Investigation of short-range cedar pollen forecasting

J.-J. Delaunay,* C. Seymour, and V. Fouillet
NTT Energy and Environment Systems Laboratories, 3-I Morinosato Wakamiya, Atsugi, 243-0198 Kanagawa, Japan
(Received 5 December 2003; revised manuscript received 28 July 2004; published 23 December 2004)

Pollen forecasting is of increasing interest as a way to help the general public avoid contact with allergy-inducing pollen. It was recently reported that the dynamics underlying pollen concentration series is very similar to that of low-dimensional deterministic chaos, thus opening up new avenues of development in local pollen forecasting. Our analysis of hourly cedar pollen series for two seasons showed evidence of a small degree of determinism underlying the pollen time-series dynamics. However, we could not confirm that our pollen series was generated by a low-dimensional chaotic system. The nearest-neighbor method using local constant prediction applied to hourly pollen forecasting with a 1-h lead time was effective for small to medium pollen variations, but failed to reproduce large and intermittent pollen bursts. The performance of the nearest-neighbor model was significantly improved by applying a nonlinear filter to the source dataset. Standard time-series techniques such as neural networks did not improve upon these results. The difficulty in fully characterizing and accurately forecasting the pollen series was thought to originate in the nonstationarity of the series and in the large and intermittent pollen bursts that were found to have no apparent time structure. Thus the dynamics of hourly pollen series is probably not strongly tied to a low-dimensional chaotic system.

DOI: 10.1103/PhysRevE.70.066214 PACS number(s): 05.45.−a, 87.10.+e

I. INTRODUCTION

Cedar pollen is known to be a source of potent allergens and as such has attracted interest in recent years. In Japan, cedar trees *(Crytomeria japonica)* have been extensively cultivated since World War II and used as lumber for construction nationwide. Airborne cedar pollen has become one of the major sources of allergens in the ambient air and is reported to be the main cause of pollinosis [1]. Today, more than one in ten inhabitants of the Kanto region are reported to suffer from this condition [2]. The medical community is clear in their advice that the most effective way to prevent pollinosis is to avoid inhaling allergy-inducing pollen. Knowledge of current and projected pollen concentrations would enable a person to take precautionary measures such as wearing a mask or taking appropriate medication. The ability to accurately predict the spatio-temporal variations in airborne pollen could be used to provide allergy sufferers with pollen alerts.

Unfortunately the problem of pollen forecasting is a very complex one [3,4], as it involves the simulation of time-varying three-dimensional concentrations over very large areas (~100 km), which is an inherently high-dimensional system involving many interacting variables [5]. A different approach to forecasting short-range pollen variations near a pollen observation station could come from time-series forecasting techniques. It has been reported that pollen series dynamics can be described as low-dimensional deterministic chaos [6–9], thus opening the way for short-range forecasts using nonlinear forecasting techniques such as artificial neural network models.

In this paper, we investigated hourly pollen series for two seasons with the aim of characterizing the pollen series dynamics and assessing the short-range forecast capability of time-series techniques. We first describe the pollen concentration measurement procedure and the relationship between a pollen series and a standard meteorological parameter series. We then analyze the pollen series using linear and nonlinear analysis techniques, searching for evidence of determinism. Finally, we report the forecast performance obtained with the nearest-neighbor prediction method and discuss previously reported results.

II. POLLEN AND METEOROLOGICAL DATA

The pollen concentration data were collected using an automatic pollen sampler, KH3000, from Yamato Corporation [10]. The device was placed about 10 meters above the ground on the roof of the Forestry and Forest Products Research Institute of Japan in Hachioji, Tokyo (139.2829 deg. east and 35.6422 deg. west) and was operated by T. Yokoyama for two successive years, 2000 and 2001. The measurement site is close to major cedar plantations located in the mountainous region west of Tokyo. The automatic pollen sampler used is of the particle counter type. In this type of sampler, a defined volume of air is circulated through a fine pipe that is intersected by a laser beam. When a particle passes through the laser beam, a scattered signal is detected whose intensity is related to particle size and optical index. In the KH3000, two laser beams and their respective detectors are used to compare scattered intensities from the same particle but measured at different angles. If the particle is perfectly spherical and homogeneous, the two intensities will be the same. It is known that Japanese cedar pollen grains are spherical with a diameter of about 30 μm, so this method allows for a selective particle count. Only spherical particles with a similar optical index are counted as cedar pollen grains. Although this device is better than a standard particle counter, it cannot completely distinguish between

*Corresponding author.
Email address: jean@mech.t.u-tokyo.ac.jp

 ©2004 The American Physical Society

TABLE I. Matrix of linear correlation coefficients between pollen concentration, ambient temperature (T), relative humidity (RH), wind speed (WS), maximum wind speed (WSmax), and precipitation (Preci) for the 2001 Takao dataset (a total of 1416 hourly values).

	Pollen	T	RH	WS	WSmax	Preci
Pollen	1	0.37	−0.33	0.22	0.33	−0.07
T		1	−0.57	0.48	0.58	0.005
RH			1	−0.53	−0.65	0.25
WS				1	0.78	−0.14
WSmax					1	−0.06
Preci						1

pollen types, because all spherical particles that are the same size as cedar pollen grains and with a similar optical index are counted as cedar pollen grains. We only used data collected during the months of February and March, which is when the amount of cedar pollen greatly surpasses that of all other pollen types in this area. The sampling time was 1 h, the total air flow was 0.246 m^3, and the concentration was recorded in grains per m^3.

The meteorological data we used were collected at a height of 10 meters at a nearby meteorological station operated by the same institute. The distance between the pollen sampler and the station was less than 50 meters. The station takes an hourly record of ambient air temperature, average and maximum wind speed and wind direction, relative humidity, sunshine duration, and precipitation.

For the months of February and March of 2000 and 2001, the mean (standard deviation) pollen concentrations were 240 (720) and 125 (350) grains/m^3 and the mean temperatures were 4.8 (5.4) and 5.4 (5.6) °C, respectively. Table I shows the linear correlation coefficients between the meteorological parameters and the pollen concentration. We found high positive correlation values with the temperature and the wind speed. The highest correlations were obtained for the ambient temperature and the maximum wind speed. We also noted negative correlations with the relative humidity and precipitation amount. The negative correlation with precipitation is explained by the washout effect that rain has on the atmosphere, usually modeled as an exponential decay over time [11,5]. The effect of the relative humidity is more difficult to determine as humidity is correlated with precipitation and yet also strongly anticorrelated with ambient temperature (see Table I). In the literature, the flowering of some species is often described as being induced by rising temperature and falling humidity [12]. Cedar trees are probably affected in a similar way, with falling humidity leading to an increase in the pollen emission rate and consequently raising the airborne pollen concentration too. Another explanation for the effect of relative humidity might be that the average pollen grain mass increases with relative humidity because of water absorption by the hydrophilic pollen grains. As heavy particles settle faster in the atmosphere, the lifetime (suspension time in the atmosphere) of the pollen grains would thus decrease and therefore the observed concentration would decrease as well.

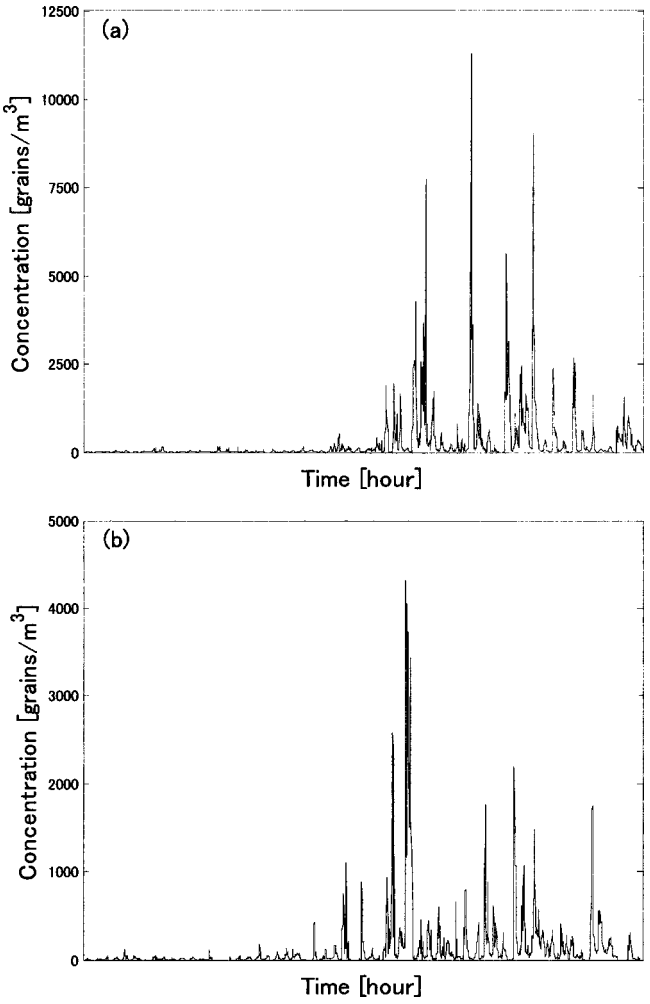

FIG. 1. Pollen time series starting February 1 and ending March 31 for(a) 2000 and (b) 2001. Hourly concentration values are plotted as a function of time.

III. POLLEN TIME-SERIES ANALYSIS

Figure 1 shows the pollen time series for the years 2000 and 2001. The series shown in Fig. 1 does not vary about a fixed level (constant mean) and exhibits some intermittent behavior typical of a nonstationary time series. Here nonstationarity is understood to be the occurrence of statistically significant variations in the estimate of parameters (such as the mean, the variance, and power spectrum) as a function of the part of the series on which they are estimated. Since the conditions that define the pollen emission (flowering and wind) and subsequent transport (mainly wind and rain) cannot be assumed to be stationary over a large time-scale range (several hours to months), nonstationarity is expected in this system. The nonstationarity of the series was confirmed by plotting the cross-sectional prediction error as described by Schreiber [13], a technique that examines nonstationarity for nonlinear series. In this method, different sections of a series are used to predict each other. If the predictions are of sufficiently varying accuracy, the series varies qualitatively over time, and thus is found to be nonstationary.

Autocorrelation curves and Fourier power spectra of the series are shown in Figs. 2 and 3, respectively. The autocor-

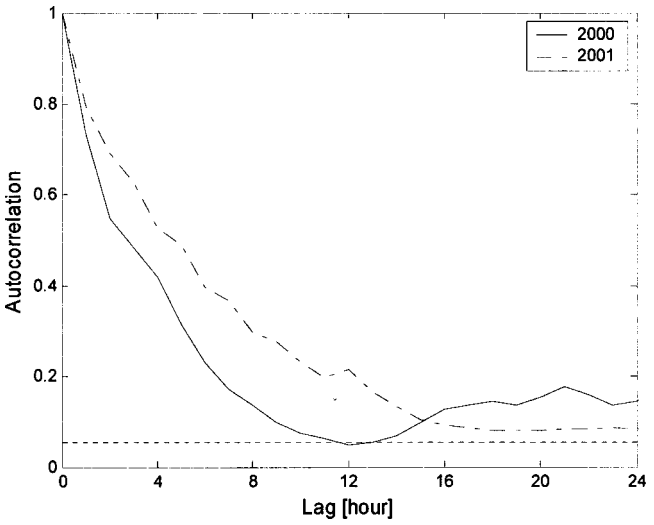

FIG. 2. Autocorrelation function of the pollen time series as a function of the lag for 2000 and 2001. The horizontal dotted line approximates the 95% confidence interval (given as the upper bound of the $\pm 2/\sqrt{N}$ interval, outside of which values are statistically significant).

relation, which indicates the similarity between adjacent values in the series, decays rapidly within the first 10 lags and then slowly decreases to a constant value larger than zero. For short lags (<3), the high autocorrelation values suggest that predicition of short-range pollen variations may be possible. The power spectrum for 2001 in Fig. 3 shows an almost continuous shape that could indicate chaotic dynamics underlying the pollen variations or colored noise. The nonlinearity of the equations of motion of the atmosphere permits chaotic solutions, so finding chaos in the variation of the pollen series could be a possibility, as previously reported [6–9].

Unfortunately, nonstationary systems are known to be very difficult to analyze and predict, and very few statistical tests are readily available for their analysis. In contrast, a large body of literature is dedicated to the analysis of deterministic chaos and the reconstruction of its dynamics (for a review, see [14]). In the following, we first briefly introduce the time-delay reconstruction method that is widely used in the analysis of chaotic time series, and then we investigate our pollen series using the correlation integral technique, the largest Lyapunov exponent technique, and Casdagli's test for evidence of deterministic chaos.

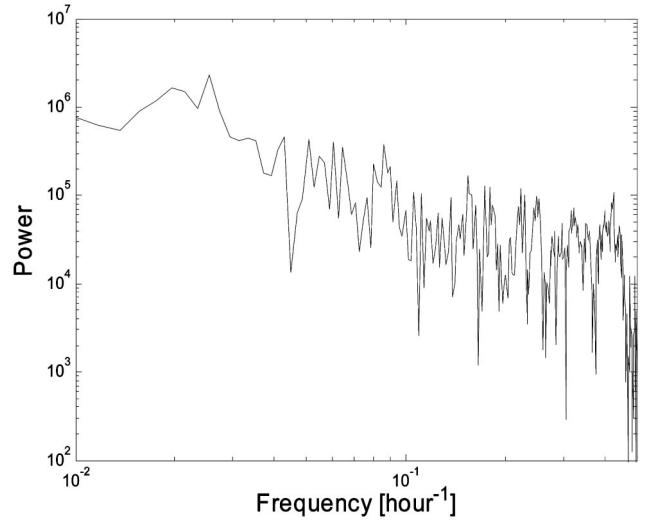

FIG. 3. Fourier power spectrum for 2001.

A. Time-delay reconstruction

Techniques for analyzing the dimensionality of the chaotic dynamics of a scalar time series are based on the time-delay reconstruction method introduced by Packard *et al.* [15] and later formalized by Takens [16]. With this method, state vectors in an embedding space are formed from time-delayed values of the N- point time series $\{x_1, x_2, \ldots, x_N\}$ as such,

$$X_t(m) = \left(x_t, x_{t-\tau}, \ldots, x_{t-(m-1)\tau}\right), \tag{1}$$

where m is the embedding dimension and τ is the delay time or lag.

B. Correlation integral

The correlation integral introduced by Grassberger and Procaccia [17] is one of the most commonly used techniques for detecting the presence of low-dimensional deterministic chaos in data. It is defined as

$$C_m(r) = \frac{2}{(N-m)(N-m+1)} \sum_{i=m}^{N} \sum_{j<i} \theta\left(r - |X_i(m) - X_j(m)|\right), \tag{2}$$

where θ is the Heaviside function.

$C_m(r)$ is interpreted as the fraction of pairs of points that are separated by a distance less than or equal to r. For a deterministic chaotic series, $C_m(r)$ behaves as the power law of r for small r (the scaling law),

$$C_m(r) \propto r^{\gamma(m)}, \tag{3}$$

where $\gamma(m)$, for a large enough m, tends to the correlation dimension of the system. The correlation dimension provides an estimate of the number of degrees of freedom excited in the system.

Applying the $C_m(r)$ technique to our data, we first check for the existence of a scaling region and then investigate the behavior of $\gamma(m)$ in relation to m. Figure 4 shows a log-log plot of $C_m(r)$ for the two seasons as a function of the embedding dimension m. We found that, for a limited scaling region ($r=10$ to 50 grains/m^3), the $C_m(r)$ curve slopes converge to a more or less constant value as the embedding dimension is increased. In addition, the $C_m(r)$ sequences for the two seasons behave differently, with Takao 2001 agreeing more closely with the scaling law than Takao 2000. The embedding dimension is estimated to be in the 5–10 range for both seasons, as the slope of the $C_m(r)$ sequences becomes roughly independent of the embedding dimension for $m \gtrsim 6$. The correlation dimension estimated from the slopes

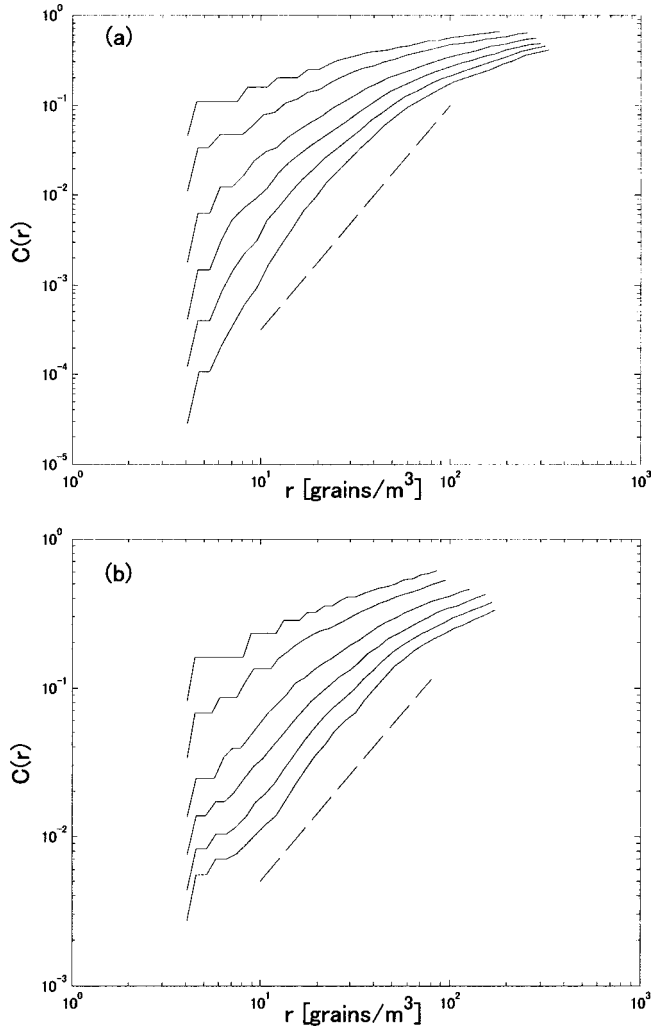

FIG. 4. Correlation integrals computed with $m=1$, 2, 4, 6, 8, and 10 for (a) Takao 2000 and (b) Takao 2001. Higher curves correspond to lower embedding dimensions. The dashed straight line corresponds to the scaling law with its slope set at 2.5 and 1.5 for Takao 2000 and 2001, respectively.

taken at large m was found to vary between 1.5 and 2.5. We also found that applying the techinique to the first part of the data sets (before the onset of the large peaks) resulted in correlation integral plots that have a greater similarity to those generated by chaotic processes and follow the scaling law more closely than plots generated using the complete datasets. From this we concluded that the linear parts seen in the correlation integral plots were generated by small to medium variations (<500 grains/m^3) in the pollen series.

The correlation integral techinique has been developed and largely tested on time series generated from the numerical integration of equations and, as such, presents the following difficulties when applied to real-world time series: (1) The number of points avalilable from a neutral times series is often too small compared with the ideal number of samples required by the technique, (2) the stationarity hypothesis for natural series is often unfulfilled, (3) the level of noise is high in natural series (here it is on the order of 10 grains/m^3), and (4) the range of dynamics available for

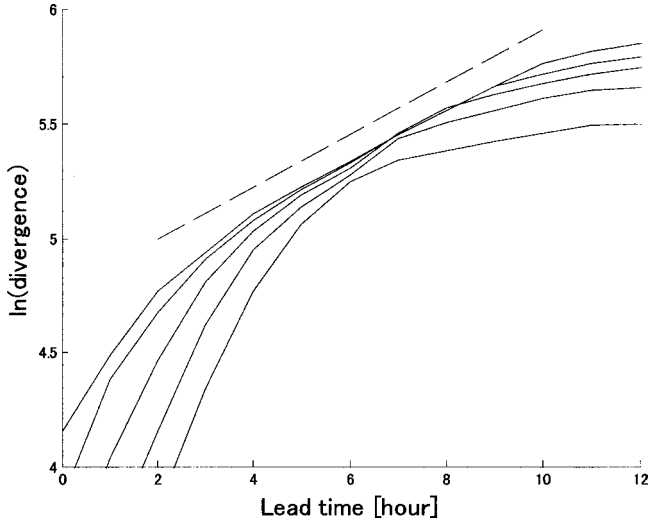

FIG. 5. Variations of the logarithm of the divergences as a function of the lead time with $m=6$, 7, 8, 9, and 10. Lower curves correspond to lower embedding dimensions. The dashed straight line is a guide to the eye indicating the linear part of the divergence curves and has its slope set at 0.11.

analysis is narrow, as in our case the variation span is less than three decades from 10 to 1000. It should also be noted that the steps seen in our graphs for small r result from the discretization of the pollen time series [18].

In conclusion, the small to medium variations of the pollen series behave similarly to that of a low-dimensional chaos dynamic system, but the dimension of the system dynamics underlying the data could not be estimated with confidence using the correlation integral technique. Finally, it should be noted that a positive outcome with the correlation integral technique does not necessarily indicate a chaotic process, as nonchaotic processes have been reported to follow the scaling law (e.g., time series generated by a colored stochastic process [19]).

C. Largest Lyapunov exponent

The Lyapunov exponents measure the rate of divergence of trajectories having nearby initial conditions, that is, they measure sensitivity to initial conditions [20]. For a chaotic series, the largest Lyapunov exponent is positive and provides an estimate of the level of chaos in a dynamical process.

We applied the method described by Rosenstein *et al.* to our data to estimate the largest Lyapunov exponent. This particular method was selected because it is suited to small data sets [21]. The results we obtained with this method are shown in Fig. 5. We found that the graphs of the logarithm of the divergences as a function of time showed some linear regions for $m \gtrsim 7$, with a slope of ~ 0.1. The graph resembled that of a low-dimensional chaotic series, but also showed some similarities with series generated by a stochastic process. It is notoriously difficult to distinguish between these two types of processes using this technique. As with the correlation integral technique, the extraction of the linear region is somewhat arbitrary and the technique delivers results that

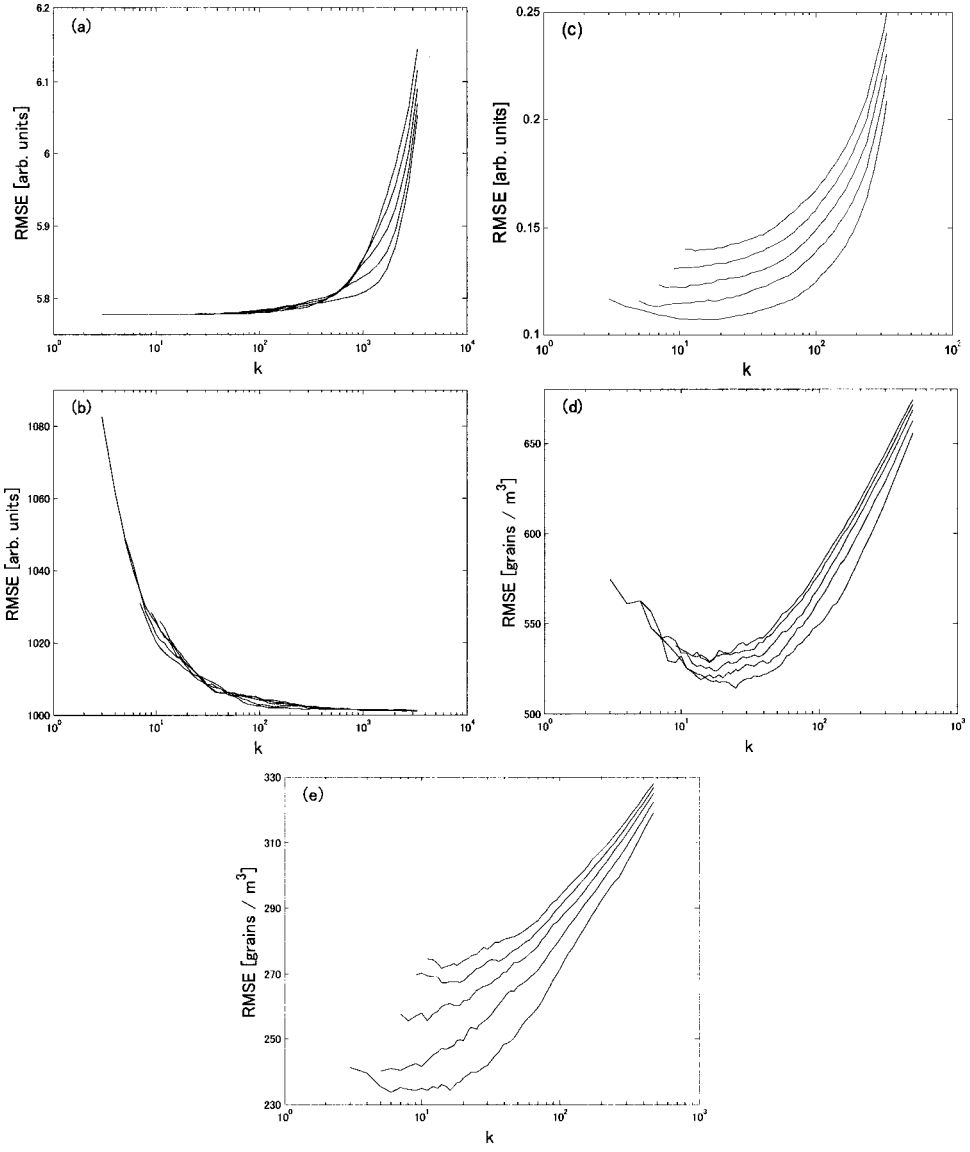

FIG. 6. Modified Casdagli prediction error plots for (a) Lorenz series, (b) white noise (c) colored noise (d) Takao 2000 pollen series, and (e) Takao 2001 pollen series. The RMSE is shown as a function of k in a semilog plot for embedding dimensions of 2, 4, 6, 8, and 10. Lower curves correspond to lower embedding dimensions.

are difficult to interpret for our data. Nevertheless, this analysis indicates the posssibility that our pollen series may be chaotic.

D. Casdagli's test

Casdagli [22] developed an exploratory technique for investigating the underlying dynamics that generates a time series and for detecting low-dimensional deterministic chaos as opposed to stochastic or high-dimensional behavior. The strategy is to investigate variations in the prediction error of the k-nearest-neighbor (local linear) forecasting algorithm in relation to k, the number of neighbors. Prediction at low k is close to (nonlinear) deterministic modeling, while high k corresponds to stochastic linear modeling. A general increase in the prediction error with increasing k is strong evidence of nonlinear determinism, whereas the opposite trend is evidence of a stochastic process. In his paper, Casdagli uses local-linear prediction as forecasting algorithm, although he suggests that other techniques could also be used. We used zeroth-order (average of the k-nearest-neighbor) prediction in our study, as the local linear algorithm involves solving a set of linear equations that may be ill-defined in our pollen series due to discrete variations in the pollen concentration and the occurrence of large spikes. In our case, the average of the k-nearest neighbors proved to be more robust than the local-linear technique. Further, we used a causality window with a width of 10 h to reduce the effect of self-correlation [23]. To test our slight modification of Casdagli's technique, we generated prediction error plots for reference series such as a low-dimensional chaotic series produced by Lorenz equations (parameters as in [14]), a white noise series, and a colored noise series, shown in Figs. 6(a)–6(c), respectively. With the Lorenz series, the root-mean-square error (RMSE) is small for small k values and increases significantly as k is increased, revealing the low-dimensional determinism underlying the series. For white noise, the opposite trend is observed, with a reduction in error as k is increased, demonstrating that these data are better modeled as a stochastic process than as a low-dimensional deterministic one. Among the different types of colored noise, we chose red noise because its autocorrelation function decays with the lag, a

property also observed with the pollen series. The red noise series was generated by integrating the Ornstein-Uhlenbeck linear differential equation, which combines a deterministic exponential decay with additive normally distributed noise [24]. We found that the red noise series behaved similarly to the Lorenz series in the Casdagli's test. The red noise series' increase in error with k is explained by the presence of some autocorrelation in the series. The variations of the red noise series are best represented as a local model. Thus, certain types of colored noise can mimic the behavior of low-dimensional chaos in the plot from Casdagli's test, which renders the distinction between low-dimensional chaos and colored noise difficult. Figures 6(d) and 6(e) show prediction error plots for the Takao 2000 and 2001 pollen time series, respectively. For the pollen time series, the error variations with k show a dip at k values of about 10 to 30. The initial fall in prediction error is due to the noise-canceling effect of the nearest-neighbor technique. The difference between the minimum error at the dip and the larger errors as k is increased indicates that our data are better modeled by a local approach than a global one. These results suggest that our data feature at least some level of determinism and that it is worth trying to use the k-nearest-neighbor technique to forecast pollen series.

IV. FORECASTING MODEL

We used various configurations of the k-nearest-neighbor prediction scheme to compute forecasts with a horizon of 1 h. We varied the embedding dimension m and the number of nearest neighbors k of the prediction scheme, tried regular zeroth-order and distance weighted zeroth-order approaches, and applied a nonlinear filtering technique and dithering to the data. Pseudofalse nearest neighbors were not included in the calculations, as these points representing true neighbors in the reconstructed state space lie on different trajectories of the map of the system [25].

We optimized and validated our nearest-neighbor model using a two-step method. First we calculated in-sample results by running the nearest-neighbor model at various settings for m and k, using data from the first half of 2000 as the training set (database), and data from the second half as the test set. Then, we used the best m and k settings from the previous in-sample testing to compute the final prediction accuracy (out-of-sample results). For this stage we used the data from the first half of 2000 as the training set (database) and the data from 2001 as the validation set. In both stages, the database was expanded with vectors which had already been used for prediction in the test/validation set, as predictions were generated sequentially in time.

The time periods used for the training, test, and validation sets were February 1, 2000–March 19, 2000, March 20, 2000–March 31, 2000, and February 1, 2001–March 31, 2001, respectively. We used the RMSE and the linear correlation coefficient (CORR) as the performance criteria. The best performing model was obtained for $m=4$ and $k=12$. This model gave an RMSE of 562 grains/m^3 and a CORR of 0.69 on the test set, and an RMSE of 214 grains/m^3 and a CORR of 0.80 on the validation set. For comparison purposes, it should be noted that a naïve model ($x_{i+1}=x_i$) gives an RMSE of 225 grains/m^3 and a CORR of 0.79 for the same validation set. The nearest-neighbor model has better performance than the naïve model in terms of the RMSE and the CORR coefficient criteria.

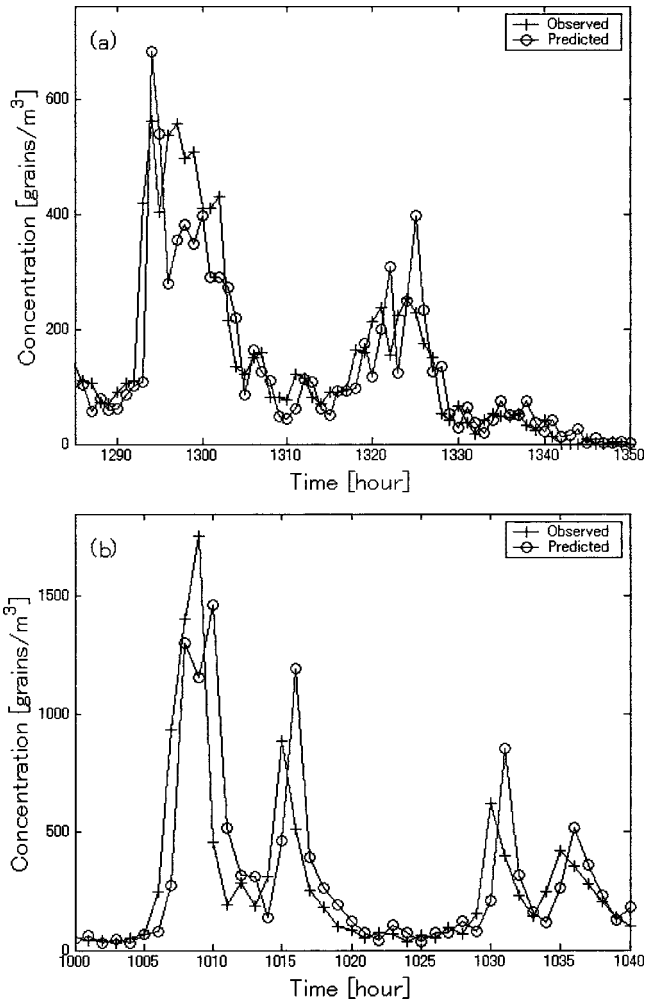

FIG. 7. Variations in the observed and predicted pollen concentration as a function of time for two data regions (a) a region showing good agreement and (b) an intermittent region showing relatively poor agreement.

We obtained improvements in model performance when using the GHKSS nonlinear filter from the TISEAN package [26]. This filter attempts to reduce the dimensionality (noise) of the data without smoothing out the peaks, a result that cannot be achieved with standard linear filtering techniques such as a moving average. We obtained the best test set results when we used the GHKSS filter, configured to reduce the dimensionality from 7 to 5, with $m=4$ and $k=18$. This model gave an RMSE of 558 grains/m^3 and a CORR of 0.70 on the test set, and an RMSE of 207 grains/m^3 and a CORR of 0.81 on the validation set. The improvement in RMSE when using the GHKSS filter is evidence for better modeling as opposed to averaging. Figure 7 shows pollen concentration time series predicated using this model together with the observed series. Good agreement is observed in Fig. 7(a) for a region of the series that varied relatively slowly, whereas

Fig. 7(b) reveals poor agreement for an intermittent region of the series.

We also tried dithering the data with Gaussian noise to avoid problems arising from the discrete nature of the data [27], but it did not significantly improve the model performance of the GHKSS filtered data. For the unfiltered data, however, we observed a slight improvement with dithering.

We confirmed that the results obtained with the *k*-nearest-neighbor model and the GHKSS filter were not improved upon by instead using the standard linear autoregressive integrated moving average model [28] or the feed forward artificial neural network model (e.g., [29,30]). Here we used the same data division as in the nearest-neighbor model: training, test, and validation subsets, according to the cross-correlation data division method [31], with time periods of February 1, 2000–March 19, 2000, March 20, 2000–March 31, 2000, and February 1, 2001– March 31, 2001, respectively. The best linear model gave an RMSE of 580 grains/m^3 and a CORR of 0.68 on the test set, and an RMSE of 218 grains/m^3 and a CORR of 0.8 on the validation set. The best neural network gave an RMSE of 587 grains/m^3 and a CORR of 0.65 on the test set, and an RMSE of 207 grains/m^3 and a CORR of 0.81 on the validation set. The best neural network model result was the same as that of the *k*-nearest-neighbor model when used in combination with a GHKKS filter. The main difficulty with the neural network approach originated in the training phase, when the neural network learned to reproduce the large peaks, since they transferred large errors to the back-propagation scheme. The resulting neural networks had poor generalization ability. This was improved somewhat by changing the error function used by the back-propagation scheme from a mean-square error to an absolute mean error function, which is less sensitive to large errors. Moreover, for the neural network approach to work effectively, more input data such as meteorological data should be used in the modeling.

In the end, we found that the *k*-nearest-neighbor model provided robust hourly forecasts of satisfactory quality for periods of time without large pollen bursts. This model has the advantage of relative simplicity compared with the neural network approach. The predictability of sections of the series lacking in burst supports our earlier assertion that our pollen series contain some determinism for parts with small to medium variation. Unfortunately, information on the occurrence of the pollen bursts is the most important with regard to helping people to take precautions such as avoiding contact with large amounts of airborne pollen and their associated allergens.

V. DISCUSSION

Reports by Bianchi *et al.* [6] and Arizmendi *et al.* [7,8] conclude that pollen time series feature low-dimensional chaotic dynamics with an attractor dimension of 0.66, as estimated by the correlation integral and further confirmed by wavelet analysis. It was also reported that the neural network forecast technique with an embedding dimension of 6 provides accurate forecasts for 1- and 12-h lead times without any significant degradation in forecast performance for the extremely large lead time of 12 h. In light of our study, we should like to comment on these results.

Arizmendi and Bianchi's study of pollen series dynamics is based on total pollen counts covering different pollen species (18 were quoted) over their respective seasons. By analyzing the complete set simultaneously, one is looking at different systems (pollen species), which may not necessarily share the same dynamics. Different plants may produce and emit their pollen in different ways in terms of their response to weather conditons.

Also, the estimate of the correlation dimension using the Grassberger and Procaccia algorithm is notoriously difficult and should be interpreted with great care (see [18,32] for examples of erroneous conclusions that can be drawn when using the method). In particular, the intermittent variations found in pollen series call for caution when interpreting the correlation integral. The above references did not provide the correlation integral graphs for different embedding dimensions, making it difficult to judge the resemblance of the correlation integral graphs with that of low-dimensional chaotic dynamics. For our data, the graphs were not in perfect agreement with the scaling law, so we could not have great confidence in the estimated correlation dimensions. The low dimension of 0.66 reported in the above studies means that the observed pollen variations could potentially be explained by a model with only a very few degrees of freedom, in this case $\simeq 1$, which seems unreasonable for such a complex natural system.

The forecast results generated using a neural network with a 1-h lead time (see Figs. 2 and 3 of Ref. [7]) are doubtful as they agree almost perfectly for the intermittent pollen variations, whose occurrence demonstrates no apparent time structure. Furthermore, there seems to be no significant degradation in the performance when the lead time is increased from 1 to 12 h (see Figs. 4–7 in [7]), which contradicts the reported chaotic nature of the series. A fundamental property of chaotic time series is the significant degradation in the forecast accuracy when the lead time is increased within the limit given by the return of skill of the system dynamics; see, for example, [33].

In a different study, Arizmendi *et al.* [8] published results obtained by applying the surrogate data test to pollen data. The surrogate data method tests for a nonlinear process underlying the data. These results apparently reinforced their previous conclusion on the chaotic nature of pollen time series. Their data (Fig. 1) reveal intermittent large spikes in the time series, marking it as nonstationary, since statistics such as the mean and variance no doubt vary significantly in different sections of the data set. As such, their application of the surrogate data testing technique is inappropriate [34]. Timmer [35] shows that nonstationarity can produce spurious results in surrogate data testing and concludes that a positive test result does not necessarily indicate a nonlinear or chaotic process, and can result from a nonstationary process instead.

Recently, the same group applied wavelet-based fractal analysis to the same pollen series [9] and concluded that the strange attractor underlying the pollen series is better characterized by the spectrum of generalized fractal dimensions [36] obtained by the wavelet technique than by the Grass-

berger correlation dimension. The main conclusion relates to the loss of information with lead time. It is unclear whether or not the previously reported low-dimensional chaotic dynamics of the pollen series was confirmed by this method.

We believe that the body of evidence supporting the low-dimensional chaotic dynamics of pollen concentration time series should be enlarged before a definitive conculsion can be drawn. The nonstationarity of the phenomenon means that most of the existing analysis techniques do not strictly apply. Nevertheless, we found some similarities in the overall behavior of our pollen series to those generated by low-dimensional chaotic dynamics. It is possible that the pollen time series could be described as a low-dimensional chaotic system with high-dimensional noise, which is also consistent with the fact that pollen emission is chiefly governed by a few meteorological parameters (temperature, wind, and rain) but, at the same time, also features very large degrees of freedom that cannot realistically be accounted for in a deterministic way (e.g., wind gusts and turbulence generated by complex terrain). The intermittent pollen bursts probably originate in these variables. We felt that the large number of pollen burst events observed during one season made it difficult to describe the series as low-dimensional chaos, although Casdagli's test apparently revealed a deterministic aspect of the pollen series. The nearest-neighbor prediction method proved useful in forecasting the pollen series with a 1–h lead time, but we do not think that hourly forecasts with lead times longer than 3 h are a realistic objective in relation to the investigated time-series techniques. In particular, we found the onset of large pollen bursts to be a limiting factor in forecasting hourly pollen concentration series.

VI. CONCLUSION

We studied the natural variations in the concentration of airborne cedar pollen grains using both standard linear and nonlinear prediction techniques. We found the Fourier spectrum of the pollen time series to have an almost continuous shape and its largest Lyapunov exponent to be positive, both properties suggesting a possibly chaotic behavior of the time series, as previously reported. However, the nonstationarity of the pollen series made it difficult to use the existing analysis tools and to reach a positive conclusion about the nature of the pollen series. Moreover, the existence of large intermittent pollen peaks, which could not be predicted accurately because of an apparent lack of time structure, argues against the pollen series arising from low-dimensional chaotic dynamics. Nevertheless, the results of Casdagli's test indicated that our data may feature some level of determinism, which led us to attempt short-range forecasting of pollen concentrations.

We applied the nearest-neighbor technique to forecasting pollen series with a 1-h lead time and found that it could accurately predict small to medium variations but that the large and intermittent pollen bursts could not be correctly predicted on an hourly basis. Filtering the data using a nonlinear filter was found to improve the overall model accuracy. The nearest-neighbor model combined with a nonlinear filter (GHKSS) is a good candidate to apply in short-term pollen forecasting, because it exhibited good performance and was relatively simple to implement compared with other methods.

Further research is required to clarify the short-term dynamics underlying variations in pollen concentration time series.

ACKNOWLEDGMENT

We are very grateful to Toshitaka Yokoyama of the Forestry and Forest Products Research Institute of Japan for providing the pollen and meteorological data.

[1] T. K. Ishizaki, R. Koizumi, Y. Ikemori, Y. Ishiyama, and E. Kushibiki, Ann. Allergy **58**, 265 (1987).
[2] Tokyo Metropolitan Public Health Office. 2002. Pollen allergy (in Japanese).
[3] S. Kawashima and Y. Takahashi, Actual Pediatr. (Granada) **34**, 142 (1995).
[4] J.-J. Delaunay, K. Fedra, and M. Kubat, Arch. Complex Environmental Studies (to be published).
[5] P. Arya, *Air Pollution Meteorology and Dispersion* (Oxford University Press, New York, 1999).
[6] M. M. Bianchi, C. M. Arizmendi, and J. R. Sanchez, Int. J. Biometeorol **36**, 172 (1992).
[7] C. M. Arizmendi, J. R. Sanchez, N. E. Ramos, and G. I. Ramos, Int. J. Biometeorol **37**, 139 (1993).
[8] C. M. Arizmendi, J. R. Sanchez, and M. A. Foti, Fractals **3**, 155 (1995).
[9] M. E. Degaudenzi and C. M. Arizmendi, Phys. Rev. E **59**, 6569 (1999).
[10] Yamato Corporation, Japanese Patent No. 10-318908 (1998).
[11] A. C. Chamberlain, *The Calculation of Precipitation Scavenging*, in *Meteorology and Atomic Energy*, edited by D. N. Slade (U. S. AEC Technical Information Center, Oak Ridge, TN, 1953).
[12] L. Moseholm, E. R. Weeke, and B. N. Petersen, Pollen Spores **29**, 305 (1987).
[13] T. Schreiber, Phys. Rev. Lett. **78**, 843 (1997).
[14] H. D. I. Arbabanel, Rev. Mod. Phys. **65**, 1331 (1993).
[15] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).
[16] F. Takens, in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 1981).
[17] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[18] J. Theiler, J. Opt. Soc. Am. A **7**, 1055 (1990).
[19] A. R. Osborne and A. Provenzale, Physica D **35**, 357 (1989).
[20] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, Physica D **16**, 285, (1985).
[21] M. T. Rosentein, J. J. Collins, and C. J. De Luca, Physica D **65**, 117 (1992).
[22] M. Casdagli, J. R. Stat. Soc. Ser. B. Methodol. **54**, 303 (1991).

[23] J. Theiler, Phys. Rev. A **34**, 2427 (1986).
[24] G. E. Uhlenbeck and L. S. Ornstein, Phys. Rev. **36**, 823 (1930).
[25] G. Xiaofeng and C. H. Lai, Phys. Rev. E **60**, 5463 (1999).
[26] R. Hegger, H. Kantz, and T. Schreiber, Chaos **9**, 413 (1999).
[27] M. Möller, W. Lange, F. Mitschke, N. B. Abraham, and U. Hübner, Phys. Lett. A **138**, 176 (1989).
[28] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis*, 3rd ed. (Prentice Hall International, Englewood Cliffs, NJ, 1994).
[29] W. W. Hsieh and B. Tang, Bull. Am. Meteorol. Soc. **79**, 1855 (1998).
[30] H. R. Maier and G. C. Dandy, Environ. Modelling Software **15**, 101 (2000).
[31] M. Stone, J. R. Stat. Soc. Ser. B. Methodol. **36**, 111 (1974).
[32] P. Grassberger, Nature (London) **323**, 609 (1986).
[33] G. Sugihara and R. M. May, Nature (London) **344**, 734 (1990).
[34] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[35] J. Timmer, Phys. Rev. E **58**, 5153 (1998).
[36] H. G. E. Hentschel and I. Procaccia, Physica D **8**, 435 (1983).