RC22958 (W0311-002) November 3, 2003
Mathematics

IBM Research Report

A Hierarchical Framework for Modeling and Forecasting Web Server Workload

Ta-Hsin Li
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

A Hierarchical Framework for Modeling and Forecasting Web Server Workload

Ta-Hsin Li

Department of Mathematical Sciences

IBM T. J. Watson Research Center

Yorktown Heights, NY 10598 USA

e-mail: thl@watson.ibm.com

March 24, 2003

Abstract

Proactive management of web server farms requires accurate prediction of workload. An exemplary measure of workload is the amount of service requests per unit time. As a time series, the workload exhibits not only short-term random fluctuations but also prominent periodic (daily) patterns that evolve randomly from one period to another. A hierarchical framework with multiple time scales is proposed in this paper to model such time series. It leads to an adaptive procedure that provides both long-term (in days) and short-term (in minutes) predictions with simultaneous confidence (prediction) bands which accommodate not only serial correlation but also heavy-tailedness, heteroscedasticity, and nonstationarity of the data.

Key Words: Adaptive, autoregressive, computer network management, filterbank, heavy tail, heteroscedasticity, Internet, non-Gaussian, nonparametric, nonstationary, prediction, seasonal time series, web traffic.

I. INTRODUCTION

A web server farm is a cluster (or system) of servers shared by several web sites and maintained by a host service provider. Computing resources are dynamically (rather than statically) allocated according to the specific needs at different web sites. A service-level agreement is used to define the quality of service (QoS) for different classes of service requests and the revenue/cost model. In such an environment, accurate online prediction of server workload, with sufficiently long horizon, is key to ensuring that adequate computing resources are allocated *timely* to meet temporary increases of service requests of some classes at some sites while achieving certain system-wide performance objectives such as maintaining the QoS requirements of the entire server farm or maximizing the total revenue of the server farm under the QoS constraints [1]–[5].

A key measure of server workload is the amount of service requests per unit time (or the request arrival rate). It can be, for example, the total size or number of files requested per unit time; it can also be the total number of operations requested per unit time. A time series of such requests usually fluctuates randomly over time, with different statistical characteristics over different time windows in different time granularities: A time series of requests can be stationary but self-similar (long-range dependent) and/or heavy-tailed over a small time window (e.g., in seconds or minutes) at a fine time granularity (e.g., in milliseconds) [6]–[9]; it can also exhibit strong daily patterns and nonstationarity over a bigger time window (e.g., in days or weeks) at a coarser time granularity (e.g., in minutes), where the nonstationarity may occur during weekends and holidays as well as in different time periods of the day [8]–[11]. Whereas the first type of data has been intensively studied for the purpose of improving the performance of routing and scheduling (load balancing), the second type of data plays an important role in designing optimal strategies of dynamic resource allocation for the purpose of efficiently and smoothly managing web server farms. The focus of this paper is on the second type of data.

It has been shown that short-term prediction, with prediction horizons less than a few minutes, is useful in detecting anomalies of workload for early warning of service disruptions [12]–[14] and in mitigating temporary QoS deterioration by dynamic resource allocation [2]. In this paper, a hierarchical approach is developed to produce not only short-term prediction but also long-term prediction, with lead time equaling one day or more, by exploiting the daily patterns in the time

series of requests. Similar problems occur in other applications, such as electric utility management [15]–[21]. While the short-term prediction is needed to accommodate rapid fluctuations, the long-term prediction can be used to enhance the flexibility of dynamic resource management because it reduces the magnitude (hence complexity) of short-term adjustment and provides sufficient lead time for strategic planning.

Although several methods have been considered for workload (or traffic) modeling and/or prediction based on traditional techniques such as seasonal autoregressive integrated moving-average (SARIMA) [22]–[26], the method proposed in this paper has the following desirable properties:

1. It provides not only point predictions but also simultaneous confidence bands (joint prediction intervals) that can be used to support flexible service-level agreements (e.g., probability-based service guarantees) and the corresponding optimal strategies of resource allocation.

2. It has the capability of automatically handling nonstationarity in both daily patterns and short-term fluctuations because it allows the model parameters to change with time.

3. It is computationally efficient because it employs fast algorithms to compute the estimates of time-varying parameters online.

4. Its hierarchical approach allows easy diagnostic checking of model adequacy.

In the proposed method, the time series of requests is decomposed into a long-term component and a short-term component, where the former is represented by a linear combination of certain basis functions with random amplitudes and the latter by a traditional time series model. The basis functions in the long-term component are selected to compress the daily patterns into a low-dimensional space suitable for efficient modeling and computation, and the amplitudes of the basis functions are modeled as random processes to handle the trend and fluctuation of daily patterns. A simple multi-regime model is employed to deal with abrupt changes due to predetermined events such as weekdays and weekends. Online adaptive algorithms are employed to compute the estimates of time-varying parameters in both long-term and short-term models that are optimized specifically for the given prediction horizons. Key characteristics of service requests, such as the time-dependent variability (heteroscedasticity), the heavy-tailed non-Gaussian distribution,

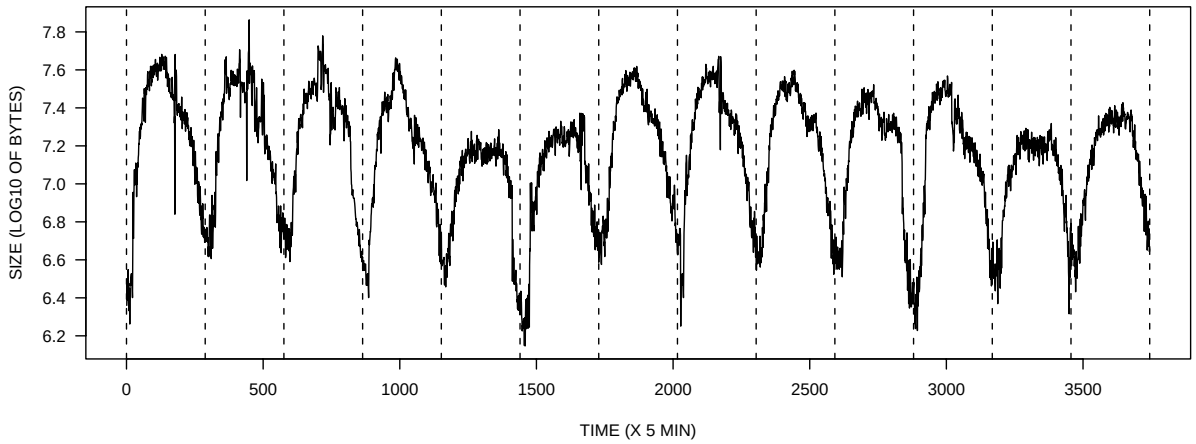


Fig. 1. *Total file sizes (in logarithm of bytes) of HTTP requests in five-minute intervals received at a web server of an online retail store over 13 days. Vertical dashed lines mark the daily boundary.*

and the residual serial dependence (correlation), are taken into account to construct simultaneous confidence bands for both long-term and short-term predictions.

II. Hierarchical Framework

Although the proposed method is applicable in more general situations, let us consider, for the simplicity of presentation, an exemplary server farm where service requests are measured at the server level. Assume that the requests are recorded at each server and aggregated in non-overlapping time intervals of length $\Delta > 0$ to form a time series. An example is shown in Fig. 1 where the time series comprises the logarithm of the total file sizes of Hypertext Transfer Protocol (HTTP) requests within every five-minute interval (i.e., $\Delta = 5$ minutes) over a period of 13 days at a commercial web site. This time series will be employed as a working example throughout the paper. Note that the log transform is employed here to eliminate, or reduce, the dependence of the (local) variability on the (local) mean of the untransformed file sizes (the former tends to rise with the later). In general, let $x_k(t) \in \mathbb{R}$ denote the amount of service requests measured at server k in time interval Δt ($k = 1, \cdots, s$; $t = 1, 2, \cdots$), where s is the total number of servers available.

As shown in Fig. 1, the time series contains prominent periodic (but not deterministic) daily patterns. This is a typical characteristic that exists at the web sites of a wide range of industries

(e.g., retail, finance, and insurance) and for many types of request measurements (e.g., total file sizes or file counts, total HTTP operations). The periodicity of such periodic patterns (the number of samples per period) will be denoted by p. For example, $p = 288$ for the daily patterns shown in Fig. 1. Note that once p is given the time index t can be expressed as $t = (\tau-1)p + r$ ($r = 1, \cdots, p$; $\tau = 1, 2, \cdots$), meaning that t is the rth time interval (of length Δ) in period τ.

Motivated by the exemplary time series in Fig. 1, let us consider the following model of $x_k(t)$ that handles both the periodic patterns and the more rapid fluctuations:

$$x_k(t) = y_k(t) + z_k(t), \quad y_k(t) := \sum_{j \in \mathscr{C}_k} \xi_{jk}(\tau)\, \phi_{jk}(t), \tag{1}$$

where, for each given k, $\mathscr{C}_k$ is a subset of $\mathscr{I}_p := \{1, \cdots, p\}$, the $\phi_{jk}(t)$ are p-periodic functions of t such that $\boldsymbol{\phi}_{jk} := \mathrm{vec}\{\phi_{jk}(t)\}_{t=1}^{p}$ ($j = 1, \cdots, p$) form a basis (not necessarily orthogonal) of $\ell_2(\mathbb{R}^p)$, the $\xi_{jk}(\tau)$ are random variables, and $z_k(t)$ is a random process with zero mean.

In this model, the periodic, long-term, patterns are represented by $y_k(t)$ as a linear combination of a subset of basis functions whose amplitudes fluctuate randomly from one period to another; the remaining intra-period, short-term, fluctuations are represented by $z_k(t) = x_k(t) - y_k(t)$ as a zero-mean random process. A general discussion on this model and its comparison with traditional models such as SARIMA and periodic AR (PAR) can be found in [27].

Although more sophisticated models (e.g., general state-space models and nonlinear dynamic models) are possible alternatives, it is assumed in this paper that $\boldsymbol{\xi}(\tau) := \mathrm{vec}\{\xi_{jk}(\tau)\colon\ j \in \mathscr{C}_k, k = 1, \cdots, s\}$ and $\mathbf{z}(t) := \mathrm{vec}\{z_k(t)\}_{k=1}^{s}$ are both AR processes, not only because the AR models can approximate more complicated linear models but also because the optimal AR parameters required for prediction can be easily obtained by computationally efficient algorithms.

If $\boldsymbol{\xi}(\tau)$ is a stationary process, then the long-term component $y_k(t)$ is a cyclostationary process with a p-periodic mean function, i.e., it satisfies

$$E\{y_k(t+p)\} = E\{y_k(t)\} \ \text{ and } \ \mathrm{Cov}\{y_k(t+p), y_k(t'+p)\} = \mathrm{Cov}\{y_k(t), y_k(t')\}$$

for all t and t'. This constraint on $y_k(t)$ can be relaxed for greater flexibility by allowing the parameters of $\boldsymbol{\xi}(\tau)$ to change with time. For example, by allowing the mean of $\boldsymbol{\xi}(\tau)$ to vary with τ, one can introduce in $y_k(t)$ a periodic or nonperiodic trend that evolves more slowly relative to p, even though the model in (1) does not explicitly contain an additive term to represent such a trend

(e.g., a linear growth function or an annual cycle). Similarly, certain nonstationary characteristics in the short-term fluctuation can be handled by allowing the parameters of $\mathbf{z}(t)$ to change over time.

Once the parametric model structure is specified, it is possible, at least in theory, to estimate the unknown parameters jointly from observations of $x_k(t)$ by, for example, the Gaussian maximum likelihood method or the Bayesian method. A drawback of this joint estimation approach is that the computation of the parameter estimates can be difficult, even under the Gaussian and stationary assumptions.

In this paper, a simpler, hierarchical, approach is taken because it requires less computation and is easier for diagnostic checking. In the hierarchical approach, the long-term and short-term components in (1) are dealt with separately in a hierarchy: the long-term component $y_k(t)$ is considered first under some simplified working assumptions about the short-term component $z_k(t)$; after the long-term component is modeled and predicted, the short-term component is modeled on the basis of the long-term prediction error. Moreover, adaptive algorithms are employed to track the model parameters, which may change slowly with time, without explicitly modeling them.

It is not unusual that n a time series of service requests the daily patterns of weekdays behave significantly differently from the daily patterns of weekends, as is shown clearly in Fig. 1. One way of handling such a phenomenon of multiple regimes, in which the time of regime shift is known *a priori*, is to consider the weekly periodicity instead of the daily one; but in doing so, the computational cost will grow exponentially with the increase of p. Another possibility is to model $\boldsymbol{\xi}(\tau)$ as an AR process in which the parameters take different values when τ belongs to different regimes; however, unless the order of the AR model is very high (i.e., higher than the maximum run-length of the regimes), this method has the disadvantage of possibly relying solely on the immediate observations of the previous regime to predict a new regime, as it happens when predicting Saturday's requests with an AR(1) model for $\boldsymbol{\xi}(\tau)$, while in fact the data in the new regime are more resemblant to the observations of the same regime many periods older (e.g., observations from the last weekend).

In this paper, different regimes are modeled separately to handle the between-regime change while taking advantage of the within-regime resemblance. It is done by cascading the original data that belong to the same regime to form a set of new time series, one for each regime. For example, all weekday observations shown in Fig. 1 are collected to form a new weekday time series and

all weekend observations are collected to form a new weekend time series. Each time series is modeled by (1) with a regime-dependent long-term component, some parameters of which, e.g., the mean of $\xi_{jk}(\tau)$ for some values of j, may be shared among all regimes to ensure a smoother transition when the regime shifts. Long-term predictions are made by using one of the regime-specific models, depending on which regime the prediction target belongs to. The same method can be applied to the short-term modeling and prediction. However, it is more convenient, and in many cases justifiable, to merge the short-term components of all regimes in the original order to form a single time series for modeling and prediction under the assumption that the combined component does not change significantly with regime shift.

The remainder of the paper is devoted to detailed exposition of model identification, estimation, and forecasting steps. The case of a single server $(s = 1)$ is considered for simplicity. In this case, the subscript k can be dropped so that (1) becomes

$$x(t) = y(t) + z(t), \quad y(t) = \sum_{j \in \mathscr{C}} \xi_j(\tau)\, \phi_j(t). \tag{2}$$

Extension of the methodology to the case of $s > 1$ is straightforward and therefore is omitted.

III. Long-Term Modeling and Prediction

Since multiple regimes are handled separately for long-term modeling and prediction, it suffices to consider the case of a single regime and assume that the statistical properties of $y(t)$ and $z(t)$ in (2) do not change abruptly with regime shift. Assume further that the basis functions $\phi_j(t)$ in (2) are given *a priori* (e.g., sinusoids, wavelets, polynomials, etc., depending on the particular type of daily patterns) and the subset $\mathscr{C}$ does not change with time. In the following, two remaining issues are discussed: (a) determination of $\mathscr{C}$ from historical observations of $x(t)$ and (b) long-term prediction of $x(t)$ via modeling and prediction of the $\xi_j(\tau)$.

A. Determination of $\mathscr{C}$

The purpose of selecting a subset $\mathscr{C} \subset \mathscr{I}_p$ is two fold. First, it is to reduce the dimensionality and hence the required computational and data resources in the subsequent procedure of modeling and prediction. Second, it is to reduce the sensitivity to estimation errors and ensure more robust

modeling and prediction. The first point is obvious because the higher the dimension of

$$\boldsymbol{\xi}(\tau) := \text{vec}\{\xi_j(\tau),\, j \in \mathscr{C}\}. \tag{3}$$

the more parameters are there to be estimated. The need for dimension reduction is especially prominent in web server management, as compared to, for example, electric utility management [15]–[21] or economic forecasting [28] [29], because it operates on a much smaller time granularity (in the magnitude of minutes rather than hours or months) which results in a very large value of p (e.g., $p = 288$ for five-minutely data and $p = 1440$ for minutely data, as compared to $p = 24$ for hourly data and $p = 12$ for monthly data). The second point is justified by the statistical theory of regression analysis [30, pp. 268–269] which asserts that in the presence of inherent statistical error in parameter estimation, the mean-square error of both modeling and prediction can be reduced by simply ignoring the "minor" components even if they exist in reality.

To select $\mathscr{C}$, the filterbank approach proposed in [27] is readily applicable. First, consider the "complete" model that includes all basis functions. In this case, it is obvious from (2) that

$$\boldsymbol{\xi}_0(\tau) := \text{vec}\{\xi_j(\tau)\}_{j=1}^{p} = \boldsymbol{\Phi}_0^{-1}\mathbf{x}(\tau) \quad (\tau = 1, 2, \cdots), \tag{4}$$

where $\boldsymbol{\Phi}_0 := [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_p]$ is the complete matrix of basis functions and

$$\mathbf{x}(\tau) := \text{vec}\{x((\tau-1)p + r)\}_{r=1}^{p}$$

is the vector of observations obtained in period τ. Eq. (4) defines a transform from $x(t)$ to $\xi_j(\tau)$ that has a filterbank interpretation. Indeed, letting $\boldsymbol{\psi}_j' := [\psi_j(p-1), \cdots, \psi_j(0)]$ be the jth row of $\boldsymbol{\Phi}_0^{-1}$ so that $\boldsymbol{\Phi}_0^{-1} = [\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_p]'$, Eq. (4) can be expressed as

$$\xi_j(\tau) = \sum_{i=0}^{p-1} \psi_j(i)\, x(\tau p - i) \quad (j = 1, \cdots, p;\ \tau = 1, 2, \cdots), \tag{5}$$

which shows that $\xi_j(\tau)$ can be obtained by applying to $x(t)$ a filterbank that consists of p finite impulse response (FIR) filters, with $\{\psi_j(i)\}_{i=0}^{p-1}$ being the impulse response of the jth filter, and subsampling (decimating) the output of the jth filter at time $t = \tau p$ (i.e., one sample per period taken at the end of each period).

With $\xi_j(\tau)$ so defined, $x(t)$ can be expressed as

$$x(t) = \sum_{j=1}^{p} w_j(t), \quad w_j(t) := \xi_j(\tau)\, \phi_j(t), \tag{6}$$

for $t := (\tau-1)p+r$. In this expression, $x(t)$ is decomposed into p component waveforms $w_j(t)$ $(j = 1,\cdots,p)$; the amplitude of these waveforms is constant within each period and is varying randomly from one period to another; the basic shape of these waveforms are determined by the basis functions $\phi_j(t)$.

Since the long-term model will be employed for long-term prediction, a proper subset $\mathscr{C}$ should be selected by examining the long-term behavior of $\xi_j(\tau)$. Given n periods of historical observations $\{\mathbf{x}(\tau)\}_{\tau=1}^n$, one can obtain $\{\boldsymbol{\xi}_0(\tau)\}_{\tau=1}^n$ from (4) or (5). The long-term effect of $\xi_j(\tau)$ as a function of τ can be quantified by the *coherence* measure [31]

$$\hat{c}_j := \frac{\hat{\mu}_j^2}{\hat{\mu}_j^2+\hat{\sigma}_j^2}, \tag{7}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the sample mean and variance of $\{\xi_j(\tau)\}_{\tau=1}^n$ defined as

$$\hat{\mu}_j := n^{-1}\sum_{\tau=1}^n \xi_j(\tau), \quad \hat{\sigma}_j^2 := n^{-1}\sum_{\tau=1}^n \{\xi_j(\tau)-\hat{\mu}_j\}^2. \tag{8}$$

Of the p component waveforms of $x(t)$ defined by (6), some are characterized by high coherence, i.e., large value of $\hat{c}_j$; these waveforms, called the high-C components of $x(t)$, are good candidates for use in the long-term prediction because they have long-lasting effect due to the fact that for a high-C component, the corresponding $\hat{\mu}_j$, as an estimator of the long-term average

$$\mu_j := \lim_{n\to\infty} n^{-1}\sum_{\tau=1}^n \xi_j(\tau),$$

is significantly different from zero. A statistical test was suggested in [27] to help identify the high-C components. In addition, one may also consider the component waveforms of $x(t)$ that have large amplitude. These waveforms can be identified by the *energy* measure

$$\hat{e}_j := n^{-1}\sum_{\tau=1}^n \{\xi_j(\tau)\}^2 = \hat{\mu}_j^2+\hat{\sigma}_j^2 \tag{9}$$

and may be referred to as the high-E components of $x(t)$. The subset $\mathscr{C}$ should contain both high-C and high-E components in order to effectively model the inter-period dynamics of $x(t)$.

As an example, Fig. 2 shows the plot of $\hat{c}_j$ and $\hat{e}_j$ associated with the filterbank of discrete Fourier transform (DFT) for the weekday data in Fig. 1. Note that the DFT-based $\hat{e}_j$ is just the average of the periodograms of $\mathbf{x}(\tau)$ $(\tau = 1,\cdots,n)$. As Fig. 2 suggests, the subset $\mathscr{C}$ should contain

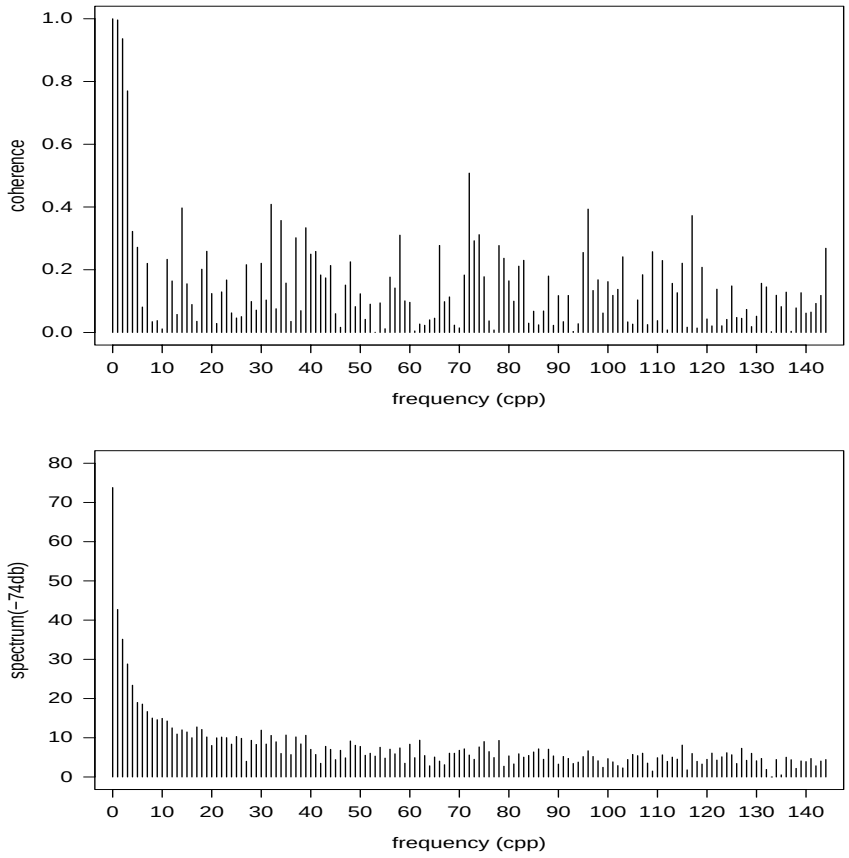


Fig. 2. *Fourier coherence (top) and periodogram (bottom) of the weekday data in Fig. 1, where cpp stands for cycles per period.*

at least the mean (zero frequency) and the first three frequencies (the fundamental frequency and its first two harmonics). Since each nonzero frequency corresponds to two coefficients, one for sine and one for cosine, the dimensionality of $\mathscr{C}$ and $\boldsymbol{\xi}(\tau)$ is equal to 7. This represents a tremendous compression of the original dimension $p = 288$.

One may also determine $\mathscr{C}$ by employing various model-selection techniques developed for regression analysis [30] [32] under the working assumption of white noise. Once $\mathscr{C}$ is determined, the $\xi_j(\tau)$ ($j \in \mathscr{C}$) can be simply obtained from (4) or (5) as before. Alternatively, they can be obtained by least squares (LS). Clearly, the LS solution coincides with that defined by (5) if the basis functions are orthonormal.

B. Modeling of $\xi_j(\tau)$ for Long-Term Prediction

Because Eq. (4) defines an equivalence relationship between $\boldsymbol{\xi}_0(\tau)$ and $\mathbf{x}(\tau)$, the d-period-ahead (long-term) prediction of $\{x(t)\}$ at time $t = np$ can be obtained from the d-step-ahead prediction of $\{\boldsymbol{\xi}_0(\tau)\}$ at $\tau = n$, i.e., $\hat{\mathbf{x}}(n+d\,|\,n) = \boldsymbol{\Phi}_0 \hat{\boldsymbol{\xi}}_0(n+d\,|\,n)$, or equivalently,

$$\hat{x}((n+d-1)p+r\,|\,np) = \sum_{j=1}^{p} \hat{\xi}_j(n+d\,|\,n)\,\phi_j(r) \quad (r = 1, \cdots, p).$$

This predictor can be simplified as

$$\hat{x}((n+d-1)p+r\,|\,np) = \sum_{j \in \mathscr{C}} \hat{\xi}_j(n+d\,|\,n)\,\phi_j(r) \quad (r = 1, \cdots, p) \tag{10}$$

under the assumption that $\boldsymbol{\xi}_1(\tau) := \text{vec}\{\xi_j(\tau),\, j \notin \mathscr{C}\}$ is a zero-mean white noise process and is statistically uncorrelated with the process $\boldsymbol{\xi}(\tau)$ in (3). Since $\mathscr{C}$, by construction, contains all components of $\mathbf{x}(\tau)$ that are significantly relevant to long-term prediction, the simplified predictor (10) should be able to capture the periodic patterns in $x(t)$ as long as $\boldsymbol{\xi}(\tau)$ is well predicted. Moreover, if the $\phi_j(t)$ are chosen judiciously, the dimension of $\boldsymbol{\xi}(\tau)$, denoted by m, is much smaller than p. Therefore, the simplification enables us to focus on the modeling and forecasting of a low-dimension vector process and thus considerably reduces the complexity of the problem. The following is devoted to the simplified predictor (10) and the modeling of $\boldsymbol{\xi}(\tau)$.

The purpose of modeling $\boldsymbol{\xi}(\tau)$ is to derive optimal predictions. It is desirable that the model be easy to identify, estimate, and derive predictions from. Although standard AR models satisfy

these requirements, it is advantageous, as explained later, to employ the *horizon-specific multistep AR*(d,κ) *model* of the form

$$\boldsymbol{\xi}(\tau) = \boldsymbol{\mu} + \sum_{i=1}^{\kappa} \mathbf{A}_i \{\boldsymbol{\xi}(\tau - d - i + 1) - \boldsymbol{\mu}\} + \boldsymbol{\varepsilon}(\tau), \quad \{\boldsymbol{\varepsilon}(\tau)\} \sim \mathrm{IID}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{11}$$

where $d \geq 1$ is the given horizon of prediction and κ is the order of the model. Under this assumption, the best (linear) prediction of $\boldsymbol{\xi}(n+d)$ given $\{\boldsymbol{\xi}(\tau)\}_{\tau=1}^{n}$ $(n > \kappa)$ can be expressed as

$$\hat{\boldsymbol{\xi}}(n+d \,|\, n) := \boldsymbol{\mu} + \sum_{i=1}^{\kappa} \mathbf{A}_i \{\boldsymbol{\xi}(n - i + 1) - \boldsymbol{\mu}\} \tag{12}$$

and the covariance matrix of the prediction error $\boldsymbol{\varepsilon}(n+d) = \boldsymbol{\xi}(n+d) - \hat{\boldsymbol{\xi}}(n+d \,|\, n)$ is equal to $\boldsymbol{\Sigma}$. This, coupled with the assumption that $\{\boldsymbol{\xi}_1(\tau)\}$ is white noise and is uncorrelated with $\{\boldsymbol{\xi}(\tau)\}$, implies that Eq. (10) gives the best (linear) prediction of $\mathbf{x}(n+d)$ based on $\{\mathbf{x}(\tau)\}_{\tau=1}^{n}$.

Note that without the white-noise assumption of $\boldsymbol{\xi}_1(\tau)$, Eq. (10) only gives the best prediction of the long-term component $\mathbf{y}(n+d) := \mathbf{x}(n+d) - \mathbf{z}(n+d)$. Therefore, the white-noise assumption is equivalent to the assumption that $z(t)$ does not have long-term dependence. In practice, it suffices that the sample autocovariances of $\{z(t)\}_{t=1}^{np}$ become negligible when the lag of correlation exceeds $p-1$. Checking this condition is another way of assessing the adequacy of the long-term model.

The parameters in (11) need to be estimated from historical data in order to compute the predictions given in (12) and (10). Let the historical data be available up to time $t = np$. Then, $\boldsymbol{\mu}$ can be estimated simply by the sample mean

$$\hat{\boldsymbol{\mu}} := n^{-1} \sum_{\tau=1}^{n} \boldsymbol{\xi}(\tau). \tag{13}$$

With $\hat{\boldsymbol{\mu}}$ in place of $\boldsymbol{\mu}$, $\mathbf{A} := [\mathbf{A}_1, \cdots, \mathbf{A}_\kappa]$ can be estimated by the LS method that minimizes $\sum_{\tau=1}^{n} \boldsymbol{\varepsilon}(\tau)' \boldsymbol{\varepsilon}(\tau)$. Note that the LS estimator also minimizes $\sum_{\tau=1}^{n} \boldsymbol{\varepsilon}(\tau)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}(\tau)$ for any given $\boldsymbol{\Sigma}$ [33, p. 76] and can be expressed as

$$\hat{\mathbf{A}} := \boldsymbol{\Xi} \boldsymbol{\Gamma}^{-1}, \tag{14}$$

where

$$\boldsymbol{\Xi} := \sum_{\tau=1}^{n} \tilde{\boldsymbol{\xi}}(\tau) \boldsymbol{\gamma}(\tau)', \quad \boldsymbol{\Gamma} := \sum_{\tau=1}^{n} \boldsymbol{\gamma}(\tau) \boldsymbol{\gamma}(\tau)', \tag{15}$$

$$\boldsymbol{\gamma}(\tau) := \mathrm{vec}\{\tilde{\boldsymbol{\xi}}(\tau - d - i + 1)\}_{i=1}^{\kappa},$$

with $\tilde{\boldsymbol{\xi}}(\tau) := \boldsymbol{\xi}(\tau) - \hat{\boldsymbol{\mu}}$. Given $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{A}}$, $\boldsymbol{\Sigma}$ can be estimated by $\hat{\boldsymbol{\Sigma}} := n^{-1}\sum_{\tau=1}^{n}\hat{\boldsymbol{\varepsilon}}(\tau)\hat{\boldsymbol{\varepsilon}}(\tau)'$, where $\hat{\boldsymbol{\varepsilon}}(\tau) := \tilde{\boldsymbol{\xi}}(\tau) - \hat{\mathbf{A}}\boldsymbol{\gamma}(\tau) = \tilde{\boldsymbol{\xi}}(\tau) - \sum_{i=1}^{\kappa}\hat{\mathbf{A}}_i\tilde{\boldsymbol{\xi}}(\tau-d-i+1)$.

As in ordinary AR models [33, p. 129], the order κ in (11) can also be chosen to minimize $\mathrm{AIC}(\kappa) := -2\log L(\kappa) + 2(\kappa m^2+1)$, where

$$L(\kappa) := (2\pi)^{-mn/2}|\hat{\boldsymbol{\Sigma}}|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{\tau=1}^{n}\hat{\boldsymbol{\varepsilon}}(\tau)'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}(\tau)\right\}$$

is the Gaussian likelihood. Since $\sum_{\tau=1}^{n}\hat{\boldsymbol{\varepsilon}}(\tau)'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\varepsilon}}(\tau) = mn$, it follows that

$$\mathrm{AIC}(\kappa) = n\log|\hat{\boldsymbol{\Sigma}}| + 2(\kappa m^2+1) + \{1+\log(2\pi)\}mn. \tag{16}$$

It is clear that AIC attempts to balance the goodness of fit, measured by $n\log|\hat{\boldsymbol{\Sigma}}|$, and the complexity of the model, measured by $2(\kappa m^2+1)$, which penalizes high-order models. Under the assumption that $\boldsymbol{\xi}(\tau)$ is, in reality, a nondegenerate AR(∞) process, it can be shown [34] [35] that AIC is asymptotically efficient in selecting κ for optimal d-step-ahead linear prediction (in the sense of minimum mean-square error). Several alternative criteria for order selection, including FPE, AICC, and BIC, are given in [33, Sec. 4.3] and [36, Sec. 9.3]. They are distinguished from AIC by the employment of different penalty functions. To check for model adequacy, one may analyze $\hat{\boldsymbol{\varepsilon}}(\tau)$ for possible serial correlation [36, Sec. 9.4].

Note that if $\boldsymbol{\xi}(\tau)$ in reality is an ordinary AR(κ) process satisfying

$$\boldsymbol{\xi}(\tau) = \boldsymbol{\mu} + \sum_{i=1}^{\kappa}\mathbf{A}_i\{\boldsymbol{\xi}(\tau-i)-\boldsymbol{\mu}\} + \boldsymbol{\varepsilon}(\tau), \quad \{\boldsymbol{\varepsilon}(\tau)\} \sim \mathrm{IID}(\mathbf{0},\boldsymbol{\Sigma}), \tag{17}$$

then, instead of (12), the best d-step-ahead prediction should be

$$\hat{\boldsymbol{\xi}}(n+d\,|\,n) = \boldsymbol{\mu} + \sum_{i=1}^{\kappa}\mathbf{A}_i\{\hat{\boldsymbol{\xi}}(n+d-i\,|\,n)-\boldsymbol{\mu}\}, \tag{18}$$

where $\hat{\boldsymbol{\xi}}(n+d-i\,|\,n) := \boldsymbol{\xi}(n+d-i)$ for $i \geq d$. Both being linear predictors, the major difference between (18) and (12) lies in the way in which the parameters are optimized in practice. Assuming that both models are estimated by LS, it is easy to see that the $\mathbf{A}_i$ in (18) are optimized by minimizing the *one-step* prediction error while the d-step prediction is derived from forward projections of the assumed AR model (17) with zero forcing. On the other hand, the $\mathbf{A}_i$ in (12) are optimized by *directly* minimizing the d-step prediction error without using model-based projections. When the AR model (17) is incorrect, as will be most likely the case in practice, the model-based projections

can result in poor d-step prediction for $d > 1$ even though the model is optimized for one-step prediction. On the other hand, the d-step predictions given by (12) tend to be more reliable (e.g., more robust to modeling errors), as demonstrated in many analytical and empirical studies [37]–[43].

The model (11), which is for $\boldsymbol{\xi}(\tau)$, can be transformed into a model for $y(t)$. As an example, consider the case where $d = 1$ and $\mathscr{C} = \{1, \cdots, m\}$ for some $0 < m < p$. In this case, one can write $\mathbf{y}(\tau) = \boldsymbol{\Phi}\boldsymbol{\xi}(\tau)$ and $\boldsymbol{\xi}(\tau) = \boldsymbol{\Psi}\mathbf{y}(\tau)$, where $\boldsymbol{\Phi} := [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m]$ and $\boldsymbol{\Psi} := [\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_m]'$. If $\boldsymbol{\xi}(\tau)$ satisfies (11) with $d = 1$, then $\mathbf{y}(\tau)$ must satisfy

$$\mathbf{y}(\tau) = \boldsymbol{\nu} + \sum_{i=1}^{\kappa} \mathbf{B}_i \mathbf{y}(\tau - i) + \boldsymbol{\zeta}(\tau), \quad \{\boldsymbol{\zeta}(\tau)\} \sim \mathrm{IID}(\mathbf{0}, \boldsymbol{\Lambda}), \tag{19}$$

where

$$\boldsymbol{\nu} := \boldsymbol{\Phi}\left(\mathbf{I} - \sum_{i=1}^{\kappa} \mathbf{A}_i\right)\boldsymbol{\mu}, \quad \mathbf{B}_i := \boldsymbol{\Phi}\mathbf{A}_i\boldsymbol{\Psi}, \quad \boldsymbol{\Lambda} := \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}'. \tag{20}$$

Therefore, $\mathbf{y}(\tau)$ is a special AR(κ) process in which the parameters $\boldsymbol{\nu}$, $\mathbf{B}_i$, and $\boldsymbol{\Lambda}$ are constrained by (20) in lower-dimensional manifolds.

An alternative long-term model, which is widely used in econometrics, is the seasonal random walk [28, p. 143] [39]

$$y(t) = \nu_0 + \tilde{y}(t), \quad \sum_{i=0}^{p-1} \tilde{y}(t - i) = \zeta(t), \quad \{\zeta(t)\} \sim \mathrm{IID}(0, \sigma_\zeta^2). \tag{21}$$

Note that if $y(t)$ is a deterministic p-periodic function, then it would satisfy (21) with $\nu_0 := p^{-1}\sum_{i=1}^{p} y(t)$ and $\zeta(t) \equiv 0$. In general, a random perturbation $\zeta(t)$ is employed in (21) to model the fluctuation of the periodic patterns. To compare (21) with (19), it suffices to note that (21) can be rewritten in the form of (19), with $\kappa = 1$ and with

$$\boldsymbol{\nu} := \nu_0(\mathbf{I} - \mathbf{B}_1)\mathbf{1}, \quad \mathbf{B}_1 := \begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{22}$$

$$\boldsymbol{\Lambda} := \sigma_\zeta^2 \begin{bmatrix} 1 & & \text{symmetric} & \\ -1 & 2 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & -1 & 2 \end{bmatrix}. \tag{23}$$

As can be seen, although the random-walk model (21) does not make explicit assumptions on the basic shape of the periodic patterns, it is, in effect, less flexible than the model in (19)–(20) because

all, except two, parameters in (22)–(23) are fully specified in advance, whereas more parameters in (19)–(20), including κ, are allowed to be determined from the data.

C. Adaptive Estimation

The LS estimator is preferred over the joint maximum likelihood estimator described in [36, Sec. 8.7] because (a) the LS estimator can be easily modified for adaptive estimation and (b) the LS estimator can be easily computed by fast online algorithms that simply update the current value upon arrival of a new observation rather than repeatedly solving the optimization problem using the entire set of historical data.

One way of modifying the LS estimator for adaptive estimation is to put more emphasis on the newer data by replacing the constant weight n^{-1} in (13) with an exponential weight function $c(\lambda,n)\,\lambda^{n-\tau}$ $(\tau=1,\cdots,n)$, where $\lambda\in(0,1]$ and

$$c(\lambda,n):=\begin{cases}1-\lambda & \text{if } 0<\lambda<1,\\ n^{-1} & \text{if } \lambda=1.\end{cases}$$

The resulting estimator can be computed online by the recursive algorithm

$$\hat{\boldsymbol{\mu}}(\tau)=d(\lambda,\tau)\,\hat{\boldsymbol{\mu}}(\tau-1)+c(\lambda,\tau)\,\boldsymbol{\xi}(\tau)\quad(\tau=1,\cdots,n) \tag{24}$$

with the initial value $\hat{\boldsymbol{\mu}}(0):=\mathbf{0}$, where $d(\lambda,\tau):=\lambda\,c(\lambda,\tau)/c(\lambda,\tau-1)=1-c(\lambda,\tau)$ can be interpreted as a discount factor of the previous estimator $\hat{\boldsymbol{\mu}}(\tau-1)$ and $c(\lambda,\tau)$ can be regarded as a contribution factor of the new observation $\boldsymbol{\xi}(\tau)$. For $\lambda=1$, $\hat{\boldsymbol{\mu}}(n)$ becomes the sample mean in (13), and for $0<\lambda<1$, it is known as the exponentially weighted moving-average (EWMA) estimator of $\boldsymbol{\mu}(n):=E\{\boldsymbol{\xi}(n)\}$. Note that EWMA can also be initialized by a known nonzero value $\hat{\boldsymbol{\mu}}(0):=\boldsymbol{\beta}_0$ and the resulting $\hat{\boldsymbol{\mu}}(n)$ can be interpreted as the maximum *a posteriori* (MAP) estimator of $\boldsymbol{\mu}(n)$ under the assumption that the prior distribution of $\boldsymbol{\mu}(n)$ is $\mathrm{N}(\boldsymbol{\beta}_0,\lambda^{-n}\mathbf{V})$ and that given $\boldsymbol{\mu}(n)$, the $\boldsymbol{\xi}(\tau)$ $(\tau=1,\cdots,n)$ are independently distributed with $\boldsymbol{\xi}(\tau)\sim\mathrm{N}(\boldsymbol{\mu}(n),(1-\lambda)^{-1}\lambda^{\tau-n}\mathbf{V})$, where $\mathbf{V}$ is an arbitrary positive definite matrix.

Similarly, an exponential weight function $\lambda^{n-\tau}$ can be introduced into (15) with $\tilde{\boldsymbol{\xi}}(\tau):=\boldsymbol{\xi}(\tau)-\hat{\boldsymbol{\mu}}(\tau)$. The resulting estimator, denoted by $\hat{\mathbf{A}}(n):=[\hat{\mathbf{A}}_1(n),\cdots,\hat{\mathbf{A}}_\kappa(n)]$, satisfies (14) and can be computed online by the recursive least-squares (RLS) algorithm [44]:

$$\hat{\mathbf{A}}(\tau)=\hat{\mathbf{A}}(\tau-1)+\mathbf{e}(\tau)\,\mathbf{G}(\tau), \tag{25}$$

where

$$\mathbf{e}(\tau) := \tilde{\boldsymbol{\xi}}(\tau) - \hat{\mathbf{A}}(\tau-1)\boldsymbol{\gamma}(\tau), \tag{26}$$

$$\mathbf{G}(\tau) := \frac{\boldsymbol{\gamma}(\tau)'\mathbf{P}(\tau-1)}{\lambda + \boldsymbol{\gamma}(\tau)'\mathbf{P}(\tau-1)\boldsymbol{\gamma}(\tau)}, \tag{27}$$

$$\mathbf{P}(\tau) := \lambda^{-1}\mathbf{P}(\tau-1) - \lambda^{-1}\mathbf{P}(\tau-1)\boldsymbol{\gamma}(\tau)\mathbf{G}(\tau). \tag{28}$$

The corresponding estimator of $\boldsymbol{\Sigma}$ can also be computed online by

$$\hat{\boldsymbol{\Sigma}}(\tau) = d(\lambda,\tau)\hat{\boldsymbol{\Sigma}}(\tau-1) + c(\lambda,\tau)\hat{\boldsymbol{\varepsilon}}(\tau)\hat{\boldsymbol{\varepsilon}}(\tau)', \tag{29}$$

where $\hat{\boldsymbol{\varepsilon}}(\tau) := \tilde{\boldsymbol{\xi}}(\tau) - \hat{\mathbf{A}}(\tau)\boldsymbol{\gamma}(\tau)$. To avoid inverting the matrix $\boldsymbol{\Gamma}(1) = \boldsymbol{\gamma}(1)\boldsymbol{\gamma}(1)'$, which is required to obtain the initial value $\mathbf{P}(1)$, RLS is usually started at $\tau = 1$ (instead of $\tau = 2$) with predetermined initial values $\hat{\mathbf{A}}(0) := \mathbf{A}_0$ and $\mathbf{P}(0) := \mathbf{P}_0$. The resulting $\hat{\mathbf{A}}(n)$ satisfies (14) with $\boldsymbol{\Xi} := \sum_{\tau=1}^{n} \lambda^{n-\tau}\tilde{\boldsymbol{\xi}}(\tau)\boldsymbol{\gamma}(\tau)' + \lambda^n \mathbf{A}_0\mathbf{P}_0^{-1}$ and $\boldsymbol{\Gamma} := \sum_{\tau=1}^{n} \lambda^{n-\tau}\boldsymbol{\gamma}(\tau)\boldsymbol{\gamma}(\tau)' + \lambda^n \mathbf{P}_0^{-1}$. This estimator coincides with the MAP estimator of $\mathbf{A}(n) := [\mathbf{A}_1(n), \cdots, \mathbf{A}_\kappa(n)]$ under the assumption that $\tilde{\boldsymbol{\xi}}(\tau) = \mathbf{A}(n)\boldsymbol{\gamma}(\tau) + \boldsymbol{\varepsilon}(\tau)$ $(\tau = 1, \cdots, n)$, where the prior distribution of $\mathrm{vec}\{\mathbf{A}(n)\}$ is $\mathrm{N}(\mathrm{vec}(\mathbf{A}_0), \lambda^{-n}\mathbf{P}_0 \otimes \boldsymbol{\Sigma})$ and, conditioning on $\mathbf{A}(n)$, the $\boldsymbol{\varepsilon}(\tau)$ are independently distributed with $\boldsymbol{\varepsilon}(\tau) \sim \mathrm{N}(\mathbf{0}, \lambda^{\tau-n}\boldsymbol{\Sigma})$. Note that the smaller is $\mathbf{P}_0$ the more important is the role of the prior mean $\mathbf{A}_0$. A convenient choice for $\mathbf{P}_0$ is $\mathbf{P}_0 = (1/\sigma_0^2)\mathbf{I}$, where $\sigma_0^2 > 0$ is a tuning parameter which, ideally, would approximate $\mathrm{Var}\{\xi_j(\tau)\}$ $(j \in \mathscr{C})$.

As an example, Fig. 3 shows the one-day-ahead predictions made at the end of each day for the time series in Fig. 1 where the weekday and weekend are modeled separately. The predictions are derived from the DFT filters with $\mathscr{C}$ containing, as suggested by Fig. 2, the zero frequency, the fundamental frequency and its first two harmonics. For each regime, the corresponding 7 coefficient series $\xi_j(\tau)$ are modeled as independent AR processes of the form (11) with $d = 1$ and $\kappa = 1$. The remaining parameters of each series are estimated separately by the one-dimensional version of EWMA in (24) and RLS in (25)–(28), where the former is initialized by the DFT of the average daily pattern and the latter by $\mathbf{A}_0 = 0$ and $\mathbf{P}_0 = (1/0.01)\mathbf{I}$. The forgetting factor λ is equal to 0.2 for the zero frequency and 0.99 for the nonzero frequencies. The resulting predictions account for 89.6% of the total variability of the original time series, with the root mean-square error (RMSE) equal to 0.1035.

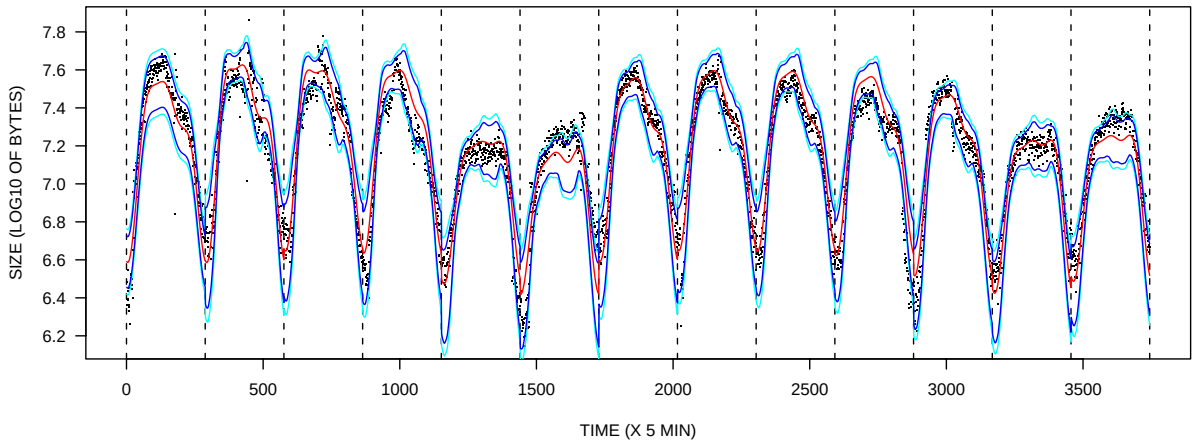


Fig. 3. *One-day-ahead predictions (red) made at the end of each day with 80% (blue) and 90% (cyan) daily simultaneous confidence bands. Dots represent actual observations. RMSE of the predictions is equal to 0.1035. Actual coverage of the confidence bands is equal to 76% and 86%, respectively. Median width of the confidence bands is equal to 0.2338 and 0.2970, respectively.*

D. Long-Term Prediction Error and Confidence Band

In addition to the point predictions, uncertainty measures such as the variance of prediction error and the confidence band are also needed to obtain a more complete picture of future requests which is critical to web server management. The proposed hierarchical approach provides a simple framework for deriving the uncertainty measures for long-term prediction.

First, consider the variance of prediction error. Let $t := (n+d-1)p+r$ $(r = 1, \cdots, p)$ and let the error of d-period-ahead prediction of $x(t)$ be defined by

$$e_L(t) := x(t) - \hat{x}(t \,|\, np). \tag{30}$$

Then, it follows from (2) and (10) that $e_L(t)$ can be decomposed as

$$e_L(t) = z(t) + \eta(t), \tag{31}$$

where $z(t)$ represents the error of long-term modeling and

$$\eta(t) := y(t) - \hat{y}(t \,|\, np) \tag{32}$$

is the error of d-step-ahead prediction of the long-term component $y(t)$, with

$$\hat{y}(t \,|\, np) := \sum_{j \in \mathscr{C}} \{\hat{\xi}_j(n+d \,|\, n)\}\, \phi_j(t)$$

denoting the prediction of $y(t)$ based on the model (11). Under the assumptions in Section III-B, $z(t)$ and $\eta(t)$ are statistically uncorrelated. Therefore,

$$\mathrm{Var}\{e_L(t)\} = \mathrm{Var}\{z(t)\} + \mathrm{Var}\{\eta(t)\}. \tag{33}$$

Given the historical data up to time np, $\mathrm{Var}\{z(t)\}$ can be predicted by the EWMA estimator

$$\hat{\sigma}_z^2(t) := c(\lambda, n) \sum_{\tau=1}^{n} \lambda^{n-\tau}\, \tilde{\sigma}_z^2((\tau-1)p+r), \tag{34}$$

where $\tilde{\sigma}_z^2((\tau-1)p+r)$ is an estimator of $\mathrm{Var}\{z((\tau-1)p+r)\}$ and may be obtained nonparametrically by smoothing the observed long-term modeling error squares $\{z^2((\tau-1)p+r)\}_{r=1}^{p}$. Note that $\hat{\sigma}_z^2(t)$ can be computed online by a recursive algorithm similar to (24). Moreover, under the model assumption (11), $\eta(t) = \boldsymbol{\varphi}(t)'\boldsymbol{\varepsilon}(n+d)$, so that

$$\mathrm{Var}\{\eta(t)\} = \boldsymbol{\varphi}(t)'\boldsymbol{\Sigma}\boldsymbol{\varphi}(t),$$

where $\boldsymbol{\varphi}(t) := \mathrm{vec}\{\phi_j(t),\, j \in \mathscr{C}\}$. Therefore, $\mathrm{Var}\{\eta(t)\}$ can be predicted by

$$\hat{\sigma}_\eta^2(t) := n(n - \kappa m^2 - 1)^{-1} \boldsymbol{\varphi}(t)' \hat{\boldsymbol{\Sigma}}(n) \boldsymbol{\varphi}(t), \tag{35}$$

with $\hat{\boldsymbol{\Sigma}}(n)$ given by (29). Combining the results in (33)–(35) suggests that

$$\hat{\sigma}_L^2(t) := \hat{\sigma}_z^2(t) + \hat{\sigma}_\eta^2(t) \tag{36}$$

can be employed as a predictor of $\mathrm{Var}\{e_L(t)\}$.

As an example, Fig. 4 shows $\hat{\sigma}_L(t)$, the predicted standard deviation (SDV) of long-term prediction error, for the one-day-ahead predictions given in Fig. 3, where $\hat{\sigma}_L^2(t)$ is calculated according to (34)–(36). Also shown in Fig. 4 is the actual SDV, denoted by $\sigma_L(t)$, of long-term prediction error obtained by applying to the time series $e_L^2(t)$ a nonparametric procedure of adaptive smoothing, known as *super smoother* [45], with smoothing bandwidth $0.1p$ (2.4 hours). The super smoother is also used to obtain $\tilde{\sigma}_z^2((\tau-1)p+r)$ in (34), where the smoothing bandwidth is selected adaptively by cross-validation. The forgetting factor λ in (34) is equal to 0.9, which is chosen to minimize the average value of the symmetrized Kulback-Leibler divergence $f(\hat{\sigma}_L(t), \sigma_L(t))$, with $f(x,y) := \frac{1}{2}(x/y + y/x) + 1$ for $x, y > 0$, which serves as a goodness-of-fit measure of $\hat{\sigma}_L(t)$. Moreover, $\hat{\boldsymbol{\Sigma}}(n)$ in (35) is a diagonal matrix in which the diagonal elements are calculated independently according to the one-dimensional version of (29) with $\lambda = 0.9$.

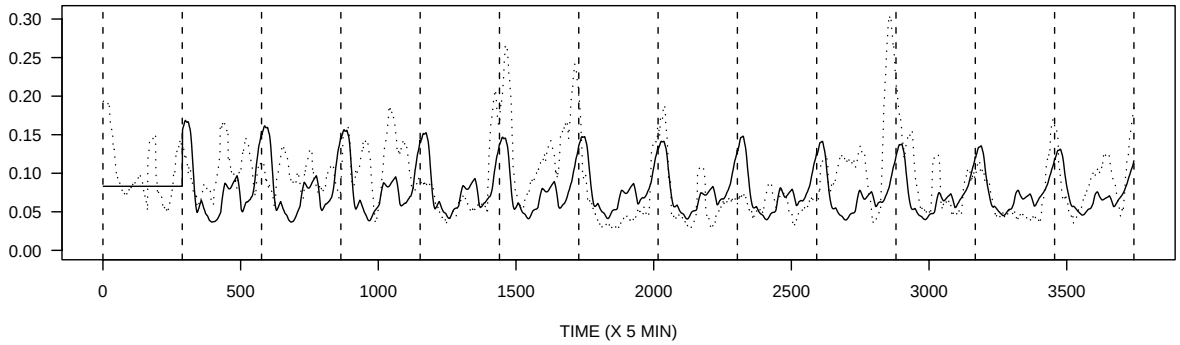


Fig. 4. *Solid line represents the predicted SDV of the long-term prediction error $e_L(t)$ obtained according to (36). Dotted line represents the actual SDV of the long-term prediction error obtained by smoothing $e_L^2(t)$.*

To construct simultaneous confidence bands (or joint prediction intervals, to be more precise) for the long-term prediction, it suffices to consider the following "signal-plus-noise" model with time-varying noise variance (heteroscedasticity):

$$x(t) = \hat{x}(t\,|\,np) + \hat{\sigma}_L(t)\,\tilde{e}_L(t),$$

where $\tilde{e}_L(t) := e_L(t)/\hat{\sigma}_L(t)$ is the *standardized* error. Under this model, a symmetric simultaneous confidence band of $x(t)$ for $t = (n+d-1)p+r$ and $r = 1,\cdots,p$ is given by

$$\hat{x}(t\,|\,np) \pm \theta(\alpha)\,\hat{\sigma}_L(t), \tag{37}$$

where $1-\alpha$ is the prescribed level of confidence and $\theta(\alpha)$ satisfies

$$\Pr\big\{|\tilde{e}_L((n+d-1)p+r)| < \theta(\alpha),\ \forall r = 1,\cdots,p\big\} = 1-\alpha. \tag{38}$$

Asymmetrical confidence bands can be constructed similarly.

The critical value $\theta(\alpha)$ depends on the *joint* distribution of the random vector

$$\tilde{\mathbf{e}}_L(n+d) := \mathrm{vec}\{\tilde{e}_L((n+d-1)p+r)\}_{r=1}^{p}.$$

This distribution can be estimated by the empirical distribution of the historical data $\{\tilde{\mathbf{e}}_L(\tau)\}_{\tau=1}^{n}$. A disadvantage of this method is that it requires a tremendous amount of training data, especially when p is large. To alleviate the predicament, one may assume that the components in $\tilde{\mathbf{e}}_L(\tau)$ are independent and identically distributed (i.i.d.), so that only the one-dimensional marginal distribution

of $\tilde{e}_L(t)$ needs to be estimated. Unfortunately, the i.i.d. confidence band may result in erroneous coverage probabilities in situations where the components in $\tilde{\mathbf{e}}_L(\tau)$ are actually correlated.

As a trade-off between the feasibility of computational and data resources and the need for accommodating correlated data, consider the following method which may be termed as *analysis-by-synthesis* or ABS. In the ABS method, the intra-period serial correlation of $\tilde{e}_L(t)$ is handled by a one-dimensional AR model fitted to the historical data $\{\tilde{e}_L(t)\}_{t=1}^{np}$. Pseudo random samples of $\tilde{\mathbf{e}}_L(n+d)$ are generated from the model by using simulated excitations which are now i.i.d. random variables drawn from a distribution, selected in a library of parametric families of distributions, that best fits the one-dimensional marginal distribution of the historical residuals of the model. Of course, when historical data are sufficiently abundant, the i.i.d. excitations can also be simulated nonparametrically by bootstrapping (e.g., sampling historical residuals with or without replacement). Finally, the critical value $\theta(\alpha)$ is determined according to (38) from the empirical distribution of the pseudo random samples of $\tilde{\mathbf{e}}_L(n+d)$. Note that by the theory of binomial distributions, the standard error in estimating the probability α would be less than a given precision δ if one uses more than $\alpha(1-\alpha)/\delta^2$ random samples.

Fig. 5 shows the time series $\tilde{e}_L(t)$ and its partial autocorrelation function corresponding to the long-term predictions in Fig. 3 and the SDV estimates in Fig. 4. As can be seen, $\tilde{e}_L(t)$ is not a white noise process, but the serial correlation of $\tilde{e}_L(t)$ is predominantly short-term and of the AR type; the regime shift, which is so apparent in $x(t)$, does not seem to play a significant role in $\tilde{e}_L(t)$. This justifies the application of the ABS method with a single AR model for all regimes.

The confidence bands shown in Fig. 3 are constructed by the ABS method according to (37) on the basis of 10,000 simulation runs of a zero-mean AR(9) process fitted to the time series $\tilde{e}_L(t)$. The i.i.d. excitations of the AR(9) model are simulated from the T distribution with 5 degrees of freedom, denoted by T_5. The order of the AR(9) model is determined by AIC. The T_5 distribution is selected from the families of T and Gaussian distributions by first estimating the parameters of these distributions to obtain the best fit in each family and then choosing the distribution that has the minimum absolute quantile deviation from the data. As shown by the QQ plot in Fig. 6(b), the T_5 distribution fits the residuals reasonably well, although a distribution with heavier tails may provide a better fit. Note that the T distributions are always rescaled to match the variance of the data and the degrees of freedom are estimated from the data by equating the kurtosis.

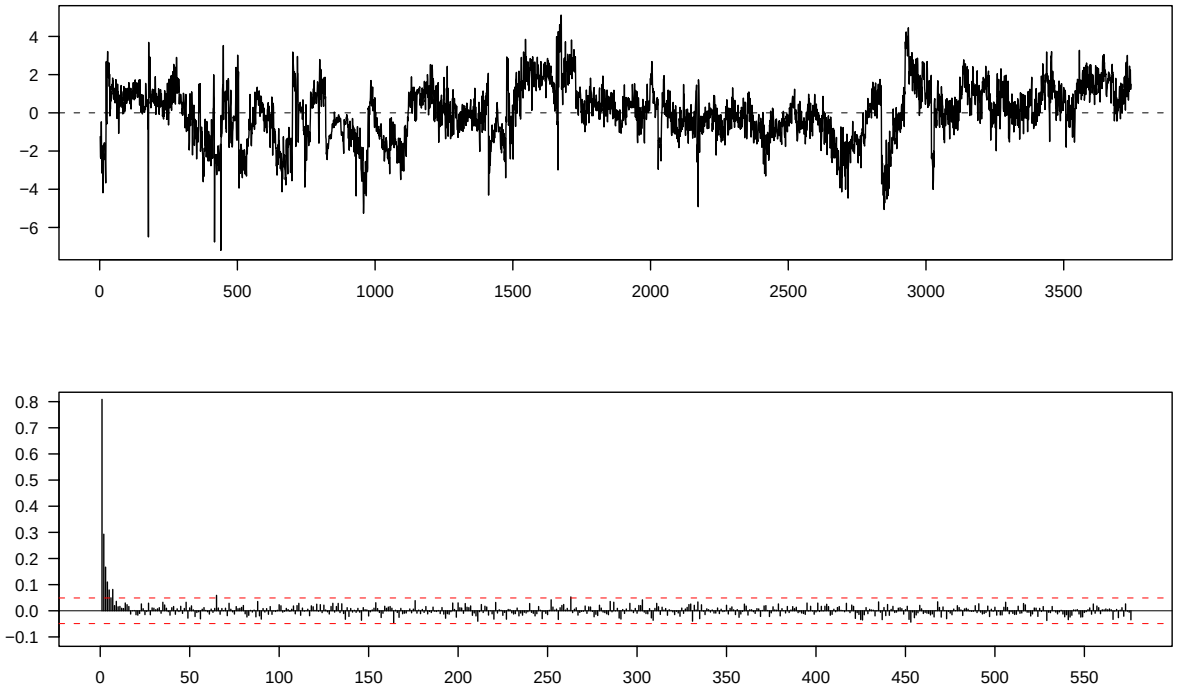


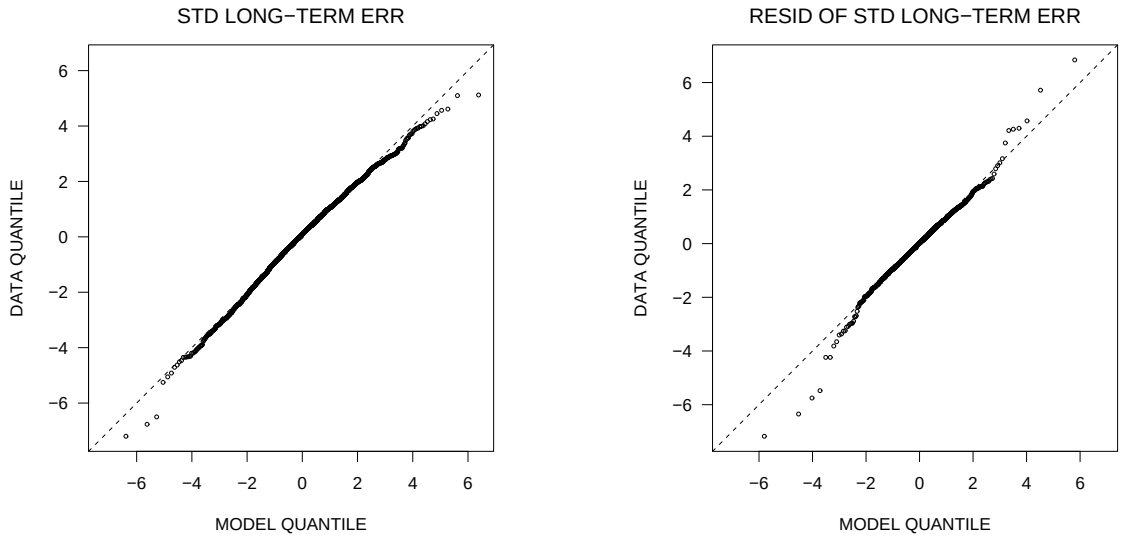
Fig. 5. *Top, standardized error $\tilde{e}_L(t)$ of one-day-ahead prediction. Bottom, partial autocorrelation function of $\tilde{e}_L(t)$ with 90% pointwise confidence interval (dashed line).*

The actual coverage of the ABS confidence bands in Fig. 3 is 76% and 86%, respectively. These numbers are much closer to the nominal values than the actual coverage of 64% and 76% for the confidence bands (not shown) constructed under the i.i.d. assumption of $\tilde{e}_L(t)$ using the T_{14} distribution, even though T_{14} fits the marginal distribution of $\tilde{e}_L(t)$ very well, as can be seen from Fig. 6(a). The i.i.d. confidence bands are too narrow. The utilization of serial correlation is largely responsible for the better performance of the ABS confidence bands.

Note that the confidence bands do not necessarily have a constant width because $\hat{\sigma}_L(t)$ varies with t, even if $\tilde{e}_L(t)$ is a stationary process. Note also that the event of actual $x(t)$ exceeding a confidence band can occur in clusters over time because of serial correlation.

IV. Short-Time Modeling and Prediction

Since the short-term component $z(t)$ in (2) cannot be obtained from partial observations of $x(t)$ within a period, the long-term prediction error $e_L(t)$ defined by (30) is employed instead as the raw



(a) (b)

Fig. 6. *(a) QQ plot of the standardized error $\tilde{e}_L(t)$ of one-day-ahead prediction versus the T_{14} distribution. (b) QQ plot of the residuals from an AR(9) fit of $\tilde{e}_L(t)$ versus the T_5 distribution.*

material for short-term modeling and prediction. Note that $e_L(t)$ is an unbiased estimator of $z(t)$ if $\boldsymbol{\xi}(\tau)$ is assumed to satisfy (11) and the error of parameter estimation is ignored.

As remarked in the introduction, one can handle the short-term components separately for each regime or combine them to form a single time series under the assumption that the statistical properties do not change abruptly with regime shift. A visual examination of the long-term prediction error shown in Fig. 7 seems to justify the latter approach for this example.

Fig. 7 also shows that the serial correlation of $e_L(t)$ is predominantly short-term and of the AR type. The absence of the strong periodicity, which is contained in the original time series, indicates that the long-term prediction is quite adequate in handling the periodic patterns. As a consequence, the procedure of short-term modeling and prediction becomes straightforward.

More specifically, it suffices to model and predict $e_L(t)$ in the same way that $\boldsymbol{\xi}(\tau)$ is modeled and predicted except that (11) becomes a horizon-specific multistep AR(h,q) model

$$e_L(t) = \mu + \sum_{i=1}^{q} a_i \{e_L(t-h-i+1) - \mu\} + \varepsilon(t), \quad \{\varepsilon(t)\} \sim \text{IID}(0, \sigma_\varepsilon^2), \tag{39}$$

where $h \geq 1$ is the given horizon of short-term prediction. Under this model, the best h-step-ahead

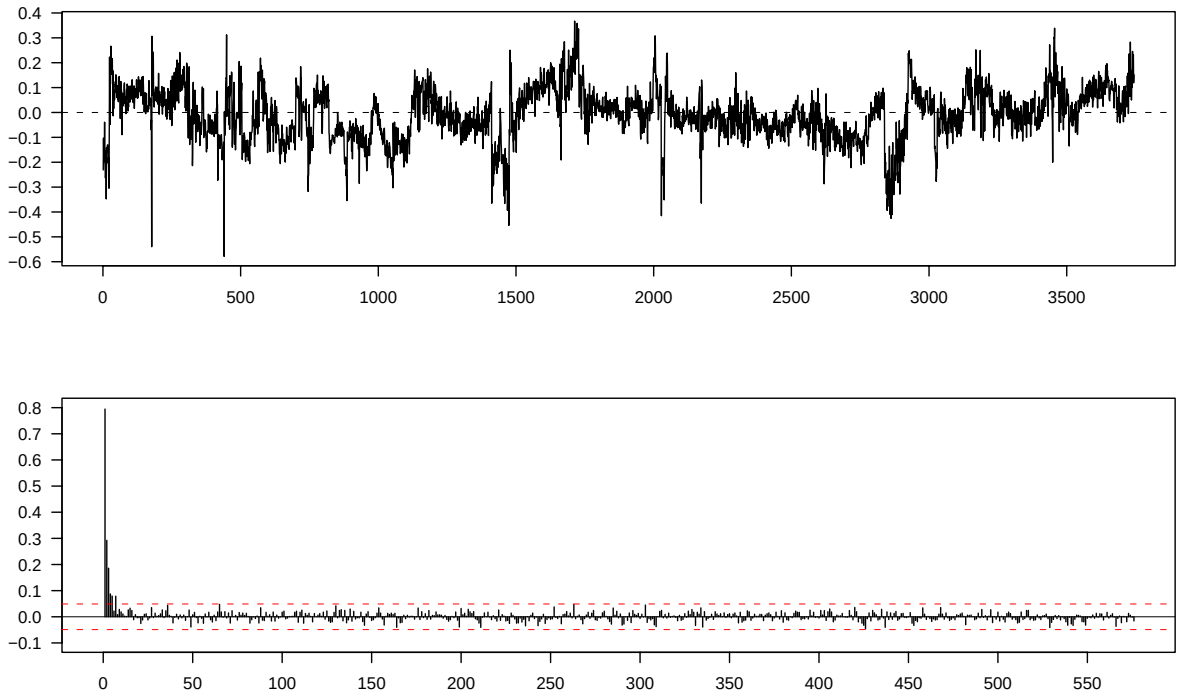


Fig. 7. *Top, one-day-ahead prediction error $e_L(t)$. Bottom, partial autocorrelation function of $e_L(t)$ with 90% pointwise confidence interval (dashed line).*

(short-term) prediction of $e_L(t+h)$ on the basis of $\{e_L(1),\cdots,e_L(t)\}$ is given by

$$\hat{e}_L(t+h\,|\,t) := \mu + \sum_{i=1}^{q} a_i\{e_L(t-i+1)-\mu\}, \tag{40}$$

with σ_ε^2 being the variance of the prediction error. The estimation procedures discussed in Section III, including the order selection and the adaptive estimation, can be applied directly to (39).

Given the short-term predictions of $e_L(t)$, the short-term predictions of $x(t)$ can be easily obtained by combining (40) with (10). More specifically, for $t := (n+d-1)p+r$ $(r=1,\cdots,p)$, the short-term prediction of $x(t+h)$ at time t is given by

$$\hat{x}(t+h\,|\,t) := \hat{x}(t\,|\,np) + \hat{e}_L(t+h\,|\,t). \tag{41}$$

As an example, Fig. 8 shows the five-minute-ahead predictions constructed according to (41) from the one-day-ahead predictions in Fig. 3 and the one-step-ahead predictions of $e_L(t)$ given by (40) with $h=1$ and $\mu=0$. In (40), the order $q=7$ is selected by AIC (16) and the coefficients a_i are calculated by RLS (25)–(28) with $\lambda = 0.9999$, $\mathbf{A}_0 = \mathbf{0}$, and $\mathbf{P}_0 = \mathbf{I}$. The five-minute-ahead

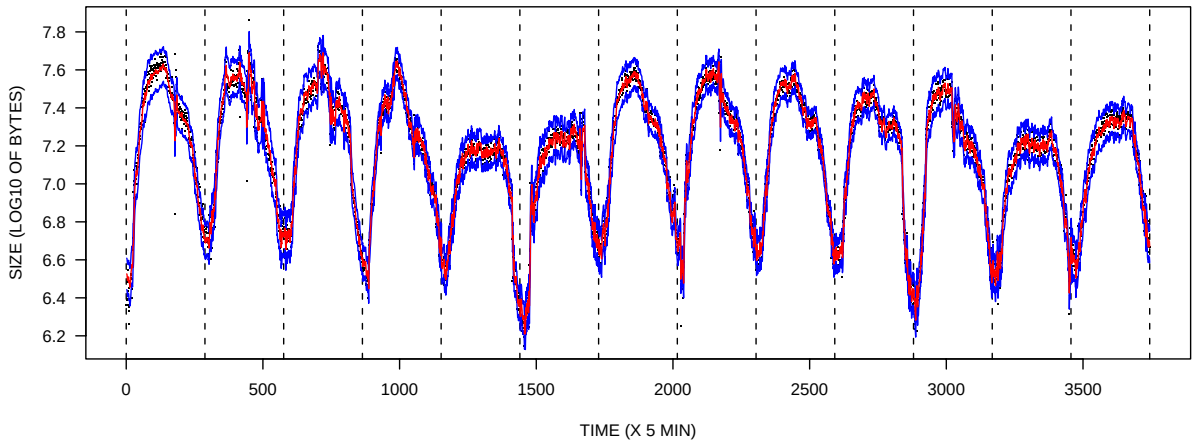


Fig. 8. *Five-minute-ahead predictions (red), obtained by revising the one-day-ahead predictions in Fig. 3 according to (41), with 90% confidence band (blue). Dots represent actual observations. RMSE of the predictions is equal to 0.0590. Actual coverage of the confidence band is equal to 89%. Median width of the confidence band is equal to 0.1618.*

predictions in Fig. 8 account for 96.6% of the total variability of the original time series in Fig. 1, with RMSE equal to 0.0590.

The effectiveness of short-term modeling and prediction can be assessed by examining the time series of short-term prediction error

$$e_S(t) := x(t) - \hat{x}(t\,|\,t-h) = e_L(t) - \hat{e}_L(t\,|\,t-h),$$

which is shown in Fig. 9, along with its partial autocorrelation function, for the five-minute-ahead predictions in Fig. 8. As can be seen, the short-term prediction is quite effective in utilizing the serial correlation remained in the long-term prediction error shown in Fig. 7.

Confidence bands for the short-term predictions of $x(t)$ can be constructed by an assessment of the statistical properties of $e_S(t)$. As in the case of long-term prediction, it is assumed that $x(t)$ satisfies the following "signal-plus-noise" model with time-varying noise variance:

$$x(t) = \hat{x}(t\,|\,t-h) + \hat{\sigma}_S(t)\,\tilde{e}_S(t),$$

where $\tilde{e}_S(t) := e_S(t)/\hat{\sigma}_S(t)$ is the standardized error and $\hat{\sigma}_S^2(t)$ is a predictor of $\mathrm{Var}\{e_S(t)\}$.

Since $e_S(t) = \varepsilon(t)$ under the assumption (39) and in the absence of estimation error, an estimator of σ_ε^2 in (39) can be used as $\hat{\sigma}_S^2(t)$ to predict $\mathrm{Var}\{e_S(t)\}$. In constructing the confidence band

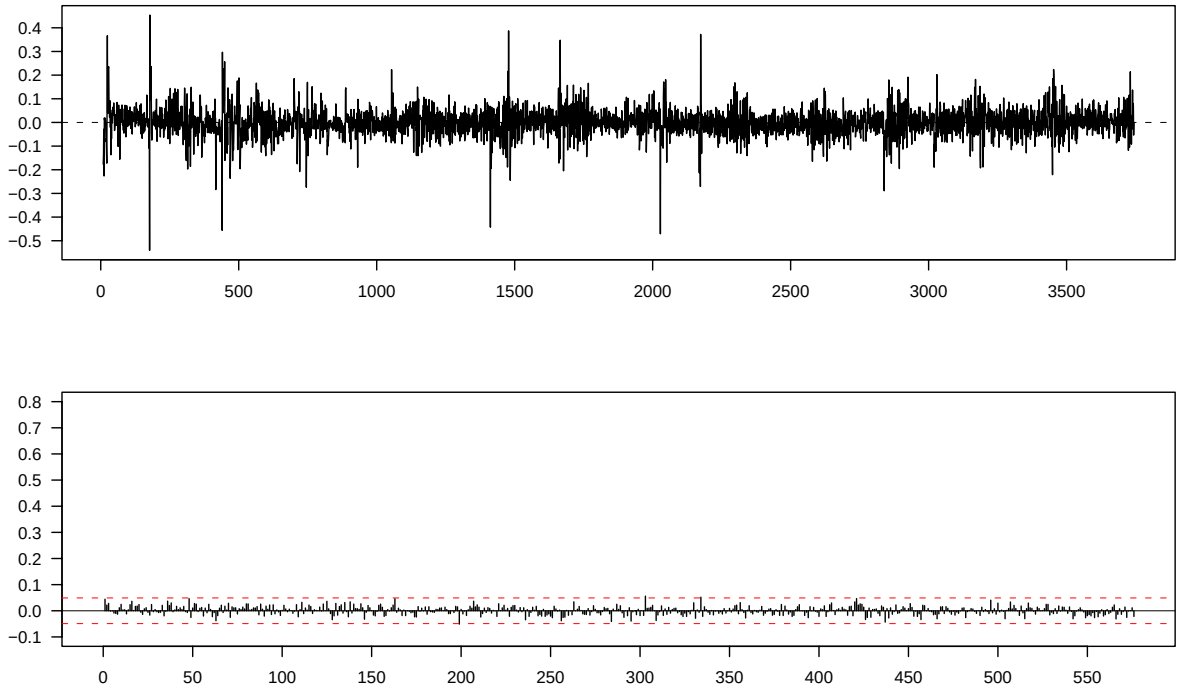


Fig. 9. *Top, five-minute-ahead prediction error $e_S(t)$. Bottom, partial autocorrelation function of $e_S(t)$ with 90% pointwise confidence interval (dashed line).*

shown in Fig. 8, $\hat{\sigma}_S^2(t)$ is obtained by RLS in a way similar to (29), with $\lambda = 0.99$. The forgetting factor is determined by minimizing the average value of $f(\hat{\sigma}_S(t), \sigma_S(t))$, where $\sigma_S^2(t)$ is the actual variance of $e_S(t)$ obtained by applying the super smoother to the time series $e_S^2(t)$ with smoothing bandwidth $0.4p$ (9.6 hours). Examination of the resulting $\tilde{e}_S(t)$ and its partial autocorrelation function (not shown but similar to Fig. 9) reveals no significant serial correlation in $\tilde{e}_S(t)$. Fig. 10 shows that T_5 fits the marginal distribution of $\tilde{e}_S(t)$ reasonably well except at the extreme of left tail. As in the case of long-term prediction, the T distribution is selected optimally from the library of T and Gaussian distributions.

The 90% confidence band in Fig. 8 is constructed using the T_5 distribution under the i.i.d. assumption of $\tilde{e}_S(t)$. The actual coverage is 89%, which is quite accurate, indicating that the i.i.d. assumption is plausible in this case. Note that the median width of the confidence band in Fig. 8 is approximately a half of that of the corresponding confidence band in Fig. 3.

As an example of multistep-ahead forecasting, Fig. 11 shows the twenty-minute-ahead predic-

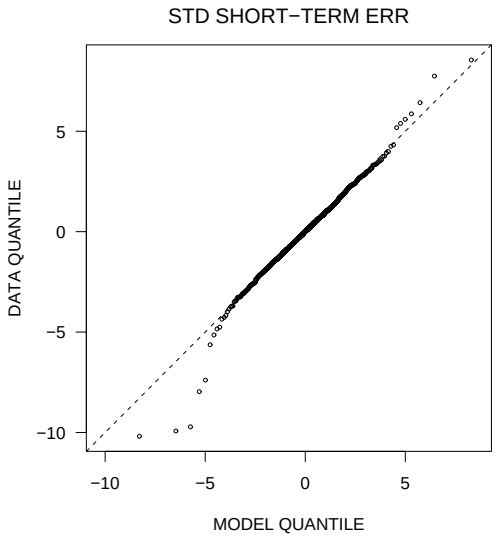


Fig. 10. *QQ plot of the standardized error $\tilde{e}_S(t)$ of five-minute-ahead prediction versus the T_5 distribution.*

tions made according to (41) from the one-day-ahead predictions in Fig. 3 and the four-step-ahead predictions of $e_L(t)$ based on (40) with $h = 4$. While the coefficients a_i are estimated adaptively by RLS, all other parameters, including the order q, the tuning parameters in RLS, and the smoothing parameters for the confidence band, remain the same as they are for $h = 1$. Moreover, the T_5 distribution is still the optimal choice for the marginal distribution of the standardized short-term error and provides an acceptable fit, as can be seen from Fig. 12(a). The corresponding 90% i.i.d. confidence band (not shown) remains reasonably accurate, yielding an actual coverage of 88%.

A more careful way of constructing the confidence band should take into account the fact that the errors of multistep-ahead prediction are no longer uncorrelated. In fact, it can be proven that the errors of h-step-ahead prediction constitute an MA($h-1$) process. Owing to this observation, the ABS method can be employed again, now with an MA($h-1$) model fitted to $\tilde{e}_S(t)$. Estimates of the MA parameters can be obtained from the sample autocovariance function of $\tilde{e}_S(t)$ by using the innovations algorithm [36, p. 172]. Residuals of the MA model can be obtained by inverse (AR) filtering. Fig. 12(b) shows that the T_4 distribution, which is the optimal choice among T and Gaussian distributions, fits the residuals reasonably well except for some extreme values. The resulting 90% confidence band, based on 10,000 simulation runs, is shown in Fig. 11, which yields an actual coverage of 91% — one percentage point better than the i.i.d. confidence band.

Finally, Fig. 13 shows the RMSE of short-term prediction for different values of h. In obtaining

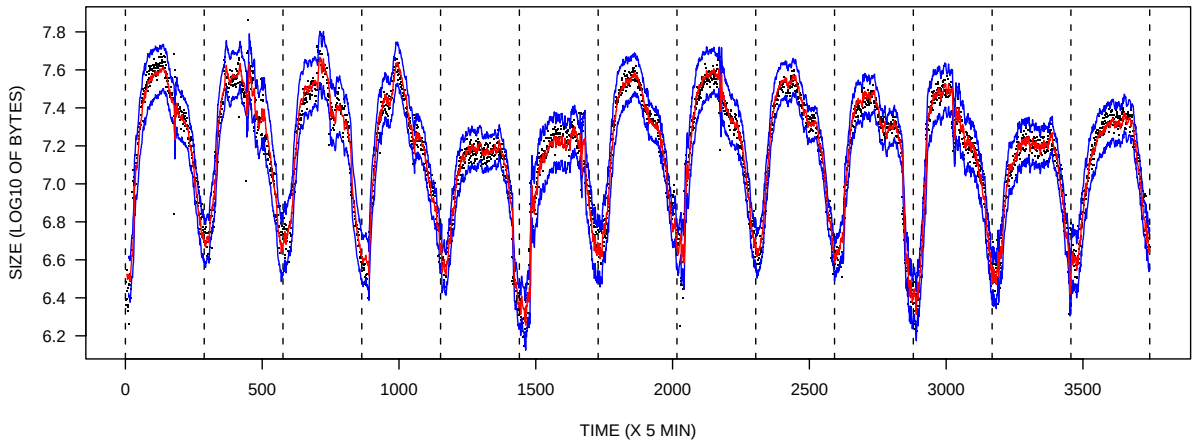


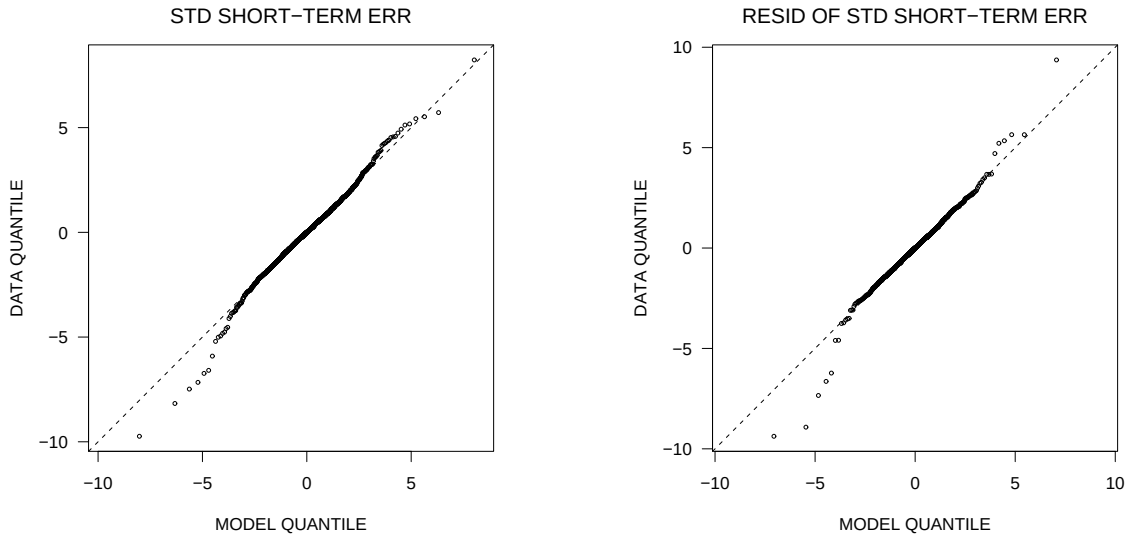
Fig. 11. *Twenty-minute-ahead predictions (red) with 90% ABS confidence band (blue). Dots represent actual observations. RMSE of the predictions is equal to 0.0726. Actual coverage of the confidence band is equal to 91%. Median width of the confidence band is equal to 0.2259.*

the predictions for $h > 1$, all parameters, except the a_i which are estimated by RLS, remain the same as they are for $h = 1$. As expected, the RMSE increases with the increase of h until it reaches, and possibly exceeds, the RMSE of long-term prediction. In this example, the three-hour-ahead ($h = 36$) predictions continue to achieve a smaller RMSE than the one-day-ahead predictions.

V. Concluding Remarks

A hierarchical framework has been proposed for modeling and forecasting the time series of web service requests which consists of strong daily patterns and more rapid fluctuations. In this framework, the daily patterns are modeled as parsimonious linear combinations of basis functions with random coefficients and the short-term fluctuations are represented as an autoregressive process. Both long-term (in days) and short-term (in minutes) predictions have been derived along with simultaneous confidence bands constructed by the analysis-by-synthesis method. Adaptive algorithms, EWMA and RLS, have been employed to provide online estimation of the model parameters that may vary slowly with time due to the nonstationarity of web service requests.

The proposed method is not limited to the time series of request arrival rates such as the five-minute total file sizes employed to demonstrate the method. Indeed, other statistical parameters, such as the standard deviation of file sizes shown in Fig. 14, that describe different aspects of



(a) (b)

Fig. 12. *(a) QQ plot of the standardized error $\tilde{e}_S(t)$ of twenty-minute-ahead prediction versus the T_5 distribution. (b) QQ plot of the residuals from an MA(3) fit of $\tilde{e}_S(t)$ versus the T_4 distribution.*

the stochastic behavior of service requests within the time intervals of length Δ may also contain strong daily patterns in addition to more rapid fluctuations. Online prediction of these parameters using the proposed method provides additional features useful in characterizing the future server workload and performance for optimal dynamic resource allocation.

Although RLS has been found adequate in the numerical examples of this paper, it is possible that the model parameters may change too rapidly for RLS to handle. In this case, the Kalman filtering techniques can be employed to estimate the model parameters as the hidden state vector of simple state-space models (e.g., random walk) [28] [46]. The state-space models may contain some unknown constants that may be treated as tuning parameters, similar to the forgetting factor in RLS, or estimated jointly by, for example, the Gaussian maximum likelihood method. Note that RLS is a special case of Kalman filtering where the state vector is constant and the observation noise is a zero-mean Gaussian white noise process with a time-varying variance which grows as power function of the forgetting factor λ. If, for example, the state vector is assumed to be a Gaussian random walk in which the (zero-mean) perturbation has a time-varying variance which is equal to ρ times the variance of the observation noise, then it can be shown that the corresponding Kalman filter coincides with (25)–(28) except that (28) includes an extra term $\rho\mathbf{I}$. In this case, ρ

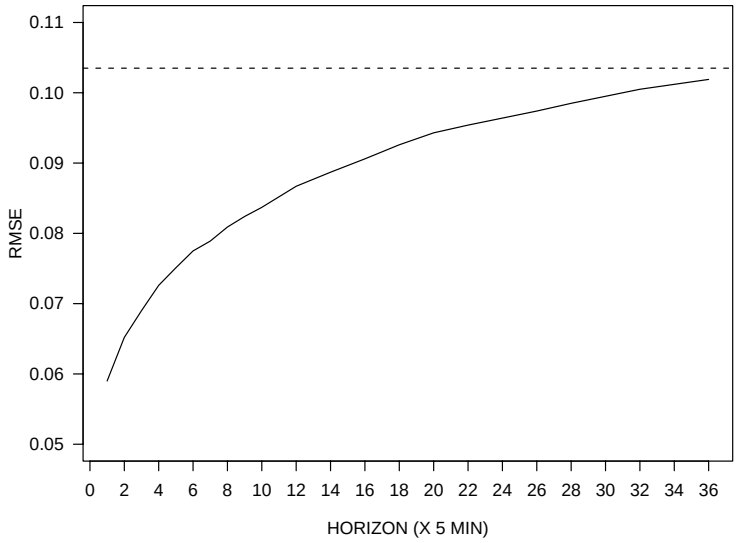


Fig. 13. *RMSE of short-term prediction as a function of horizon h. Dashed line represents RMSE of one-day-ahead prediction in Fig. 3.*

can be viewed as an additional tuning parameter interpretable as the system-noise-to-observation-noise ratio of the state-space model. For the data sets considered in this paper, the random-walk assumption did not lead to improved performance for either long-term or short-term prediction, which suggests that the fluctuation (if any) of the parameters in (12) and (40) is slower than a random walk.

In constructing the confidence bands, it is important that the time-varying variance of the prediction errors be estimated accurately, especially in long-term prediction. The estimation has been done in this paper by using nonparametric smoothing and EWMA. Parametric methods based on the autoregressive conditional heteroscedastic (ARCH) and generalized ARCH (GARCH) models [28][29] may also be used to handle the time-varying variance. Moreover, it has been found that the marginal distribution of the prediction errors tends to have heavy tails. Indeed, the T distribution has always been chosen over the Gaussian distribution by the automatic selection criterion of minimum absolute quantile deviation for both long-term and short-term prediction errors. Parametric distributions with heavier (symmetric or asymmetric) tails may be more helpful.

It is not surprising that both long-term and short-term components in a time series of service requests may contain outliers – the observations that dramatically exceed the normal range of

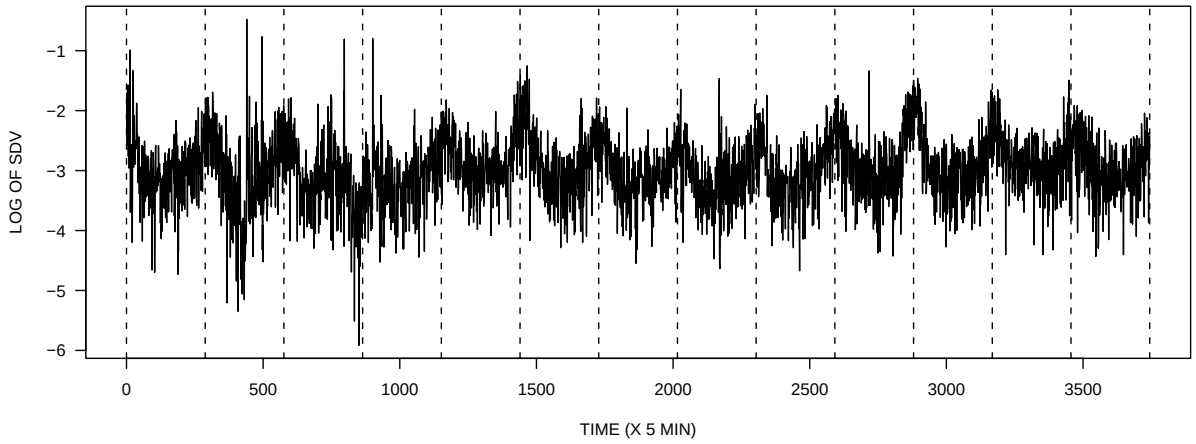


Fig. 14. *Standard deviation (in logarithm) of the log file sizes in five-minute intervals corresponding to the example of HTTP requests shown in Fig. 1.*

fluctuation for a short period of time relative to the prediction horizons. While outliers in the short-term component exist for many reasons, outliers in the daily patterns may also occur due to special events, such as sales promotions and holidays, for which the increase of requests is dramatic but temporary, as can be seen from the example shown in Fig. 15. Such outliers cannot be predicted accurately for the given prediction horizon, unless similar events occurred repeatedly in the past, or the prediction horizon is shortened, as shown in Fig. 15, where the short-term predictions dramatically improve the one-day-ahead prediction for day 8. The inaccurate predictions caused by outliers will produce anomalies in the time series of prediction error which must be dealt with judiciously in order to prevent them from contaminating the modeling and prediction in the future when the outliers disappear. Robust estimation techniques can be used to handle the outliers. For example, in EWMA and RLS, one may update the parameters only if the increments are within the "normal" range. The results in Fig. 15 are obtained in this way.

Finally, the prediction methods discussed in this paper are essentially linear. Nonlinear methods should be explored in future research, not only for short-term prediction [47]–[49], but also for long-term prediction, with proper simultaneous confidence bands, and with efficient mechanisms to handle nonstationarity.

Acknowledgment. The author thanks Tamar Eilam, Alan King, and Mark Squillante for helpful discussions. He also thanks Tracy Kimbrel for preparing the HTTP request data.

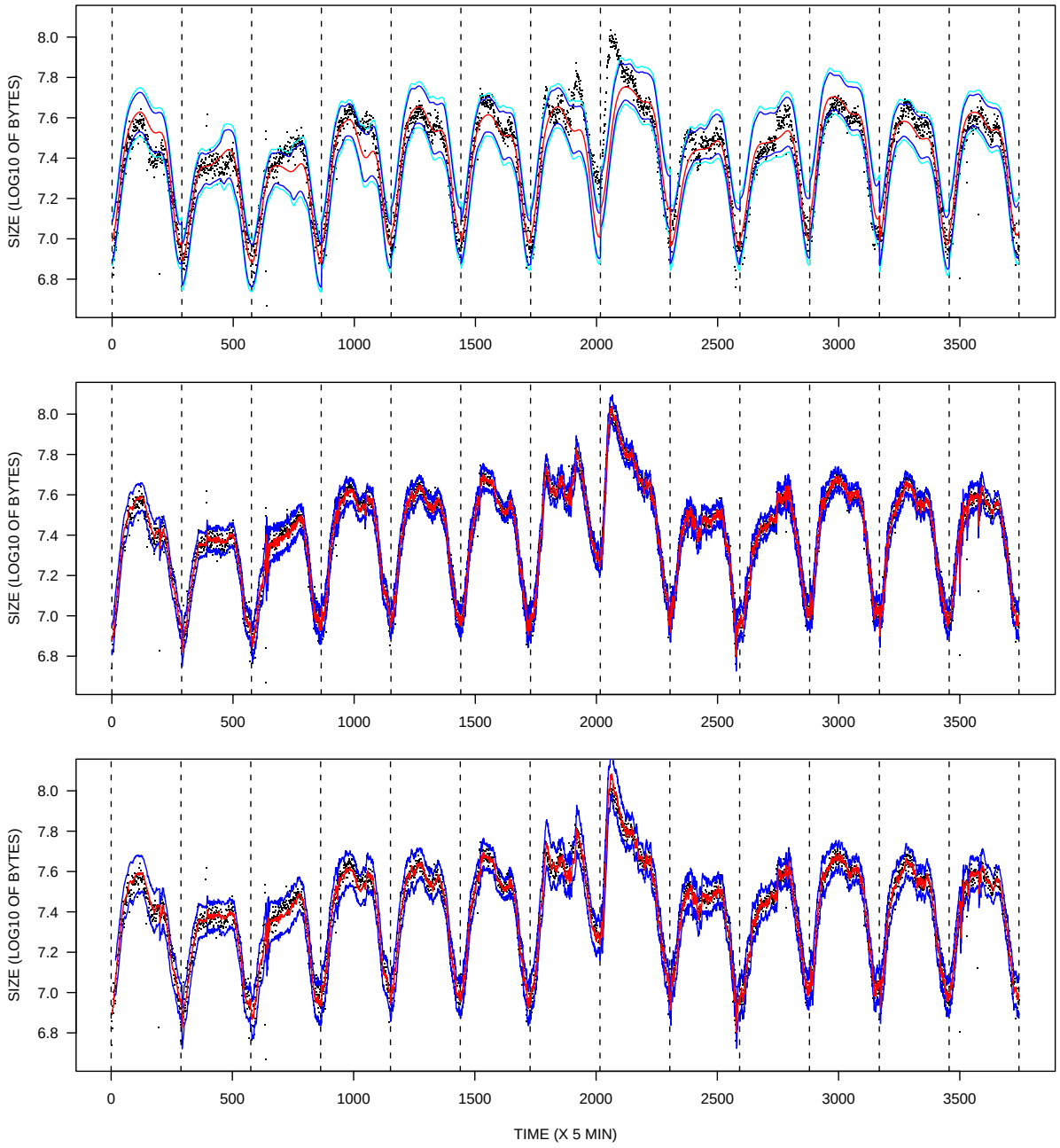


Fig. 15. *Another example at a different commercial web site. Top, total file sizes of HTTP requests (black dots) and one-day-ahead predictions (red) with 80% (blue) and 90% (cyan) confidence bands, where RMSE = 0.0967 and actual coverage = 77% and 87%, respectively, with median width 0.2010 and 0.2488. Middle, five-minute-ahead predictions (red) with 90% confidence bands (blue), where RMSE = 0.0464 and actual coverage = 86% with median width 0.1132. Bottom, twenty-minute-ahead predictions (red) with 90% confidence band (blue), where RMSE = 0.0513 and actual coverage = 92% with median width 0.1560.*

REFERENCES

[1] V. A. F. Almeida and D. A. Menascé, *Capacity Planning for Web Services: Metrics, Models, and Methods*, Upper Saddle River, NJ: Prentice Hall, 2002.

[2] A. Chandra, W. Gong, and P. Shenoy, "An online optimization-based technique for dynamic resource allocation in GPS servers," to appear in *Proceedings of the ACM Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, June 2003. Electronic version available at `http://lass.cs.umass.edu/~lass/papers/pdf/TR02-30.pdf`.

[3] R. P. Doyle, J. S. Chase, O. M. Asad, W. Jen, and A. M. Vahdat, "Model-based resource provisioning in a web service utility," to appear in *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, March 2003. Electronic version available at `http://issg.cs.duke.edu/publications/mbrp-usits03.pdf`.

[4] D. P. De Farias, A. King. M. Squillante, and B. Van Roy, "Dynamic control of web server farms," presented at *the 2nd Annual INFORMS Revenue Management Section Conference*, June 2002. Electronic version available at `http://www.demingcenter.com/html_files/roi/pdf_2annual/DeFarias_Dynamic.pdf`.

[5] D. P. Pazel, T. Eilam, L. L. Fong, M. Kalantar, K. Appleby, and G. Goldszmidt, "Neptune: a dynamic resource allocation and planning system for a cluster computing utility," in *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*, May 2002, pp. 48–55.

[6] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.

[7] J. R. Gallardo, D. Makrakis, and M. Angulo, "Dynamic resource management considering the real behavior of aggregate traffic," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 177–185, June 2001.

[8] K. Kant and Y. Won, "Server capacity planning for web traffic workload," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 5, pp. 731–747, 1999.

[9] M. Arlitt and C. L. Williamson, "Internet web servers: workload characterization and performance implications," *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, pp. 631–645, 1997

[10] M. Arlitt, D. Krishnamurthy, and J. Rolia, "Characterizing the scalability of a large web-based shopping system," *ACM Transactions on Internet Technology*, vol. 1, no. 1, pp. 44–69, Aug. 2001.

[11] M. S. Squillante, D. D. Yao, and L. Zhang, "Web traffic modeling and web server performance analysis," in *Proceedings of the 38th Conference on Decision and Control*, Dec. 1999, pp. 4432–4439.

[12] J. L. Hellerstein, F. Zhang, and P. Shahabuddin, "An approach to predictive detection for service management," in *Proceedings of the 6th IFIP/IEEE International Symposium on Integrated Network Management*, May 1999, pp. 309–322.

[13] D. Shen and J. L. Hellerstein, "Predictive models for proactive network management: application to a production web server," in *2000 IEEE/IFIP Network Operations and Management Symposium*, Apr. 2000, pp. 833–846.

[14] F. Zhang and J. L Hellerstein, "An approach to on-line predictive detection," in *Proceedings of the 8th International Symposium on Modeling, Analysis and Simulation of Computers and Telecommunication Systems*, Aug. 2000, pp. 549–556.

[15] A. P. Douglas, A. M. Breipohl, F. N. Lee, and R. Adapa, "Risk due to load forecasting uncertainty in short-term power system planning," *IEEE Transactions on Power Systems*, vol. 13, no. 4, pp. 1493–1499, Nov. 1998.

[16] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558–1573, Dec. 1987.

[17] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," *IEEE Transactions on Power Systems*, vol. 4, no. 4, pp. 1484–1491, Nov. 1989.

[18] F. J. Nogales, J. Contreras, A. J. Conejo, and R. Espinola, "Forecasting next-day electricity prices by time series models," *IEEE Transactions on Power Systems*, vol. 17, no. 2, pp. 342–348, May 2002.

[19] J. H. Park, Y. M. Park, and K. Y. Lee, "Composite modeling for adaptive short-term load forecasting," *IEEE Transactions on Power Systems*, vol. 6, no. 2, pp. 450–457, May 1991.

[20] R. Ramanathan, R. Engle, C. W. J. Granger, F. Vahid-Araghi, and C. Brace, "Short-run forecasts of electricity loads and peaks," *International Journal of Forecasting*, vol. 13, pp. 161–174, 1997.

[21] S. Sargunaraj, D. P. Sen Gupta, and S. Devi, "Short-term load forecasting for demand side management," *IEE Proceedings on Generation, Transmission, and Distribution*, vol. 144, no. 1, pp. 68–74, Jan. 1997.

[22] S. Basu, A. Mukherjee, and S. Klivansky, "Time series models for Internet traffic," in *Proceedings of the IEEE Conference on Computer Communications*, Mar. 1996, vol. 2, pp. 611–620.

[23] J.-C. Bolot and P. Hoschka, "Performance engineering of the World Wide Web: application to dimensioning and cache design," *Computer Networks and ISDN Systems*, vol. 28, no. 7/11, pp. 1397–1405, May 1996.

[24] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis, Forecasting and Control*, 3rd ed., Englewood Cliffs, NJ: Prentice Hall, 1994.

[25] K. Chandra and A. E. Eckberg, "Traffic characteristics of on-line services," in *Proceedings of the second IEEE Symposium on Computers and Communications*, July 1997, pp. 17–21.

[26] Y.-W. Chen, "Traffic behavior analysis and modeling of sub-networks," *International Journal of Network Management*, vol. 12, pp. 323–330, May 2002.

[27] T. H. Li and M. J. Hinich, "A filter bank approach for modeling and forecasting seasonal patterns," *Technometrics*, vol. 44, no. 1, pp. 1–14, 2002.

[28] A. C. Harvey, *Time Series Models*, 2nd ed., Cambridge, MA: MIT Press, 1993.

[29] P. H. Franses, *Time Series Models for Business and Economic Forecasting*, Cambridge, UK: Cambridge University Press, 1998.

[30] D. C. Montgomery and E. A. Peck, *Introduction to linear regression analysis*, New York: Wiley & Sons, 1992.

[31] M. J. Hinich, "A statistical theory of signal coherence," *IEEE Journal of Ocean Engineering*, vol. 25, no. 2, pp. 256–261, Apr. 2000.

[32] J. D. Hart, *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer-Verlag, 1997.

[33] H. Lütkepohl, *Introduction to Multiple Time Series Analysis*, 2nd ed., New York: Springer-Verlag, 1993.

[34] R. J. Bhansali, "Asymptotically efficient autoregressive model selection for multi-step prediction," *Annals of the Institute of Statistical Mathematics*, vol. 48, pp. 577–602, 1996.

[35] R. Shibata, "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process," *Annals of Statistics*, vol. 8, pp. 147–164, 1980.

[36] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed., New York: Springer-Verlag, 1991.

[37] D. R. Cox, "Prediction by exponentially weighted moving averages and related methods," *Journal of the Royal Statistical Society Series B*, vol. 23, pp. 414–422, 1961.

[38] D. F. Findley, "On some ambiguities associated with the fitting of ARMA models to time series," *Journal of Time Series Analysis*, vol. 5, pp. 217–227, 1984.

[39] W. Gersch and G. Kitagawa, "The prediction of time series with trends and seasonalities," *Journal of Business and Economic Statistics*, vol. 1, pp. 253–264, 1983.

[40] J. Haywood and G. T. Wilson, "Fitting time series models by minimizing multistep-ahead errors: a frequency domain approach," *Journal of the Royal Statistical Society Series B*, vol. 59, no. 1, pp. 237–254, 1997.

[41] P. Stoica and A. Nehorai, "On multi-step prediction error methods for time series models," *Journal of Forecasting*, vol. 8, pp. 357–368, 1989.

[42] G. C. Tiao and T. Xu, "Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case," *Biometrika*, vol. 80, pp. 623–641, 1993.

[43] A. A. Weiss, "Multi-step estimation and forecasting in dynamic models," *Journal of Econometrics*, vol. 48, pp. 135–149, 1991.

[44] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Chapter 13, Upper Saddle River, NJ: Prentice Hall, 1996.

[45] J. H. Friedman, "A variable span scatterplot smoother," Technical Report No. 5, Laboratory for Computational Statistics, Stanford University, 1984.

[46] P. C. Young, D. J. Pedregal, and W. Tych, "Dynamic harmonic regression," *Journal of Forecasting*, vol. 18, pp. 369–394, 1999.

[47] M. Hasegawa, G. Wu, and M. Mizuno, "Applications of nonlinear prediction methods to the Internet traffic," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 2001, vol. 3, pp. 169–172.

[48] M. Wu, R. A. Joyce, H.-S. Wong, and S.-Y. Kung, "Dynamic resource allocation via video content and short-term traffic statistics," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 186–199, June 2001.

[49] C. You and K. Chandra, "Time series models for Internet data traffic," in *Proceedings of the 24th Conference on Local Computer Networks*, Oct. 1999, pp. 164–171.