX	7.40	9.12
2011	PCA–spec	9.18	7.74	LLR	9.42	8.59
	PCA–SVR	10.65	6.64	NNX	9.61	8.64
2012	PCA–spec	12.31	15.41	LLR	12.72	17.10
	PCA–SVR	13.46	12.68	NNX	12.28	12.24
2023	PCA–spec	5.55	7.55	LLR	6.59	9.64
	PCA–SVR	7.08	6.35	NNX	5.90	7.64
2030	PCA–spec	5.10	6.74	LLR	5.89	7.33
	PCA–SVR	7.68	4.23	NNX	5.69	6.13
2033	PCA–spec	4.78	5.06	LLR	5.05	5.36
	PCA–SVR	6.46	3.41	NNX	5.22	4.92
2052	PCA–spec	8.24	9.46	LLR	9.36	12.53
	PCA–SVR	11.19	8.54	NNX	9.49	10.75
3034	PCA–spec	6.40	9.37	LLR	6.38	8.97
	PCA–SVR	8.77	9.27	NNX	6.68	8.34
3035	PCA–spec	6.07	9.68	LLR	6.01	10.15
	PCA–SVR	7.89	11.26	NNX	6.07	9.11
4004	PCA–spec	8.01	12.68	LLR	6.93	10.10
	PCA–SVR	8.12	8.91	NNX	7.11	9.24
4005	PCA–spec	5.38	7.06	LLR	5.82	7.60
	PCA–SVR	8.20	5.51	NNX	6.24	7.05
5062	PCA–spec	10.67	11.90	LLR	9.24	10.45
	PCA–SVR	10.62	9.91	NNX	9.71	10.24

DOI: 10.1002/atr