Pattern-Based Evaluation of Coupled Meteorological and Air Quality Models

SCOTT BEAVER AND SAFFET TANRIKULU

Bay Area Air Quality Management District, San Francisco, California

AHMET PALAZOGLU AND ANGADH SINGH

University of California, Davis, Davis, California

SU-TZAI SOONG, YIQIN JIA, AND CUONG TRAN

Bay Area Air Quality Management District, San Francisco, California

BRUCE AINSLIE AND DOUW G. STEYN

The University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 13 January 2010, in final form 8 April 2010)

ABSTRACT

A novel pattern-based model evaluation technique is proposed and demonstrated for air quality models (AQMs) driven by meteorological model (MM) output. The evaluation technique is applied directly to the MM output; however, it is ultimately used to gauge the performance of the driven AQM. This evaluation of AQM performance based on MM performance is a major advance over traditional evaluation methods. First, meteorological cluster analysis is used to assign the days of a historical measurement period among a small number of weather patterns having distinct air quality characteristics. The clustering algorithm groups days sharing similar empirical orthogonal function (EOF) representations of their measurements. In this study, EOF analysis is used to extract space–time patterns in the surface wind field reflecting both synoptic and mesoscale influences. Second, simulated wind fields are classified among the determined weather patterns using the measurement-derived EOFs. For a given period, the level of agreement between the observation-based clustering labels and the simulation-based classification labels is used to assess the validity of the simulation results. Mismatches occurring between the two sets of labels for a given period imply inaccurately simulated conditions. Moreover, the specific nature of a mismatch can help to diagnose the downstream effects of improperly simulated meteorological fields on AQM performance. This pattern-based model evaluation technique was applied to extended simulations of fine particulate matter ($PM_{2.5}$) covering two winter seasons for the San Francisco Bay Area of California.

1. Introduction

Photochemical air quality model (AQM; Russell and Dennis 2000) simulations are increasingly used for regulatory purposes (Fine et al. 2003). They provide technical information to support air quality planning decisions. The resulting policies can affect billions of dollars worth of public health and economic activity annually (Yang et al. 2005). Because of the large stakes involved, policy makers require confidence that simulation results are valid. For use in policy making, AQM simulations must go beyond merely reproducing observed pollutant levels. They must additionally represent atmospheric processes with sufficient fidelity to allow inferences about dominant pollutant buildup mechanisms. Understanding the major buildup pathways for regulated pollutants provides the only sound basis for optimizing emission control strategies. Thus, modelers must additionally evaluate AQM inputs such as meteorological fields, pollutant and precursor emissions inventories, and land use, as well as the inner workings of the model to track ambient conditions.

AQMs need to accurately reproduce dominant pollutant buildup pathways across the full range of meteorological

Corresponding author address: Saffet Tanrikulu, Bay Area Air Quality Management District, 939 Ellis St., San Francisco, CA 94109.
E-mail: stanrikulu@baaqmd.gov

DOI: 10.1175/2010JAMC2471.1

conditions experienced during the modeled period. The atmospheric processes responsible for pollutant buildup are represented in gridded meteorological fields input to AQMs. These fields are often prepared using separate meteorological models (MMs). Thus, biases in the MM output are propagated through the AQM. If MM biases can be identified, they can provide strong signals to assess the accuracy and diagnose the shortcomings of an AQM.

Traditionally, simulated variables such as wind speed and direction, temperature, and humidity are directly compared against routine measurements (e.g., Seaman 2000) in an attempt to gauge the accuracy of the MM output. Such an operational evaluation (Tesche 2002) is simple to implement: error and bias statistics are computed between corresponding modeled variables and measured parameters that are paired in space and time. These paired comparison statistics assay the magnitude and sign of the discrepancy between simulation and observation. Large errors and/or biases indicate MM outputs that are unacceptable for use in AQMs. Smaller error and bias levels, however, do not guarantee the modeled meteorological fields are acceptable for use in AQM simulations. That is, error and bias statistics are necessary but insufficient for evaluating MM outputs as AQM inputs. For example, two ozone episodes from the same summer were modeled for the Central California Ozone Study (CCOS). Operational evaluations of the simulated meteorological fields indicated similar MM performance for both episodes (Wilczak et al. 2005; Tesche et al. 2004). These fields drove otherwise identical AQM simulations. AQM performance was adequate for one episode but poor for the other.

Various technical issues limit the robustness of operational evaluation statistics. A commonly cited problem is incommensurability (Swall and Foley 2009), also known as change of support (Wilke 2003): measurements, which are point estimates, are not directly comparable with modeled quantities, which are volume averages. Also, decreasing the model grid size often fails to yield better operational performance statistics (Rife and Davis 2005, and references therein). Most surface meteorological networks are insufficiently dense to sample the localized air flows represented in a finely gridded MM (Gego et al. 2005). Additionally, error and bias statistics cannot account for stochastic fluctuations that are absent from deterministic model outputs (Hanna and Yang 2001). Finally, point-by-point operational evaluation cannot distinguish between atmospheric features that are missing altogether in the simulated fields, as opposed to those that are present but dislocated in time and/or space.

Evaluation methods based on space–time patterns (Casati et al. 2008) often provide more physical insight than operational evaluation. Generally, a statistical method is applied to extract patterns from either a spatial field or a time series for both simulated and observed values. The extracted patterns are then compared between simulation and observation. This framework to compare patterns avoids the direct pairing of observed point estimates with simulated quantities. Incommensurability issues are largely avoided. Also, the patterns are estimated using multiple data points (from either a spatial field or time series). This approach contrasts markedly with operational evaluation, which pairs single data points in space or time. Thus, simulated quantities may be more robustly compared against measurements by using a pattern-based approach instead of paired statistics. Moreover, evaluation explicitly based on space–time patterns is likely to characterize a model's ability to reproduce physically relevant atmospheric features. Alternatively, operational evaluation is purely empirical and may lack physical meaning.

There are many types of model evaluation based on space–time patterns. Spectral decomposition can determine whether important time scales are sufficiently represented in a simulation. Decompositions can be performed using linear filtering (Rao et al. 1997; Gilliam et al. 2006) or wavelets (Li and Shue 2004). Also, joint distributions between different fields (e.g., wind and temperature) can be estimated. Comparing simulated and observed joint distributions can indicate how well the temporal coherence of the respective fields has been simulated (Mueller 2009). Spatial patterns are commonly isolated from fields of model output and observations using empirical orthogonal functions (EOFs; Ludwig et al. 1995), also known as principal component analysis (PCA; Rohli et al. 2004). These spatial patterns can then be compared qualitatively and/or quantitatively to evaluate the simulated fields. Cluster analysis (Ainslie and Steyn 2007) and other data partitioning methods (Cannon et al. 2002) are also useful for extracting patterns in space and/or time. In practice, multiple space–time statistical techniques may be combined to focus on specific scales at which conceptually important phenomena occur.

Ideally, coupled MM–AQM evaluation should explicitly account for MM shortcomings that may degrade AQM performance. In practical terms, such an evaluation technique would save resources, as meteorological fields unsuitable as AQM inputs could be identified directly. This foresight would avoid the costs of running and evaluating AQM simulations destined to perform poorly. An evaluation technique that predicts AQM performance based on MM performance is also attractive from a scientific standpoint. Empirical relationships linking the weather and air quality may aid conceptual model

development for air pollution meteorology (Christakos 2003).

This paper introduces a novel model evaluation technique for a coupled MM–AQM in which the MM output is used as AQM input. It is based on comparing statistically extracted space–time weather patterns embedded in meteorological observations and MM output. The method achieves two important goals. First, it predicts how inaccuracies for MM-generated meteorological fields may degrade AQM performance. Second, it evaluates coupled MM–AQM performance across different weather patterns. It can identify and diagnose representative weather patterns for which systematically poor MM performance consistently degrades AQM performance.

2. Proposed pattern-based evaluation framework

We propose a two-stage pattern-based framework for coupled MM–AQM evaluation. First, actual meteorological conditions are categorized among, or binned into, a set of statistically defined weather patterns. Each weather pattern should reflect different spatial and temporal distributions for the analyzed meteorological parameters and should also be associated with distinct air pollution characteristics. The historical period from which the weather patterns are identified must include, but can extend beyond, the simulation period for which model evaluation will be performed. Second, MM outputs are classified into the previously identified, measurement-derived weather patterns. For a given period, agreement between the observation-based categorization and the simulation-based classification implies model validity. A mismatch in the labeling of some period between simulation and observation implies that the distribution of simulated quantities is inconsistent with observation. The nature of any mismatch allows inference as to how AQM performance may be degraded by MM shortcomings. Simulated pollutant levels are likely to resemble those associated with the mistakenly simulated weather pattern indicated by the classification, instead of the observed weather pattern reflected by the measurements. Longer durations for such mismatches will likely result in increasingly severe AQM performance degradation.

Labelings of the observed and simulated meteorological conditions are implemented using unsupervised (data driven) cluster analysis and supervised classification, respectively (Jain 2000). Clustering both identifies the measurement-derived weather patterns and labels their times of occurrence. Each weather pattern, or cluster, is associated with a distinctly parameterized statistical model that best describes the distribution of its assigned data. Classification determines which measurement-derived weather pattern most closely matches the modeled meteorological fields for a given period. The classification is performed using a statistical calculation analogous to that used by the clustering algorithm that defined the weather patterns. Here, both clustering and classification are based on EOF analysis of wind fields, as described in section 3.

Clustering of meteorological parameters can readily establish the links between measured air quality and observed meteorological conditions. These links allow prediction of AQM performance based on MM output. But, the clustering does need to be implemented on appropriate data to identify atmospheric features at scales relevant to air quality over the modeling domain. These scales may generally include the planetary, synoptic, meso-, and microscales. This model evaluation framework assumes that nonmeteorological inputs to the AQM such as emissions, chemistry, and land use are reasonably accurate. Otherwise, it may not be possible to determine the causes of AQM performance issues based on an evaluation of meteorological fields alone.

To illustrate the utility of pattern-based model evaluation, consider a simple two-pattern example. Suppose that actual conditions are clustered as stagnant, but simulated conditions are classified as windy and turbulent. In this case, the AQM would be expected to underestimate pollutant levels. On the other hand, suppose that conditions are in fact windy but are simulated as stagnant. In that case, the AQM would be expected to overestimate pollutant levels. Continuing with the same example, consider the case in which actual conditions are stagnant for an entire week. If only one day of the week is mistakenly simulated as windy, then AQM performance may not be severely degraded. If the mismatch occurs for the entire week, however, the AQM performance would be expected to be worse.

3. Theory

a. EOF analysis of wind fields

In this study, clustering of meteorological observations and classification of MM outputs are both based on EOF analysis (Lorenz 1956), also known as PCA (Jolliffe 2002). This statistical approach can extract features from meteorological parameters measured over space and time. Here, EOFs are estimated from hourly u and v wind components measured from a network of s surface weather stations. In terms of PCA, the model is applied in the S mode (Serrano et al. 1999) with the parameters (u or v at a specific station) treated as ''variables'' and the sampling times treated as ''cases.'' The u and v components for each station are scaled

by dividing by the mean observed wind speed for that station. This scaling weights each weather station roughly equally in the EOF analysis without distorting the wind directions. To account for the autocorrelation (Shumway and Stoffer 2005) in the hourly wind measurements, replicated values at 1- and 2-h delays are concatenated to the original values. Scaled values for each hour h are stacked into "measurement vectors" $\mathbf{x}(h)(1 \times 6s)$ as follows, where the subscript is an index over the s weather stations,

$$\mathbf{x}(h) = \begin{bmatrix} u_1(h), v_1(h), u_1(h-1), v_1(h-1), u_1(h-2), v_1(h-2), \ldots, \\ u_s(h), v_s(h), u_s(h-1), v_s(h-1), u_s(h-2), v_s(h-2) \end{bmatrix}. \quad (1)$$

The EOFs for any set of measurements vectors $\mathbf{x}(h)$ are estimated by applying singular value decomposition (SVD). Each EOF is associated with a singular value σ_i that is proportional to the amount of variance in the set of measurements vectors explained by that EOF. The EOFs are rank ordered by decreasing level of variability explained. The first EOF has the lowest order (1), has the largest singular value, and explains more variance than any other EOF. Of a possible $6s$ EOFs, only the first $n_{\max} \ll 6s$ EOFs are retained and stacked into the columns of $\mathbf{P}(6s \times n_{\max})$. The percentage of the variability in the decomposed data that is explained by the first $n_{\max}$ EOFs is calculated from the singular values,

$$\%\text{variance explained} = \frac{\sum_{i=1}^{n_{\max}} \sigma_i}{\sum_{i=1}^{6s} \sigma_i} \times 100. \quad (2)$$

The EOFs are orthogonal, and, when applied to time series values, reflect wind field variability within distinct frequency bands (Galin 2007). Atmospheric processes occurring at lower frequencies generally have larger spatial scales (Steyn et al. 1981). Because the EOFs are a spectral decomposition of the wind field time series, they are also associated with any atmospheric processes that are coherent (correlated in time) with the wind field.

For regional study domains in which large-scale influences dominate the weather, the EOF rank-ordering is similar to the classical concept of wavenumber for numerical weather modeling. Relatively lower ranked EOFs tend to represent features at relatively larger scales. Synoptic influences generally affect all stations in a region and thereby tend to contribute the largest amounts of variability to the measurements vectors. Thus, synoptic influences tend to be represented by the lower-order EOFs. More localized atmospheric features affect subsets of the stations and thereby tend to contribute less to the overall variability in the wind field. These mesoscale influences tend to be represented by the middle-order EOFs. Microscale influences and stochastic fluctuations may affect each weather station uniquely. They tend to be represented in the higher-order EOFs, which explain small amounts of variability. Microscale structures in the boundary layer are of little interest for model validation purposes because MMs typically do not represent such finescale processes. Thus, the user should attempt to select $n_{\max}$ to retain the lower-order (synoptic) and middle-order (mesoscale) EOFs and discard the higher-order (microscale) EOFs.

b. EOF-based cluster analysis

The nontraditional nonhierarchical clustering algorithm of Beaver and Palazoglu (2006a) is applied to measurements vectors $\mathbf{x}(h)$ to produce k clusters of days, or weather patterns. Each cluster c is represented by a distinctly parameterized set of $n_{\max}$ EOFs appearing as the columns of matrix $\mathbf{P}_c(6s \times n_{\max})$. The parameter $n_{\max}$ is determined by trial and error such that all clusters sufficiently reflect the various synoptic and mesoscale phenomena represented in their assigned measurements. The clustering algorithm is constrained to always assign the 24 h from a given day (midnight to midnight, local time) to the same cluster. This blocking of the hourly cluster assignments into 24-h windows serves as a simple low-pass filtering to generate daily labels by clustering hourly measurements. Measurement vectors $\mathbf{x}(h)$ for each day d appear as the rows of data block $\mathbf{X}(d)(24 \times 6s)$.

Initially, each cluster is randomly seeded with the daily data blocks. Then, the days are reassigned iteratively to produce an optimized set of clusters. On each iteration, an EOF model $\mathbf{P}_c$ is estimated for each cluster c from its assigned data blocks $\mathbf{X}(d)$ vertically concatenated into the rows of supramatrix $\boldsymbol{\chi}_c(24N_c \times 6s)$, where N_c is the number of days assigned to cluster c. The scalar sum-of-squares errors totaled across 24 h, $e_c(d)$, is computed for fitting the block of data for each day d into the EOF model for each cluster c,

$$e_c(d) = [\|\mathbf{X}(d)(\mathbf{I} - \mathbf{P}_c\mathbf{P}_c^{\mathrm{T}})\|_F]^2 = \sum_{h=t(d)}^{t(d)+23} [|\mathbf{x}(h)(\mathbf{I} - \mathbf{P}_c\mathbf{P}_c^{\mathrm{T}})|]^2. \quad (3)$$

The notation indicates squared Frobenius (matrix Euclidean) and L2 (vector Euclidean) norms, $\mathbf{I}$ is an identity matrix, and $t(d)$ is the first hour (midnight local time) of day d. Then, each day d is reassigned to the cluster c satisfying $\text{argmin}_c e_c(d)$. The iterative procedure continues until no further reassignments are possible.

The nonhierarchical algorithm usually converges to a local minimum for a given value of k, the number of clusters. The procedure of Beaver and Palazoglu (2006b) is applied to an ensemble of randomly initialized runs of the clustering algorithm. This randomized resampling approach, similar to bootstrapping, yields a final ensemble-averaged solution with an appropriate number of clusters that is near the global minimum of the solution space. Most days are assigned a single cluster label having a high level of confidence. Some transitional days sharing properties of two clusters may be doubly assigned with moderate confidence. A small proportion of the days cannot be assigned to any cluster with reasonable confidence and remain unlabeled.

Properly identified weather patterns should be associated with distinct aloft conditions, despite no aloft data having been input to the clustering algorithm. The corresponding aloft conditions for each cluster are determined by compositing weather maps across the days assigned to that cluster. Cluster-averaged precipitation and surface temperature fields can further characterize the weather patterns.

c. *Multiscale classification of model outputs using EOFs*

Once the weather patterns are established by clustering, MM outputs can be classified among these known categories. Simulated wind values are first interpolated from model grid points to the corresponding locations of the weather stations that provided the clustered measurements. The simulated u and v components are scaled by the modeled mean wind speed for that location. This scaling reduces discrepancies in wind speed between model and observation while still preserving the simulated wind field spatial structure. The processed model output is arranged into vectors analogously to (1). Then, the sum-of-squares errors for fitting each day of model output into each cluster's EOFs are calculated analogously to (3). Each day of model output is classified into the cluster having the EOFs that represent those simulated winds with the smallest sum-of-squares error.

The classification can be performed using different subsets of the EOFs. Here, classifications are always performed using the first n EOFs, where $n \leq n_{\max}$ is said to be the EOF model order. This hierarchical EOF model structure is used because finer-scale patterns are not well defined without being superimposed on their larger-scale settings. Classification using EOF model order n is achieved analogously to (3), except using only the first n columns of $\mathbf{P}_c$. A given day of model output may be classified into different clusters at different EOF model orders. Such behavior indicates simulated conditions corresponding to different measurement-derived weather patterns at different scales. The clustering, on the other hand, only needs to be performed once using $n_{\max}$ EOFs, as determined during execution of the algorithm. The clustering estimates a total of $n_{\max}$ EOFs for each cluster that are by definition consistent across all scales. This consistency across scales for the measurement-based clusters reflects how distinct synoptic regimes set the stage for distinct mesoscale air flows to develop.

4. Case study

a. *Description of study domain*

The proposed pattern-based model evaluation technique is demonstrated for the San Francisco Bay Area (SFBA) of California (Fig. 1) for the core fine particulate matter ($PM_{2.5}$) season of December–January. Exceedances of the 24-h $PM_{2.5}$ National Ambient Air Quality Standard (NAAQS) of 35 μg m^{-3} occurred mostly during these months. During these winter episodes, synoptic-scale stability and subsidence often trapped $PM_{2.5}$ and its precursors close to the ground. At the mesoscale, terrain-induced air flows defined the source–receptor relationships, limited pollutant dispersion, and controlled the $PM_{2.5}$ spatial distribution. The SFBA is ideal for demonstrating pattern-based model evaluation at both the synoptic scale and mesoscale.

The SFBA is part of the larger central California domain, which also includes the Sacramento Valley (SV) and the San Joaquin Valley (SJV). This pair of large, inland valleys together forms the Central Valley (CV). The SFBA, the SV, and the SJV have major connections at the Delta region to the east of the Bay. Air flows between the SFBA and the CV occurred through the narrow Carquinez Strait, the only major gap in the rims surrounding the CV. These three basins shared similar air quality characteristics because of similar emissions, coupled meteorological conditions, and interconnected terrain. During episodic winter conditions, the SV and/or the SJV were often upwind of the SFBA.

b. *Summary of previous cluster analysis results*

Cluster analysis was applied to SFBA surface wind measurements from 12 winter seasons (November–March) from 1 January 1996 to 31 March 2007 (Beaver et al. 2010). Clustering 1754 days robustly identified the

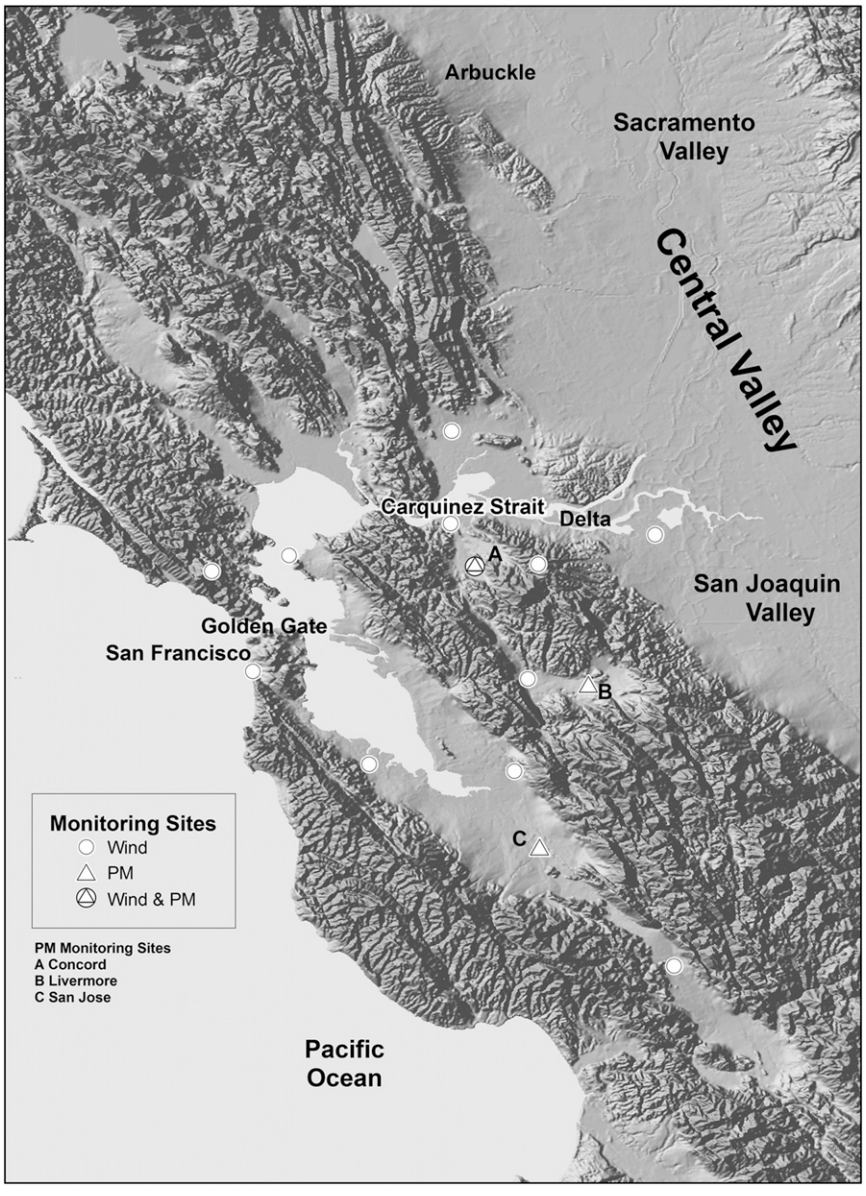

FIG. 1. SFBA and partial CV study domain showing surface wind stations used in cluster analysis, $PM_{2.5}$ monitors, the Arbuckle weather station, and important geographic features.

relevant weather patterns impacting $PM_{2.5}$ levels. The clustered surface wind measurements were from a network of 12 SFBA weather stations shown in Fig. 1. Based on previous experience, microscale structures and noise were assumed to account for around 10% to 15% of the SFBA wind field variability. Thus, the cluster analysis was performed using $n_{max} = 14$ to explain around 85%–90% of the variability in each cluster that represented mostly synoptic and mesoscale influences.

The clustering identified five weather patterns having distinct $PM_{2.5}$ characteristics. The synoptic-scale conditions were resolved by compositing gridded pressure level data up to the 500-hPa pressure level. The 500-hPa composite National Centers for Environmental Prediction (NCEP) reanalysis (http://www.esrl.noaa.gov/psd/) geopotential height fields (not shown) were used to name the clusters. The type of synoptic features impacting the SFBA, their relative strengths of forcing on the surface winds, and cluster names are indicated in Table 1. Interregional surface airflow patterns for these clusters are shown in Fig. 2. Distinct wind field patterns in the CV, outside of the clustered domain, provided further evidence that the weather patterns are real. Each cluster was also verified to exhibit a distinct surface temperature

TABLE 1. Names, number of occurrences, number of NAAQS exceedance days (any SFBA monitor exceeds 35 μg m^{-3}), and qualitative characteristics of five clusters.

Name	No. days total	No. exceedance days	$PM_{2.5}$ levels	Synoptic (500-hPa level) feature	Strength of synoptic forcing
R1	219	7	Moderate	Offshore high pressure ridge	Strong
R2	422	145	Highest	Shoreline high pressure ridge	Weak
R3	279	25	High	Inland high pressure ridge	Weakest
V	413	6	Low	Trough (ventilated)	Strong
S	489	6	Low	Storm–cyclone (zonal flow aloft)	Strong

pattern (not shown). These five weather patterns are fully described in Beaver et al. (2010).

Three clusters (named R1, R2, and R3) were associated with anticyclonic conditions and elevated $PM_{2.5}$ levels; ''R'' denotes upper-level high pressure ridges. Over 80% of the SFBA $PM_{2.5}$ 24-h exceedances occurred under R2. The rest occurred mostly under R3. Episodic weather patterns R2 and R3 both had ridges of aloft high pressure positioned over the SFBA, resulting in weak large-scale pressure gradients. Light, shallow, easterly air flows developed around the SFBA. Cluster R3 had the weakest large-scale forcing and the lowest SFBA wind speeds. It was also the only weather pattern with diurnally reversing wind directions. Both episodic weather patterns exhibited near-calm conditions throughout the CV. Like the episodic weather patterns, R1 also had easterly surface winds through the SFBA. Unlike the episodic patterns, however, the R1 airflow pattern was driven by a strong large-scale pressure gradient. The R1 flow pattern was relatively deep, and both moderate mechanical mixing (vertical dispersion) rates and mixing depths resulted in moderate $PM_{2.5}$ levels. Strong winds entered the SV from the north and flowed southward along the SV major axis. Unlike the episodic

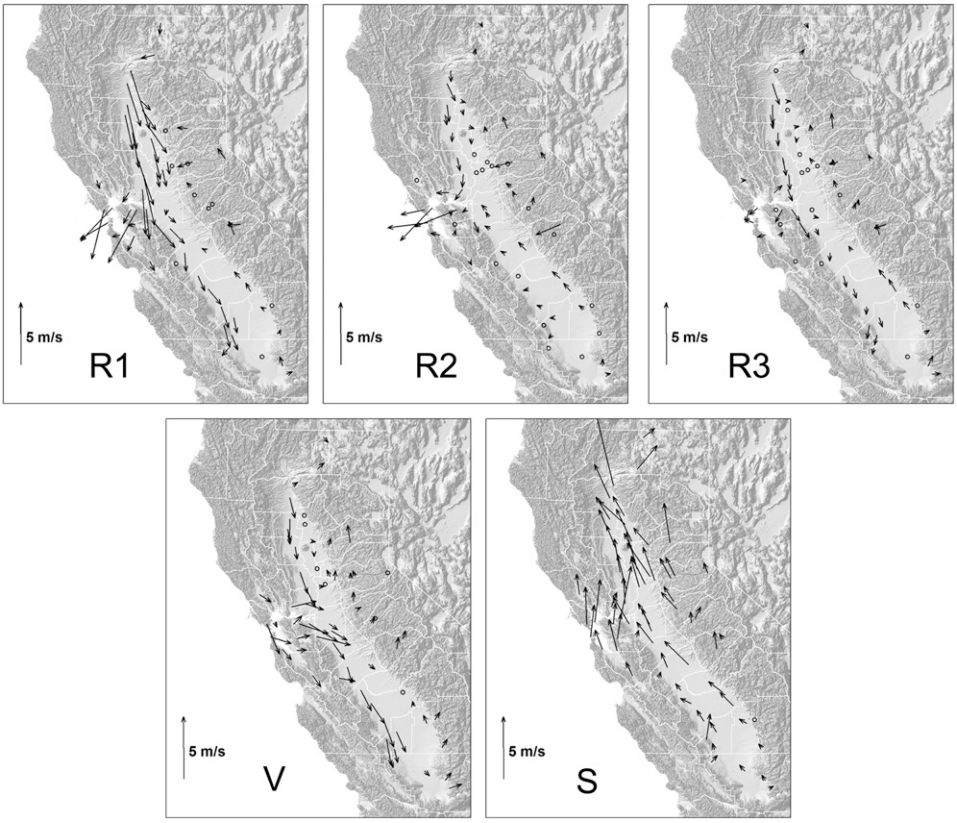

FIG. 2. Mean 0900 PST 1-h surface wind fields for five clusters. Arrow lengths are proportional to wind speed. Arrows point along direction of wind. Arrow tails are positioned at stations.

TABLE 2. Numbers of assigned days for observed clusters (first two columns). Numbers of simulated days from each cluster classified to each pattern, using lower- and middle-order EOFs. Sums across columns for the classification tables (center and right groups of columns) may not match value in same row under ''observed clusters'' because any doubly assigned days were counted as full matches to both patterns.

Observed clusters		Classification using first 3 EOFs (lower order)					Classification using first 11 EOFs (middle order)				
Name	No. days	R1	R2	R3	V	S	R1	R2	R3	V	S
R1	31	26	0	0	0	4	28	0	0	0	4
R2	51	24	19	1	0	7	34	10	1	0	7
R3	23	12	1	4	2	6	14	0	4	3	4
V	11	5	0	0	4	4	2	0	1	5	5
S	22	1	1	0	0	22	2	0	0	1	20

weather patterns, the aloft ridge for R1 was positioned offshore instead of over the SFBA. Two other cyclonic weather patterns were named V for ventilated and S for stormy. Both exhibited strong large-scale pressure gradients, strong marine winds entering the SFBA from the west, and low $PM_{2.5}$ levels. Nearly all SFBA precipitation occurred under S.

c. Description of MM–AQM simulations

Mesoscale meteorological and photochemical simulations were performed for a subset of the 1996–2007 winter cluster analysis study period. The modeling domain included the SFBA, the SV, the SJV, and remote regions over the Pacific Ocean and the Sierra Nevada. Simulations for 1 December–2 February were performed for both the 2000/01 and 2006/07 winters, for 128 days total. Meteorological fields were prepared using the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) with 4-km horizontal grid size and 30 vertical layers. Then, $PM_{2.5}$ levels were simulated using the Community Multiscale Air Quality (CMAQ) model with the Statewide Air Pollution Research Center, version 1999 (SAPRC99), chemical mechanism and the Models-3 AERO3 aerosol module with the Regional Acid Deposition Model aqueous chemistry mechanism (AE3-aq). Emissions only varied by day of week, with significant weekday–weekend differences, and by winter season.

CMAQ performance was evaluated for three key monitoring locations. Gravimetric samplers analyzed using the federal reference method (FRM) provided daily 24-h $PM_{2.5}$ measurements at Concord and San Jose. Beta attenuation method (BAM) instruments provided daily 24-h $PM_{2.5}$ measurements at Livermore and San Jose. San Jose $PM_{2.5}$ level was taken as the average of the FRM and BAM measurements. Observations were compared against the minimally deviating simulated value within a 3×3 array of first-layer grid cells centered around the monitor. Pairing the observations with simulated values in adjacent grid cells helped account for the sharp $PM_{2.5}$ gradients over the complex terrain.

5. Results

a. MM evaluation

Simulated hourly winds were interpolated from the MM5 output to locations corresponding to the surface weather stations used in the clustering. These simulated winds were used to classify each day among the five weather patterns described in section 4b. Classifications were performed using EOF model orders 1–14.

Table 2 shows the correspondence of each pair of clustering (observation) and classification (simulation) labels using selected lower- and middle-order EOFs. The selected lower-order EOF classification used the first 3 EOFs and reflected mostly large-scale (synoptic) variability. The selected middle-order EOF classification used the first 11 EOFs and reflected more localized (mesoscale) circulations. In reality, there is a continuum of scales represented across the 14 EOF model orders. Representative results for two model orders (3 and 11) near opposite ends of this spectrum demonstrate the multiscale capabilities of the EOF-based evaluation technique.

Regardless of actual conditions (cluster label), MM5 was generally unable to reproduce the R3 pattern. The cluster analysis assigned 23 days from the simulation period to this pattern. Of the 128 simulated days, only 5 and 6 (column sums in Table 2) were simulated as R3 (correctly or otherwise) at the selected lower and middle EOF model orders, respectively. Many of the R3 days were incorrectly simulated as either R1 or S, both windy patterns. This mismatch suggested CMAQ would underestimate $PM_{2.5}$ levels for most of the R3 days because simulated wind speeds were too high.

MM5 also had trouble simulating R2. At the selected lower model order, under half (19 of 51) of the R2 days were correctly simulated. Around half (24 of 51) of these R2 days were mistakenly simulated as R1, a windy pattern. At the selected middle EOF model order, MM5 performance was further degraded. More R2 days (34) were mistakenly simulated as R1, and fewer R2 days (10) were correctly simulated. MM5 performance for this episodic cluster was more degraded at finer scales.

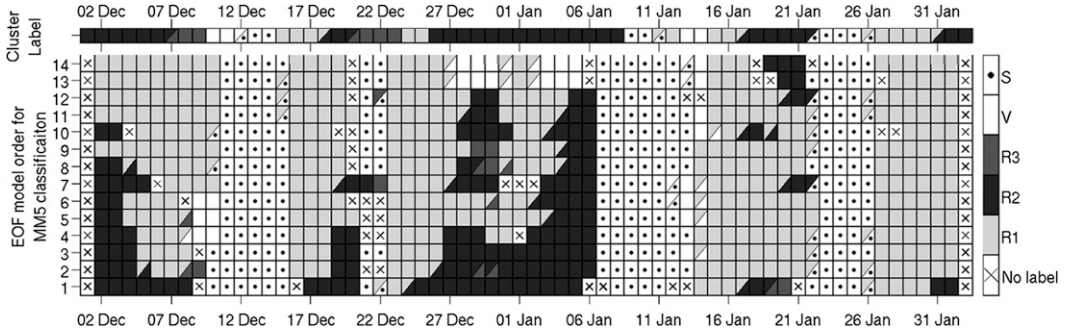

FIG. 3. Time series for cluster and classification labels for 2000/01 simulation period. Each square indicates label(s) for a single day using a given EOF model order. Squares are broken into pairs of triangles for doubly assigned days. (top) Cluster (observation) labels using $n_{max} = 14$; (bottom) classification (simulation) labels for $1 \leq n \leq n_{max}$ stacked vertically. Patterns R1, R2, and R3 are indicated by grayscale shading. Patterns V, S, and unlabeled days are shown as white with no marker, a dot, and an x, respectively.

These mismatches suggested that CMAQ would underestimate $PM_{2.5}$ levels for the majority of the R2 days because they were simulated as high wind pattern R1. This systematic MM5 bias to mistakenly simulate observed episodic R2 conditions as nonepisodic R1 conditions is termed the R2–R1 mismatch.

Across all scales, both R1 and S were usually simulated correctly. MM5 had difficulty simulating V on more than half of its occurrences. At lower order, V was likely to be confused with either R1 or S. At middle order, however, V was generally only confused with S. Simulating V as S was not likely to degrade CMAQ performance significantly, because both cyclonic weather patterns had low $PM_{2.5}$ levels.

Figure 3 shows the time series for the clustering and classification labels. Results of the classification across all 14 EOF model orders are shown for 2000/01 only. The time series for the model evaluation results indicated one significant problem with the timing of MM5. The clustering indicated that the transition R2 → S occurred over 8–9 January. MM5, however, produced this same transition two days in advance of the observed transition. This mistiming suggested that CMAQ would produce a premature decrease in simulated $PM_{2.5}$ levels before 8–9 January.

For most EOF model orders, the classification label varied smoothly with EOF model order (vertical dimension in Fig. 3, bottom panel). For example, R2 days were often simulated as R2 for lower model orders, but they were simulated as R1 for middle orders. Classifications using just the lowest lower-order EOF (1) and including the highest middle-order EOFs (13–14) were often inconsistent with those using intermediate EOF model orders (2–12).

b. AQM evaluation

The coupled MM–AQM evaluation technique generally indicated that episodic conditions (R2 and R3) were inaccurately simulated whereas nonepisodic conditions (R1, V, and S) were reasonably accurately simulated. Thus, CMAQ performance varied considerably between episodic and nonepisodic conditions. Days clustered into R1, V, or S exhibited simulated 24-h $PM_{2.5}$ levels at Concord, Livermore, and San Jose with mean biases (model minus observation; negative biases indicated model underestimation) and errors of -1.6 ± 6.4, -4.2 ± 8.8, and 0.0 ± 9.3 $\mu g\ m^{-3}$, respectively, relative to the measurements over 2000/01 and 2006/07. In comparison, days clustered into R2 or R3 had considerably poorer statistics of -7.3 ± 12.2, -13.7 ± 15.6, and -7.3 ± 14.8 $\mu g\ m^{-3}$, respectively. Despite significant biases, the observed and simulated $PM_{2.5}$ levels were well correlated. Pearson correlation coefficients between simulated and observed $PM_{2.5}$ levels for 2000/01 and 2006/07 at Concord, Livermore, and San Jose were 0.81, 0.69, and 0.82, respectively.

Time series for the CMAQ simulation results are shown in Fig. 4 for 2000/01. This simulation period included four complete episodes of elevated $PM_{2.5}$ interspersed with relatively unpolluted conditions. Three episodes occurred under persistent R2 conditions, as determined by the clustering (see cluster labels on Fig. 3): 1–7 December; 26 December–8 January; and 17–22 January. As indicated in Fig. 3, each of these episodes exhibited varying degrees of the R2–R1 mismatch. Meteorological conditions for these episodes were often simulated correctly at lower EOF model orders; however, MM5 mistakenly simulated many of the R2 days as R1 at middle EOF model orders. Moreover, the R2–R1 mismatch occurred for multiple consecutive days during each episode. Therefore, as expected, CMAQ-simulated $PM_{2.5}$ levels were underestimated, and in many cases, severely so. The third persistent R2-type episode (17–22 January) exhibited the most severe R2–R1 mismatch, with mismatches occurring for most days at most EOF

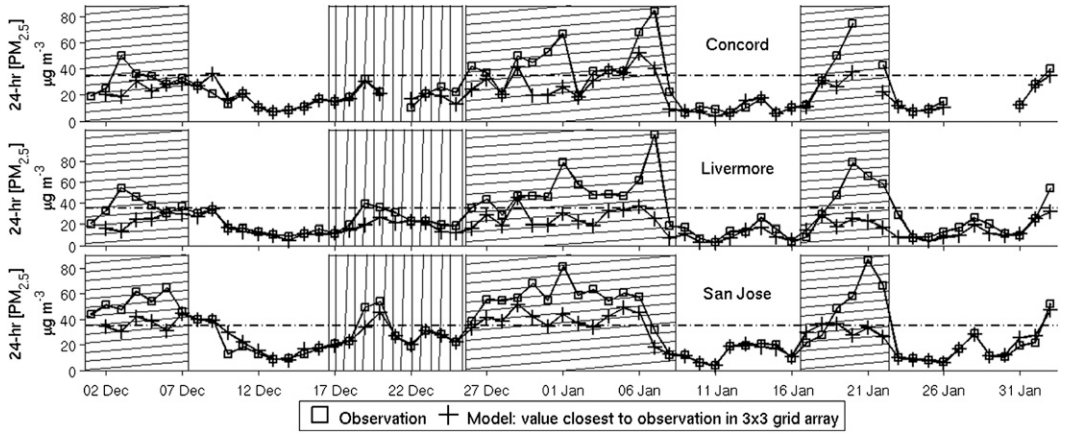

FIG. 4. Simulated (plus signs) and observed (squares) 24-h $PM_{2.5}$ levels at three SFBA monitoring locations for 2000/01 winter. Four highlighted episodes are of two classes: persistent R2 (diagonal hatch) and more transient R1 → R2 → R3 → R1 (vertical hatch). Horizontal lines are at 24-h $PM_{2.5}$ NAAQS exceedance threshold (35 μg m^{-3}).

model orders. CMAQ-simulated $PM_{2.5}$ levels were most severely underestimated for this episode. Also, the second persistent R2-type episode exhibited the R2–R1 mismatch at most scales during 31 December–2 January. Simulated $PM_{2.5}$ levels for this episode were more severely underestimated during this period of intensified R2–R1 mismatch as compared to the straddling periods having mismatches at fewer scales. During the persistent R2-type episodes, the simulated meteorological conditions were most accurate during 4–6 January. These days were correctly simulated as R2 for most EOF model orders, except for orders 13–14, which often resulted in inconsistent classifications. The downward bias in simulated $PM_{2.5}$ levels was less severe for most locations during this period.

Figure 5 provides a snapshot for a day exhibiting the R2–R1 mismatch. Simulated $PM_{2.5}$ levels and winds are shown for the central California modeling domain on R2 day 27 December 2000. In reality, this day had near-calm winds and high $PM_{2.5}$ levels throughout the CV. The simulation, however, produced winds that were too strong in the northern SV. The simulated surface airflow pattern (Fig. 5) most strongly resembled that of R1 (see Fig. 2).

Diagnosis of the R2–R1 mismatch focused on surface locations in the southern SV. Winds here were important for several reasons. First, clusters R1 and R2 were most strongly differentiated in the SV (see Fig. 2). Second, previous research has suggested that the complex SV surface flows are more sensitive to small changes in the large-scale pressure gradient driving flow through the Carquinez Strait than for the other central California basins (Bao et al. 2008). Third, the southern portion of the SV is connected with the SFBA, and direct pollutant exchange may have occurred here.

Figure 6 shows the time series for simulated and observed hourly wind speed and direction at Arbuckle (see Fig. 1) for the first two persistent R2-type episodes. Similar behavior was observed for the third persistent R2-type episode (not shown). The Arbuckle station was representative of the southwestern SV during these episodes. When the R2–R1 mismatch occurred, simulated wind speeds were too high. The model also did not appear to reproduce the timing of the observed wind speed minima that often occurred overnight. Additionally, the observations indicated diurnally shifting flows with overnight westerly winds. MM5 winds were persistently from the northwest. During 4–6 January, when the R2–R1 mismatch was minimal, the simulated Arbuckle winds tracked the observed winds reasonably well. A similar pattern appeared in the southeastern SV (not shown), except that the observed overnight flows were easterly.

A different type of episode developed over 17–25 December, during which the sequence R1 → R2 → R3 → R1 occurred. This more transient type of episode having evolving large-scale conditions over an 8-day period was reasonably well modeled by MM5. Except for the R3 days, which were almost never simulated properly, the lower-order classification labels matched the cluster labels. Mismatches occurred at middle EOF orders. CMAQ performance for this episode was less degraded than for the other episodes that occurred under persistent R2 conditions.

The mismatch in timing for the R2 → S transition observed to occur over 8–9 January appeared to significantly degrade CMAQ performance. As expected, simulated $PM_{2.5}$ levels at many locations began to decrease in advance of the observations. The $PM_{2.5}$ levels were severely underestimated on 7 January.

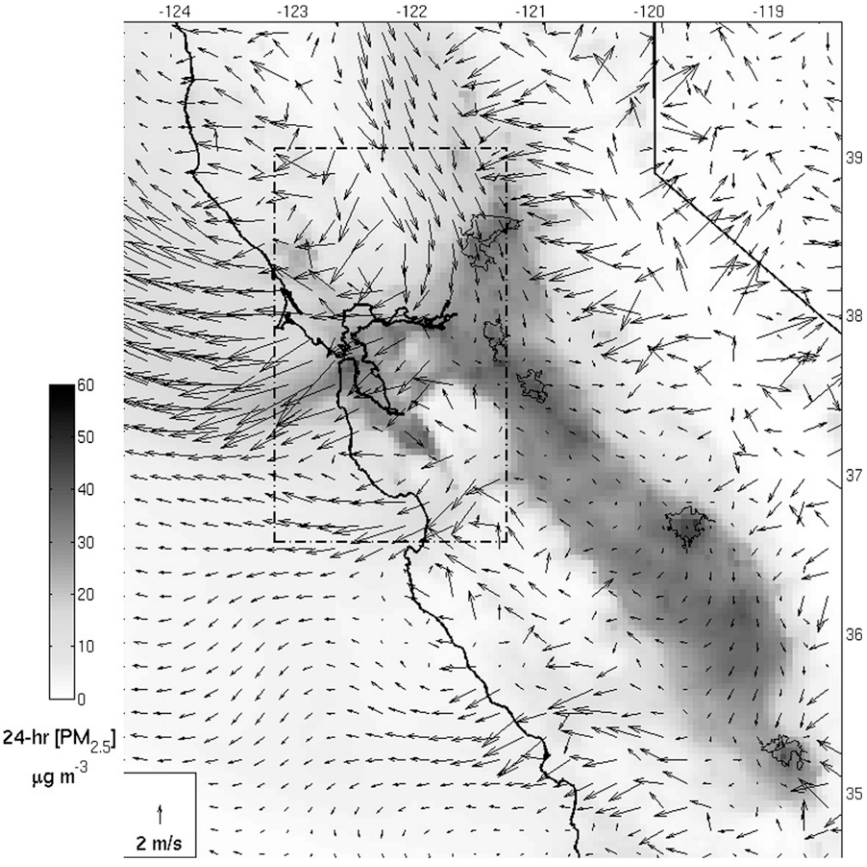

FIG. 5. Simulated surface layer 24-h $PM_{2.5}$ levels and 24-h wind field for 27 Dec 2000. Arrows point along direction of wind. The $PM_{2.5}$ levels are indicated by grayscale. California boundaries and CV municipal boundaries for Sacramento, Stockton, Modesto, Fresno, and Bakersfield (north to south) are shown for reference. The dashed box indicates the extent of Fig. 1.

6. Discussion

a. MM diagnosis

The predominant classification mismatches signaled a classic deficiency of MM5 to overemphasize the large-scale pressure gradient and underemphasize localized air flows (e.g., Hogrefe et al. 2001). MM5 generally had difficulty reproducing the conditions associated with weak synoptic forcing having an aloft ridge over the SFBA. The model was unable to produce R3, the pattern with the weakest synoptic forcing, regardless of actual conditions. The model also had difficulty reproducing R2, the pattern with the second-weakest synoptic forcing. Anticyclonic conditions (R1, R2, and R3) were generally simulated as R1, the anticyclonic pattern having the strongest synoptic forcing and moderate $PM_{2.5}$ levels. Here, R1, V, and S shared strong large-scale pressure gradients, regionally high winds, and lacked strong stability. The R1 days were usually simulated correctly. The cyclonic patterns (V and S) were typically simulated as cyclonic, although the V–S mismatch was common. The V–S mismatch was not very important for air quality applications because both patterns were windy and well ventilated; however, this finding may be important for precipitation applications because V is dry and S is rainy.

The clustering indicated that $PM_{2.5}$ episodes in the SFBA resulted upon transitions from cyclonic toward anticyclonic regimes. MM5 could distinguish between anticyclonic (R1, R2, and R3) and cyclonic (V and S) regimes having moderate-to-high and low $PM_{2.5}$ levels, respectively. Thus, the simulated meteorological conditions should be able to distinguish between days with moderate to high $PM_{2.5}$ levels and days with low $PM_{2.5}$ levels. These simulated fields would be expected to reproduce the timing of $PM_{2.5}$ episodes when used to drive an AQM. One exception would be the prematurely simulated R2 → S transition observed to occur over 8–9 January 2001. Mismatches among the anticyclonic weather patterns would be expected to result in appreciable downward biases for CMAQ-simulated peak $PM_{2.5}$ levels that occurred under episodic patterns R2 and R3.

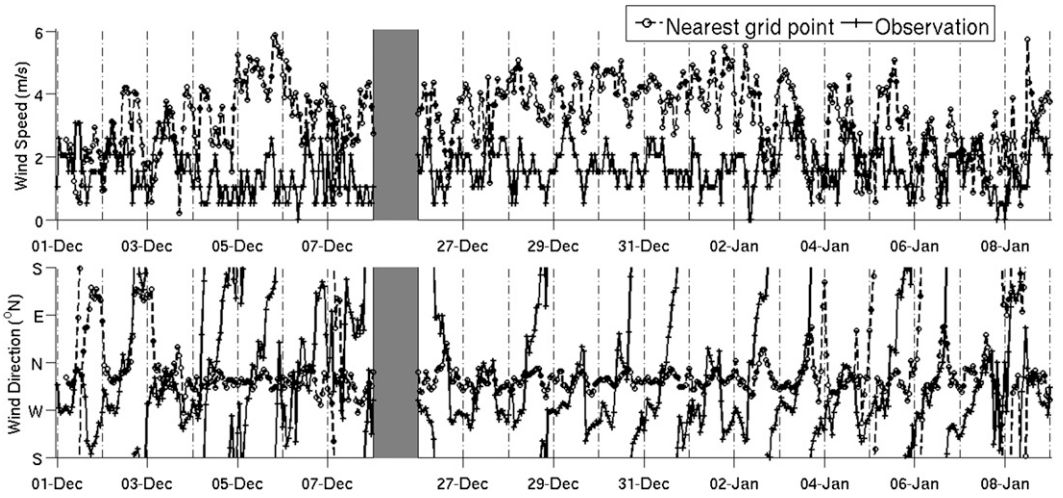

FIG. 6. Time series for observed (solid line with plus signs) and simulated (dashed line with circles) hourly 1-h (top) wind speed and (bottom) direction at Arbuckle (see Fig. 1). Two periods exhibiting R2–R1 mismatch from 2000/01 winter are separated by a gray patch: 1–7 Dec and 26 Dec–8 Jan. Hashed vertical lines appear at midnight PST beginning each day.

The R2–R1 mismatch was an important systematic bias identified for MM5 because the R2 conditions accounted for most SFBA $PM_{2.5}$ episodes. This systematic bias occurred only under certain conditions, so it was not obvious using traditional operational evaluation techniques. The observed R1 and R2 patterns were clearly distinguished by the EOF-based clustering; however, simple analyses of weather maps or surface wind fields may not have distinguished these conditions. The EOFs reflected atmospheric processes coherent with the surface wind field, and therefore helped to reveal three-dimensional features differentiating the related R1 and R2 patterns. The clustering of surface observations clearly indicated differences in the positions of the ridges aloft. For R1 the ridge was positioned offshore, whereas for R2 the ridge was positioned directly over the SFBA. MM5, on the other hand, appeared relatively insensitive to differences in the boundary conditions between R1 and R2. Both R1 and R2 were observed to produce persistent easterly surface winds through the SFBA. These airflow patterns appeared similar based on simple wind field analyses of the SFBA measurements. The EOF-based clustering, however, distinguished R1 as a relatively deep flow generated by the large-scale pressure gradient and R2 as a relatively shallow flow generated by terrain and surface heating effects.

The performance of MM5 was scale dependent, especially under conditions with pronounced terrain-induced airflow features. At the synoptic scale (lower-order EOFs), MM5 was able to reproduce the effects of the strong ridging pattern R2 about half of the time. Thus, the model was often able to replicate the bulk easterly SFBA surface air flows associated with the ridge. At the mesoscale (middle-order EOFs), however, MM5-simulated conditions that were not strongly conducive to $PM_{2.5}$ buildup. SFBA surface winds were correctly simulated as persistently from the east; however, the R2–R1 mismatch implied that wind speeds, mixing rates, and therefore overall pollutant dispersion rates were too high, especially in the SV. The scale dependency was less prevalent for the weather patterns with strong synoptic forcing. At lower orders, V was mistakenly simulated as R1 or S, the other patterns with strong synoptic forcing. At middle orders, however, V was only mistakenly simulated as S, the other pattern with westerly marine surface winds. At middle orders that reflect mesoscale influences, R1 and V were not confused because they exhibited opposite directions of bulk surface flow through the SFBA (easterly and westerly, respectively).

At the very lowest (1) and highest middle (13–14) EOF model orders, the model evaluation technique itself did not perform well. The first EOF typically represented 40%–50% of the variability in the MM5 output. The simulated conditions likely were insufficiently represented using this lone EOF. Classification using the highest middle-order EOFs (13–14) generated many labels that did not vary smoothly with EOF model order. These highest middle-order EOFs were likely explaining highly localized conditions that were not strongly connected to the organized flows that determined $PM_{2.5}$ source–receptor relationships. Also, these highest middle-order EOFs may have represented stochastic fluctuations and/or microscale structures in the ambient conditions that were not represented by MM5. The poor

performance of the model evaluation technique at extreme EOF model orders (1 and 13–14) represented effects of underfitting and overfitting, respectively, by the EOF models when classifying MM5 outputs.

b. AQM diagnosis

The predictive capability of the proposed MM–AQM evaluation technique provided a number of insights beyond those revealed by operational evaluation. As expected, episodic $PM_{2.5}$ levels were usually underestimated by CMAQ. This effect was pronounced for episodes occurring under persistent R2 conditions, as manifested by the R2–R1 mismatch. The scale dependency of MM5 to exhibit the R2–R1 mismatch appeared to explain the degree of degraded CMAQ performance. Periods with R2–R1 mismatches occurring across more EOF model orders exhibited poorer CMAQ performance. Also, longer durations for the R2–R1 mismatch produced larger downward biases for the simulated $PM_{2.5}$ levels. Presumably, a lack of simulated pollutant buildup helped cause these significantly underestimated $PM_{2.5}$ levels, especially during persistent R2 conditions. For a different type of episode with stronger synoptic forcing, CMAQ performed reasonably well. With a single exception, the timing of the episodes was reproduced accurately. A mistiming occurred for a storm observed to pass over 8–9 January 2001 that was simulated two days in advance. The otherwise accurate model timing was reflected by the high correlation coefficients between observed and simulated $PM_{2.5}$ levels, despite often severe biases.

During the R2–R1 mismatch, the simulated winds speeds in the SV were too high. Surface winds appeared to be reasonably well simulated within the SFBA; however, the inaccurately simulated conditions upwind of the SFBA in the SV appeared to considerably degrade CMAQ performance. Surface wind speeds in the SV were too high, suggesting artificially high simulated mixing rates under the stable and subsiding conditions. During these episodes, the SJV may also have been upwind of the SFBA. SJV winds appeared to be simulated reasonably well during the R2–R1 mismatch.

Beyond inaccurate wind speeds, the R2–R1 mismatch indicated that MM5 produced the wrong type of low-level airflow features. The observed overnight westerly flows at Arbuckle (Fig. 6) represented terrain-induced downslope (drainage) flows under clear-sky anticyclonic conditions. The diurnally shifting observed wind directions further evidenced the localized nature of this flow pattern. No aloft measurements were available over the SV; however, the observed overnight downslope flow pattern was presumably relatively shallow. A similar pattern along the eastern SV slopes also suggested overnight downslope flows. These observed downslope flows over the CV rims have been previously linked with SFBA exceedances (Beaver et al. 2010). MM5 winds, however, were persistently from the northwest. During the overnight hours, MM5 was producing low-level down-valley flows channeled along the SV major axis when in fact downslope flows converged toward the valley floor. The simulated northerly flows over the SV extended from the surface through the tenth model layer, or around 800 m AGL. The simulated down-valley flows were likely deeper and had higher mixing rates than the observed downslope flows. The model likely created too much dispersion in the SV, which subsequently affected downwind SFBA locations. Also, MM5 appeared to be unable to simulate the overnight calm conditions in the SV. This likely allowed for insufficient air mass aging in CMAQ, inhibiting buildup for dominant secondary $PM_{2.5}$ components such as ammonium nitrate. The underestimation of $PM_{2.5}$ levels during the R2–R1 mismatch may have also resulted from inaccurately simulated stability; R1 was far less stable than R2, allowing additional vertical dispersion of pollutants.

A second type of episode occurred under somewhat stronger synoptic forcing than episodes occurring under persistent R2 conditions. Localized terrain-induced flows were not as prevalent for this second type of episode. Thus, CMAQ performance was relatively improved because of the ability of MM5 to better handle these more synoptically forced surface air flows. Also, for episodes occurring under persistent R2 conditions, the R2 days with correct lower-order EOF classifications had somewhat improved CMAQ performance. One brief period (4–6 January 2001) having persistent R2 conditions was simulated correctly for both the lower- and most middle-order EOF classifications. MM5 reproduced diurnally shifting winds in the SV, and coupled MM5–CMAQ performance was better than for any other period exhibiting persistent R2 conditions. Nonepisode days typically occurred under patterns R1, V, and S. They had strong large-scale pressure gradients and reasonable performance for both the MM5 and CMAQ. The reasonable CMAQ performance to simulate the moderate $PM_{2.5}$ levels associated with R1 suggested that the emissions inventory and chemical mechanism were reasonably accurate. Thus, overall, the most important factor for explaining degraded CMAQ performance appeared to be the inability of MM5 to produce terrain-induced flows over the complex central California terrain during weak synoptic forcing events.

7. Conclusions

A pattern-based method for coupled MM–AQM evaluation has been developed. It was tested for $PM_{2.5}$

simulations over the SFBA for two winter $PM_{2.5}$ seasons. An EOF-based clustering of surface winds was performed using SFBA measurements. Five major weather patterns reflecting both synoptic and mesoscale variability impacting $PM_{2.5}$ levels were identified. MM5 outputs for the two winter seasons were classified among these five measurement-derived weather patterns. For each day of the simulation period, the labels for the observed winds and MM5-simulated winds were compared. The effects of the MM5 classification mismatches were used to diagnose degraded CMAQ performance.

In general, MM5 had difficulty reproducing the meteorological conditions associated with weak synoptic forcing events. CMAQ performance was especially degraded for episodes having persistent ridges of aloft high pressure over the study domain, leading to stagnating surface conditions. For such episodes, the model often incorrectly produced winds driven by the large-scale pressure gradient instead of by localized mechanisms. (This discrepancy was termed the R2–R1 mismatch.) A key shortcoming of MM5 appeared to be its inability to simulate overnight downslope flows over the complex central California terrain, especially in the SV. It was interesting to find that the CMAQ performance for the SFBA appeared to be limited by degraded MM5 performance in the upwind SV. Episodes having somewhat stronger synoptic forcing were better simulated by MM5, and CMAQ-estimated $PM_{2.5}$ levels were in closer agreement with observations. The timing of most episodes was properly simulated because MM5 could usually distinguish between anticyclonic and cyclonic conditions.

The above MM5 shortcoming is consistent with a well-known deficiency of this model. It typically provides too much synoptic push through complex terrain during periods of weak large-scale pressure gradients and light localized winds. The pattern-based evaluation technique was quite valuable to identify and diagnose the impact of this general MM5 shortcoming for a specific application. Identification of the MM5 bias would have been difficult using traditional methods. First, the meteorology-dependent bias was not obvious from operational evaluation statistics averaged across entire winter seasons. Second, the MM5 bias was not apparent using only local surface analyses. The systematic deficiency involved inaccurately simulated three-dimensional structures in the boundary layer. The evaluation of CMAQ performance based on MM5 performance was only possible because other CMAQ inputs such as emissions and chemistry appeared to be reasonable.

Identification and diagnosis of systematic model biases are critical for transferring knowledge between modelers and model developers. Such knowledge transfer is of paramount importance for collaboratively improving model performance to meet the needs of air quality planners.

REFERENCES

Ainslie, B., and D. G. Steyn, 2007: Spatiotemporal trends in episodic ozone pollution in the Lower Fraser Valley, British Columbia, in relation to mesoscale atmospheric circulation patterns and emissions. *J. Appl. Meteor. Climatol.,* **46,** 1631–1644.

Bao, J. W., S. A. Michelson, P. O. G. Persson, I. V. Djalalova, and J. M. Wilczak, 2008: Observed and WRF-simulated low-level winds in a high-ozone episode during the Central California Ozone Study. *J. Appl. Meteor. Climatol.,* **47,** 2372–2394.

Beaver, S., and A. Palazoglu, 2006a: Cluster analysis of hourly wind measurements to reveal synoptic regimes affecting air quality. *J. Appl. Meteor. Climatol.,* **45,** 1710–1726.

——, and ——, 2006b: A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay area. *Atmos. Environ.,* **40,** 713–725.

——, ——, A. Singh, S.-T. Soong, and S. Tanrikulu, 2010: Identification of weather patterns impacting 24-h average fine particulate matter pollution. *Atmos. Environ.,* **44,** 1761–1771.

Cannon, A. J., P. H. Whitfield, and E. R. Lord, 2002: Synoptic map-pattern classification using recursive partitioning and principal component analysis. *Mon. Wea. Rev.,* **130,** 1187–1206.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18.

Christakos, G., 2003: Critical conceptualism in environmental modeling and prediction. *Environ. Sci. Technol.,* **37,** 4685–4693.

Fine, J., L. Vuilleumier, S. Reynolds, P. Roth, and N. Brown, 2003: Evaluating uncertainties in regional photochemical air quality modeling. *Annu. Rev. Environ. Resour.,* **28,** 59–106.

Galin, M. B., 2007: Study of the low-frequency variability of the atmospheric general circulation with the use of time-dependent empirical orthogonal functions. *Izv. Atmos. Oceanic Phys.,* **43,** 15–23.

Gego, E., C. Hogrefe, G. Kallos, A. Voudouri, J. S. Irwin, and S. T. Rao, 2005: Examination of model predictions at different horizontal grid resolutions. *Environ. Fluid Mech.,* **5,** 63–85.

Gilliam, R. C., C. Hogrefe, and S. T. Rao, 2006: New methods for evaluating meteorological models used in air quality applications. *Atmos. Environ.,* **40,** 5073–5086.

Hanna, S. R., and R. Yang, 2001: Evaluations of mesoscale models' simulations of near-surface winds, temperature gradients, and mixing depths. *J. Appl. Meteor.,* **40,** 1095–1104.

Hogrefe, C., and Coauthors, 2001: Evaluating the performance of regional-scale photochemical modeling systems: Part I–meteorological predictions. *Atmos. Environ.,* **35,** 4159–4174.

Jain, A. K., 2000: Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.,* **22,** 4–37.

Jolliffe, I. T., 2002: *Principal Component Analysis*. 2nd ed. Springer-Verlag, 497 pp.

Li, S. T., and L. Y. Shue, 2004: Data mining and policy making in air pollution management. *Expert Syst. Appl.,* **27,** 331–340.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Scientific Rep. 1, Statistical Forecasting Project, Massachusetts Institute of Technology Defense Document Center 110268, 49 pp.

Ludwig, F. L., J. Y. Jiang, and J. Chen, 1995: Classification of ozone and weather patterns associated with high ozone concentrations in the San Francisco and Monterey Bay Areas. *Atmos. Environ.,* **29,** 2915–2928.

Mueller, S. F., 2009: Model representation of local air quality characteristics. *J. Appl. Meteor. Climatol.,* **48,** 945–961.

Rao, S. T., I. G. Zurbenko, R. Neagu, P. S. Porter, J. Y. Ku, and R. F. Henry, 1997: Space and time scales in ambient ozone data. *Bull. Amer. Meteor. Soc.,* **78,** 2153–2166.

Rife, D. L., and C. A. Davis, 2005: Verification of temporal variations in mesoscale numerical wind forecasts. *Mon. Wea. Rev.,* **133,** 3368–3381.

Rohli, R. V., M. M. Russo, A. J. Vega, and J. B. Cole, 2004: Tropospheric ozone in Louisiana and synoptic circulation. *J. Appl. Meteor.,* **43,** 1438–1451.

Russell, A., and R. Dennis, 2000: NARSTO critical review of photochemical models and modeling. *Atmos. Environ.,* **34,** 2283–2324.

Seaman, N. L., 2000: Meteorological modeling for air-quality assessments. *Atmos. Environ.,* **34,** 2231–2259.

Serrano, A., J. A. Garcia, V. L. Mateos, M. L. Cancillo, and J. Garrido, 1999: Monthly modes of variation of precipitation over the Iberian Peninsula. *J. Climate,* **12,** 2894–2919.

Shumway, R. H., and D. S. Stoffer, 2005: *Time Series Analysis and Its Applications*. Springer-Verlag, 571 pp.

Steyn, D. G., T. R. Oke, J. E. Hay, and J. L. Knox, 1981: On scales in meteorology and climatology. *McGill Climatol. Bull.,* **30,** 1–8.

Swall, J. L., and K. M. Foley, 2009: The impact of spatial correlation and incommensurability on model evaluation. *Atmos. Environ.,* **43,** 1204–1217.

Tesche, T. W., 2002: Operational evaluation of the MM5 meteorological model over the continental United States: Protocol for annual and episodic evaluation. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, 51 pp.

——, D. E. McNally, and J. G. Wilkinson, 2004: Evaluation of the 16-20 September 2000 ozone episode for use in 1-hr SIP development in the California Central Valley. California Air Resources Board, 93 pp.

Wilczak, J., J. Bao, S. Michelson, S. Tanrikulu, and S.-T. Soong, 2005: Simulation of an ozone episode during the Central California Ozone Study. Part I: MM5 meteorological model simulations. *Proc. 13th Conf. on the Application of Air Pollution Meteorology,* Pittsburgh, PA, Air and Waste Management Association, Paper J.2.1.

Wilke, C. K., 2003: Hierarchical models in environmental science. *Int. Stat. Rev.,* **71,** 181–199.

Yang, T., K. Matus, S. Paltsev, and J. Reilly, 2005: Economic benefits of air pollution regulation in the USA: An integrated approach. Rep. 113, revised January 2005, Massachusetts Institute of Technology, 29 pp.