Calibration of Spatial Forecasts from Citizen Science Urban Air Pollution Data with Sparse Recurrent Neural Networks

Matthew Bonas\textsuperscript{1} and Stefano Castruccio\textsuperscript{1}

May 10, 2021

Abstract

With their continued increase in coverage and quality, data collected from personal air quality monitors has become an increasingly valuable tool to complement existing public health monitoring systems over urban areas. However, the potential of using such ‘citizen science data’ for automatic early warning systems is hampered by the lack of models able to capture the high-resolution, nonlinear spatio-temporal features stemming from local emission sources such as traffic, residential heating and commercial activities. In this work, we propose a machine learning approach to forecast high-frequency spatial fields which has two distinctive advantages from standard neural network methods in time: 1) sparsity of the neural network via a spike-and-slab prior, and 2) a small parametric space. The introduction of stochastic neural networks generates additional uncertainty, and in this work we propose a fast approach for forecast calibration, both marginal and spatial. We focus on assessing exposure to urban air pollution in San Francisco, and our results suggest an improvement of 35.7\% in the mean squared error over standard time series approach with a calibrated forecast for up to 5 days.

\textsuperscript{1}Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA.
1 Introduction

According to the World Health Organization, exposure of the population to ambient air pollution is the cause of approximately 4.2 million premature deaths globally each year, of which approximately 60,000 occur in the United States (US, Zhang et al. (2018)). While some fatalities can be attributed to extreme events such as wildfires, most deaths in urban environments are caused by a systematic exposure to degraded air quality from local sources such as traffic, residential heating and commercial activities (Goodkind et al., 2019). Therefore, it is of high scientific interest to develop methods to understand and predict these local spatio-temporal patterns, in order to improve early warning systems and minimize population exposure (Monteiro et al., 2005; Carlsten et al., 2020).

From an applied perspective, even in developed countries such as the US the high quality monitoring network maintained by the Environmental Protection Agency (EPA) is not sufficiently dense in space to capture the urban variability of air pollution. For example, in San Francisco, which will be the focus of our applications in this paper, only one EPA site is currently available. Alternative satellite data sources such as the Moderate Resolution Imaging Spectroradiometer (MODIS) are very sparse in space and time, and such optical measures are not directly relatable to ground observations (Levy et al., 2010). Proxy methods such as land use regression, which links air pollution to other urban covariates easier to access such as traffic, altitude, temperature or other weather variables, have a varying degree of accuracy (Briggs et al., 2000). Numerical chemical transport models such as the Weather and Research Forecasting coupled with Chemistry (WRF-Chem, Grell et al. (2005)) rely instead on emission inventories at maximum resolution of a few kilometers in the US and coarser in other world regions, a scale vastly insufficient to characterize the scales of variability of urban emissions. This work aims at leveraging a more recent and substantially less scientifically investigated means to monitor urban air pollution: citizen science data. As of 2021, PurpleAir is one of the most popular air quality sensor networks, which since the commercialization of the first sensors in 2015 has seen an exponential increase in number of privately owned monitoring stations in cities in the US and beyond. This work demonstrates how this new source of data can be used to produce urban maps of population exposure which can not be obtained using only EPA data.

From a methodological point of view, a forecast and quantification of the spatio-temporal uncertainty of high-resolution, high-frequency air pollution at urban/local scale (<5km) requires accounting for the turbulent effect of local processes which cannot be effectively captured with standard dynamical models such as AutoRegressive Moving Average (ARMA), or spatial versions thereof (Huang et al., 2021). More flexible approaches to characterize
nonlinear dynamics have been developed in the field of machine learning, with the use of neural networks (NNs). While these recurrent NNs (RNNs) are considerably more flexible in characterizing multivariate (possibly spatial) data resolved in time, the computational time for inference is substantial, and approaches such as explicit computation of gradients (backpropagation) cannot be directly applied in a dynamical setting as they would lead to numerical instabilities (vanishing or exploding gradient). Additionally, a robust forecasting methodology requires an assessment and calibration of the uncertainty, both marginally at each sampling site and jointly to allow prediction of spatial averages or differences. Methods for assessing uncertainty in NNs (or deep versions thereof) such as dropout or dilution \cite{Hinton12} predicated on random deletion of nodes to assess the robustness of the network, or Bayesian approaches \cite{Graves11,Blundell16} imply an additional computational overhead. This computational burden may be acceptable in a static problem, but unfeasible in a dynamic setting with high-frequency data where decisions need to be made in real-time.

To reduce the computational burden for both forecasting and uncertainty quantification, a modification of the RNN predicated on the use of stochastic networks is proposed. This Echo-State Network (ESN, \cite{Jaeger01}) can be regarded as a hierarchical model with a spike and slab prior from a statistical perspective. The advantage of a hierarchical sparse approach compared to traditional RNNs is two-fold: 1) sparse matrices allows for fast sparse linear algebra inference and 2) the parametric space is considerably smaller than all the entries of the weight matrices in standard RNNs. While ESNs have been previously used for forecasting air pollution \cite{Xu19}, as well as hospitalizations from degraded air quality \cite{Araujo20}, these works did not focus on quantifying the prediction uncertainty and were used on a single sampling site. In a series of recent works \cite{McDermott17,McDermott18,McDermott19}, ESNs were used to forecast nonlinear spatio-temporal fields and the authors proposed the use of an ensemble of forecasts to quantify the uncertainty, but their approach was focused on short-term forecasting without calibration, and did not account for a natural expansion of the uncertainty through time expected in long-range forecasting. Besides the marginal forecasts, in many applications such as monitoring urban air pollution, it is of interest to characterize the spatial dependence as one might want to account for uncertainty of (local) spatial averages or differences. In cases of large spatial fields, a full analysis could prohibitively increase the computational burden of the ESN. In this case, the use of dimension reduction methods for spatial data in conjunction with ESNs was proposed in a series of recent works \cite{McDermott18,McDermott19,Huang21}. When the number of sampling sites is however small and spatial interpolation is of interest (as in our application with monitoring stations), an explicit characterization of the spatial dependence is necessary.
In this work we propose a faster and more explicit method for calibrating long-lead forecasts of spatial fields predicated on post hoc marginal calibration, expansion of long-lead uncertainty with monotonic splines as well as explicit spatial modeling. Our approach effectively de-couples the model predictions from their calibration and proposes a method which could be applied to any other spatio-temporal model. The spatial dependence is described through a non-stationary correlation function obtained through convolution at fixed knots, which allows to obtain spatially interpolated fields along with their uncertainty.

The work proceeds as follows. Section 2 introduces the air pollution as well as the population data for San Francisco. Section 3 presents the model, along with the inferential process, the forecasting and the calibration method. Section 4 presents a collection of simulation studies to assess the performance of the model in terms of point forecast and uncertainty quantification against standard time series methods. Section 5 presents the results of the model in terms of air pollution forecasting and population exposure. Section 6 concludes with a discussion. The code for this work is available at the following GitHub repository: github.com/MBonasND/2021ESN.

2 Data

We consider concentrations of particulate matter smaller than 2.5 µm in aerodynamic diameter (PM$_{2.5}$) from PurpleAir, one of the widest citizen science networks measuring real-time air quality conditions. The PurpleAir network uses PA-II sensors which, as of 2021, cost 249USD compared to Environmental Protection Agency (EPA) sensors, whose cost is in the range of thousands of USD (Ardon-Dryer et al., 2020), for monitoring changes in the PM$_{2.5}$ concentrations. The PurpleAir sensors use PMSX003 laser counters to measure particulate concentrations in real time, while EPA sensors collect PM$_{2.5}$ through a filter which is weighted to determine the particulate mass concentration, a system considerably more sophisticated and expensive. The instrument precision, expressed as maximum percentage difference between two PA-II sensors inside each device has been estimated as ±10% on concentrations less than 500 µgm$^{-3}$ (personal communication with the manufacturer). PA-II sensors are considered the standard for citizen science outdoor air pollution devices, and an independent study from the South Coast Air Quality Management District found overall very strong agreement ($R^2 \sim 0.93 – 0.97$) between these and Federal Equivalent Methods (South Coast Air Quality Management District, 2021).

We consider data from 44 different sensors located in San Francisco, CA, collected at hourly resolution but post-processed at 4-hour resolution forming a continuous record from
2020/01/01 to 2020/05/24, for a total of 870 time points. More than 100 additional PurpleAir stations were not considered because they had only a fraction of the data available from the study period. San Francisco was chosen due to its high coverage compared to other major US cities, as well as the availability of previous studies for validation of the sensors, as detailed below. The period of study was chosen to avoid the wildfire season in California (Miller and Safford, 2012) with May 24, 2020 being three days before the first major seasonal wildfire (CAL FIRE, 2021), so that the local urban emissions are the main source of air pollution. Figure 1 shows the locations of the PurpleAir sites as well as the EPA’s air quality monitoring station, along with the median measurement of the PM$_{2.5}$ concentration during the period of study. Previous work has validated the measurements from the PurpleAir sensors (Kelly et al., 2021; Ardon-Dryer et al., 2020; Shen et al., 2021). Ardon-Dryer et al. (2020) showed that the sensors were generally well-correlated with the EPA observations, with the $R^2$ greater than 0.65 for 6 co-located sites in the United States, including San Francisco. In this work, the PurpleAir measurements were compared to the EPA data with a simple linear regression and achieved an $R^2$ of 0.52, thus in line with the previous study. A time series plot from the EPA site and the nearest PurpleAir site is shown in Figure S1A and a scatterplot with a fitted regression line can be seen in Figure S1B. In order to provide population exposure maps, population data from 194 census tracts in San Francisco was gathered from the 2014-2018 5-year American Community Survey by the US Census Bureau.

Figure 1: Bubble plot of the median PM$_{2.5}$ concentration across the PurpleAir sensor locations (dots) and EPA site (cross) from January 1$^{st}$, 2020 through May 24$^{th}$, 2020. The orange location is the site depicted in Figure 5.
3 Methodology

We introduce the ESN in Section 3.1 and the inference in Section 3.2. We then propose a recursive forecasting algorithm for long-range or long-lead forecasting in Section 3.3. Finally, we propose an adaptive method to quantify and adjust the forecast’s uncertainty as well as its multivariate (possibly spatial) dependence in Section 3.4.

3.1 Echo-State Networks

An ESN is formulated as follows (Lukosevicius, 2012):

\[
\begin{align*}
\text{output:} & \quad Y_t = B h_t + \epsilon_t, \\
\text{hidden state:} & \quad h_t = (1 - \alpha) h_{t-1} + \alpha \omega_t, \\
\omega_t & = f_h \left( \frac{h}{|h_t|} W h_{t-1} + W_{\text{in}} x_t \right), \\
\text{matrix distribution:} & \quad W_{i,j} = \gamma W_{i,j} g(\eta W) + (1 - \gamma W_{i,j}) \delta_0, \\
& \quad W_{\text{in}} = \gamma W_{\text{in}} g(\eta W_{\text{in}}) + (1 - \gamma W_{\text{in}}) \delta_0, \\
& \quad \gamma_{i,j} \sim \text{Bern}(\pi W), \quad \gamma_{i,j} \sim \text{Bern}(\pi W_{\text{in}}).
\end{align*}
\]

The output \( Y_t \) is a \( n_t \)-dimensional vector (corresponding to the number of locations in case of spatial data) for time \( t = 1, \ldots, T \) and is expressed as a linear function \( B h_t + \epsilon_t \) in (1a), where \( h_t \) is an \( n_h \)-dimensional state vector, \( \epsilon_t \) is an error term, and \( B \) is an unknown coefficient matrix which is estimated (a more general polynomial form is investigated in the supplementary material, and did not provide a significant improvement). The state vector is a convex combination of a past state \( h_{t-1} \) and a memory term \( \omega_t \), controlled by a ‘leaking rate’ parameter \( \alpha \) as specified in (1b). The memory term \( \omega_t \) is the result of a nonlinear function \( f_h \) with some pre-specified forms such as the hyperbolic tangent or rectified linear (Goodfellow et al., 2016), which combines the past state \( h_{t-1} \) and the \( n_x \)-dimensional vector consisting of present and past input \( x_t = (x_{t-\tau}, \ldots, x_{t-m\tau}) \), as shown in (1c). The ability of the model to retain long-range information is controlled by parameters \( \alpha \) and \( m \), since the input comprises of values at time points \( t - \tau \) through \( t - m\tau \) where \( \tau \) is commonly denoted as the forecast lead time (McDermott and Wikle, 2017). Both matrices \( W \) and \( W_{\text{in}} \) in (1c) are randomly generated assuming high sparsity. Indeed, the entries of both matrices are a linear combination of \( \delta_0 \), a Dirac function at zero, and a symmetric distribution \( g(\cdot) \) centered
about zero, see (1d) and (1e). The distribution of each entry is therefore a ‘spike-and-slab’, a widespread choice for prior in Bayesian variable selection (Ishwaran and Rao, 2005; Malsiner-Walli and Wagner, 2011). The probability of each entry being zero is controlled by independent Bernoulli random variables with probabilities $\pi_W$ and $\pi_{W^{in}}$ for $W$ and $W^{in}$, respectively. The matrices $W$ and $W^{in}$ are of dimension $n_h \times n_h$ and $n_h \times n_x$ respectively. The error $\epsilon_t$ is independent identically distributed mean zero multivariate normal with some covariance structure. In this work, instead of specifying it explicitly, we account for the dependence after the forecasts have been produced, as we will show in Section 3.4.

In order to be well defined, the proposed ESN must also have the echo-state property, i.e., with a sufficiently long time sequence the model must asymptotically lose its dependence on the initial conditions (Lukosevicius, 2012; Jaeger, 2007). In order for this property to hold, the spectral radius or largest eigenvalue of $W$, denoted by $\lambda_W$, must be less than one. This is ensured by the use of the scaling parameter $\nu$. The set of hyper-parameters for this model is then $\theta = (n_h, m, \nu, \lambda_r, \eta_W, \eta_{W^{in}}, \pi_W, \pi_{W^{in}}, \alpha)$.

### 3.2 Inference and Computational Issues

In order to perform inference on the hyper-parameter vector $\theta$, a validation approach is used, where a portion of the training data is held out (validation set) and forecasts are validated over it. The distribution $g(\cdot)$ used to populate the matrices $W$ and $W^{in}$ is typically generated from either a symmetric uniform or a normal distribution centered around zero (Lukosevicius, 2012). In this work we use standard normal distributions to populate these matrices, thus we remove hyper-parameters $\eta_W$ and $\eta_{W^{in}}$. Additionally, out of the entire hyper-parameter space, the hyper-parameters $\pi_W$, and $\pi_{W^{in}}$ did not have a significant impact on the performance of the model as is shown in a preliminary analysis in the supplementary material. Hence, the hyper-parameter space is reduced to $\theta^* = (n_h, m, \nu, \lambda_r, \alpha)$. Once $\theta^*$ is fixed, the matrix $B$ in equation (1a) is estimated, predictions are obtained for the validation set, the mean squared error (MSE) is calculated and the hyper-parameters are inferred by minimizing it.

A (conditional) regularized regression such as ridge regression with penalty $\lambda_r$ mitigates the large variance of the estimators. Equation (1a) can be denoted in matrix form as $Y =$
\[ \mathbf{HB} + \mathbf{E}, \text{ where} \]

\[
\mathbf{Y} = \begin{bmatrix}
\mathbf{Y}_1^\top \\
\vdots \\
\mathbf{Y}_T^\top
\end{bmatrix}, \quad
\mathbf{H} = \begin{bmatrix}
\mathbf{h}_1^\top \\
\vdots \\
\mathbf{h}_T^\top
\end{bmatrix}, \quad
\mathbf{E} = \begin{bmatrix}
\mathbf{e}_1^\top \\
\vdots \\
\mathbf{e}_T^\top
\end{bmatrix}.
\]

A ridge regression is applied to estimate \( \mathbf{B} \) such that 
\[
\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} (\| \mathbf{Y} - \mathbf{HB} \|_F^2 + \lambda_r \| \mathbf{B} \|_F^2),
\]
where \( \| \cdot \|_F \) is the Frobenius norm and \( \lambda_r \) is the penalty parameter. This provides the ridge estimator \( \hat{\mathbf{B}} \) in closed form as 
\[
\hat{\mathbf{B}} = (\mathbf{H}^\top \mathbf{H} + \lambda_r \mathbf{I})^{-1} \mathbf{H}^\top \mathbf{Y}.
\]

Additionally, while the sparsity of the weight matrices allows for fast and memory efficient sparse linear algebra operations, the scalability of the inference and hence the generation of the ensemble for high-dimensional \( \mathbf{Y}_t \) represents a challenge, especially in cases where the time scale is fine (e.g., \(<1 \text{ day}\)) and the elapsed time for forecasts needs to be at most a fraction of the sampling rate. Generating ensemble forecasts independently in parallel greatly improves the computational efficiency, as each simulation requires linear algebra operations with high-dimensional output vectors, representing spatial locations in a typical spatio-temporal application. In the supplementary material we assess the computational speedup as a function of the number of processing units (cores) used to generate the ensembles of forecasts.

### 3.3 Long-Lead Forecasting

In order to produce long-lead forecasts, we use a simple iterative approach. Let us assume we observe data for \( t = 1, \ldots, T \), we then produce an ensemble of forecasts for \( T + 1 \) and then its average is appended in the input \( \mathbf{x}_{T+1} \) to predict \( T + 2 \). The model is then re-trained using \( T + 1 \) time points and used to predict point \( T + 2 \) and then iteratively for the next time steps. This method is detailed in Algorithm S1 in the supplementary material. One-step ahead forecasts were not strongly influenced by the leaking rate \( \alpha \) [McDermott and Wikle 2017], but in our proposed approach we found that the forecasting ability of the ESN was highly sensitive to the choice of \( \alpha \) when performing long-lead forecasting. As it will be described in detail later from Figure 2, the choice of \( \alpha = 1 \) (emphasising short-range dependence) fails to capture long-lead patterns dictated by the nonlinearity of the underlying process.
3.4 Calibrating the Forecast

One might attempt to quantify uncertainty at a future time point by feeding the previous ensemble estimates forward, thus creating an ensemble of ensembles. However, for either long time series or high-dimensional data vectors, this approach would rapidly become computationally expensive and impractical. In this work, we propose a new method to quantify the forecast uncertainty which is feasible both in terms of time and memory. The method relies on the use of a fraction of the training set to 1) compare predictions and true values 2) calibrate the marginal uncertainty of the residuals and 3) estimate their correlation. Throughout this section we assume that data are observed for \( t = 1, \ldots, T + n_w n_f \) but the model is only trained until \( T \). We generate an ensemble of forecasts for time points \( T + 1 \) through \( T + n_w n_f \) across multiple ‘windows’ using the long-lead forecasting approach detailed in Section 3.3 which are then used for calibration. More specifically, ensembles of forecasts would be generated for the time points \( T + 1, \ldots, T + n_w n_f \) in intervals of size \( n_f \) (representing the forecast horizon), where \( n_w \) represents the number of ‘windows’. We denote the forecast for time point \( T + 1 \) given \( \{ Y_T, Y_{T-1}, \ldots \} \) as \( \hat{Y}_{T+1}^T = \mathbb{E}(Y_{T+1} | Y_T, Y_{T-1}, \ldots) \), where the expectation is intended with respect to model (\[1\]). The forecasts are then grouped according to each vector element \( \ell = 1, \ldots, n_\ell \) as follows:

\[
\begin{bmatrix}
\hat{Y}_{T+1}(\ell) \\
\hat{Y}_{T+n_f}(\ell) \\
\vdots \\
\hat{Y}_{T+(n_w-1)n_f}(\ell)
\end{bmatrix}, \quad \begin{bmatrix}
\hat{Y}_{T+2}(\ell) \\
\hat{Y}_{T+n_f+2}(\ell) \\
\vdots \\
\hat{Y}_{T+(n_w-1)n_f+2}(\ell)
\end{bmatrix}, \quad \ldots, \quad \begin{bmatrix}
\hat{Y}_{T+n_f}(\ell) \\
\hat{Y}_{T+2n_f}(\ell) \\
\vdots \\
\hat{Y}_{T+(n_w-1)n_f}(\ell)
\end{bmatrix}.
\]

(2)

Since these groupings are created for each vector element independently, the length of each vector \( \{ f^{(\ell)}_j ; j = 1, \ldots, n_f \} \) is \( n_w \). As an example, if \( n_f = 20 \) and \( n_w = 10 \) then the one- and two-step ahead forecast for vector element \( \ell \) would be grouped as \( f^{(\ell)}_1 = ( \hat{Y}_{T+1}(\ell), \hat{Y}_{T+20}(\ell), \ldots, \hat{Y}_{T+180}(\ell) ) \) and \( f^{(\ell)}_2 = ( \hat{Y}_{T+2}(\ell), \hat{Y}_{T+22}(\ell), \ldots, \hat{Y}_{T+182}(\ell) ) \). The residuals associated with each of these groupings are calculated by subtracting the forecast from each associated true value, e.g., for the \( j \)-step ahead forecasts and vector element \( \ell \) we have:

\[
R^{(\ell)}_j = Y^{(\ell)}_{j-\text{step}} - f^{(\ell)}_j, \quad j = 1, \ldots, n_f,
\]

(3)

where \( Y^{(\ell)}_{j-\text{step}} = \{ Y_{T+(w-1)n_f+j}(\ell) ; w = 1, \ldots, n_w \} \) and \( R^{(\ell)}_j = \{ R^{(\ell)}_j(w) ; w = 1, \ldots, n_w \} \). The standard deviation associated with the residuals of each collection of future time points is
then calculated as:

\[
\hat{\sigma}_j^{(\ell)} = \sqrt{\frac{\sum_{w=1}^{n_w} \left( R_j^{(\ell)}(w) - \overline{R}_j^{(\ell)} \right)^2}{n_w - 1}}, \quad j = 1, \ldots, n_f,
\] (4)

where \( \overline{R}_j^{(\ell)} = \frac{\sum_{w=1}^{n_w} R_j^{(\ell)}(w)}{n_w} \). A reasonable assumption is that uncertainty about the forecasts for each vector element expands in time, i.e., \( \hat{\sigma}_j^{(\ell)} \leq \hat{\sigma}_{j'}^{(\ell)}, j \geq j' \) for every \( \ell \). The vector \( \left( \hat{\sigma}_1^{(\ell)}, \ldots, \hat{\sigma}_{n_f}^{(\ell)} \right) \) is then smoothed with a monotonic increasing spline with resulting fitted values indicated as \( \left( \tilde{\sigma}_1^{(\ell)}, \ldots, \tilde{\sigma}_{n_f}^{(\ell)} \right) \). In this work, we use a monotonic cubic spline interpolation (Wolberg and Alfy, 1999; de Boor, 1978) with the additional constraint \( \tilde{\sigma}_j^{(\ell)} \geq 0 \), where \( \tilde{\sigma}_j^{(\ell)} \) represents the slope of the spline at future point \( j \). This method would also allow for calibration of the forecasts (Gneiting et al., 2007), as the ensemble is adjusted to quantify the correct uncertainty, as will be shown by the probability integral transform (PIT) of the standardized residuals:

\[
\tilde{R}_j^{(\ell)} = \left\{ \frac{R_j^{(\ell)}}{\tilde{\sigma}_j^{(\ell)}}, j = 1, \ldots, n_f \right\}.
\] (5)

Since the residuals are assumed to follow a multivariate Gaussian distribution by the model in equation (1) and the forecasts are calibrated, the PIT will more closely resemble a standard uniform distribution after the proposed adjustment. Additionally, the estimated marginal standard deviations from equation (4) can then be used to generate the confidence intervals for each vector element, as will be shown in Sections 4 and 5. The full approach is detailed in Algorithm 1.

Finally, the vector of standardized residuals \( \tilde{R}_{j,w} = \left( \tilde{R}_j^{(1)}(w), \ldots, \tilde{R}_j^{(n_\ell)}(w) \right) \) is assumed to have a multivariate \( n_\ell \)-dimensional normal distribution across vector elements represented as:

\[
\tilde{R}_{j,w} \overset{i.i.d.}{\sim} N_{n_\ell}(0, C),
\] (6)

i.e., the \( \tilde{R}_{j,w} \) are independent and identically distributed across \( j = 1, \ldots, n_\ell \) and \( w = 1, \ldots, n_w \) with correlation matrix \( C \). In this work, we use a penalized non-parametric method to generate sparse estimates of \( C \) in the simulation study in Section 4 and a parametric model for the spatial application in Section 5.
Data: \( Y_t : t = 1, \ldots, T + n_w n_f \)

Result: Vector of standard deviations \( \tilde{\sigma} \)

Initialize: Generate forecasts \( \hat{Y}^{(\ell)}_{j-step} = \{ \hat{Y}_{T+(w-1)n_f+1}^{T+(w-1)n_f+j}(\ell); w = 1, \ldots, n_w \} \) for \( j = 1, \ldots, n_f \) and \( \ell = 1, \ldots, n_\ell \)

for \( \ell = 1, \ldots, n_\ell \) do

for \( j = 1, \ldots, n_f \) do

Compute the \( j \)-step ahead forecast \( \{ \hat{Y}_{T+(w-1)n_f+1}^{T+(w-1)n_f+j}(\ell); w = 1, \ldots, n_w \} \) and group them into \( f^{(\ell)}_{j} \), as detailed in equation (2);

Calculate the residuals for the \( j \)-step ahead forecast \( R^{(\ell)}_{j} \) from equation (3);

Compute the \( j \)-step ahead residual standard deviation \( \hat{\sigma}^{(\ell)}_{j} \) from equation (4);

end

end

Fit a monotonic increasing spline to \( \{ \hat{\sigma}^{(\ell)}_{j}; \ j = 1, \ldots, n_f \} \) and denote the vector of fitted values \( \tilde{\sigma} \).

end

Algorithm 1: Marginal forecast calibration.

4 Simulation Study

In this work, we aim to understand the predictive capabilities of the ESN on time series with non-linear dynamics against more traditional time series models. We do this in a controlled setting from a model with a well known quasi-periodic behavior. In Section 4.1 we simulate data from a Lorenz-96 model [Lorenz 1996], a model first proposed to study forecasting ability on spatial chaotic systems [Karimi and Paul 2010]. We compare the ESN with other standard time series methods in terms of their predictive ability in Section 4.2. Finally, we calibrate the marginal forecasts, estimate the multivariate dependence and compare the results with both an uncalibrated and independent approach in Section 4.3.

4.1 Simulated Data

We simulate data from the Lorenz [1996] differential equations for \( n_\ell = 40 \) variables, i.e.,

\[
\frac{dY_t(\ell)}{dt} = \{Y_t(\ell + 1) - Y_t(\ell - 2)\} Y_t(\ell - 1) - Y_t(\ell) + F_t, \quad \ell = 1, \ldots, 40.
\] (7)

The forcing is constant in time and set to \( F_t = 4.5 \), which implies a median absolute correlation of 0.42. In the supplementary material, we show an analysis with different values of \( F_t \),
implying a varying degree of dependence among vector elements. Additionally, we simulated ten realizations of this data from ten different random initial conditions to be used for this analysis. A realization of 200 time points from a selected vector element for one out of the ten realizations is shown in Figure S5. The ESN was initially trained using the first \( T = 980 \) time points and was tested using the subsequent \( h = 20 \) time points. We use the following hyper-parameters for the ESN: \( n_h = 60, \lambda_r = 0.001, \nu = 0.55, m = 4, \pi_W = \pi_{W^m} = 0.1 \) and \( g(\cdot) \) set to a standard normal distribution.

### 4.2 Point Forecasts

Forecasts for all ten realizations were produced using \( \alpha \) values across two different grids: \( \{10^{-2} \times s_\alpha : s_\alpha = 1, \ldots, 100\} \) and \( \{10^{-4} \times s_\alpha : s_\alpha = 1, \ldots, 100\} \), in order to allow extremely small \( \alpha \) values in the assessment of the predictability of the model. The value which achieved an overall minimum MSE across all vector elements and realizations was \( \hat{\alpha}_{\text{opt}} = 0.0023 \), far from the optimal choice \( \alpha = 1 \) for short-range forecasting, as shown in Figure 2A. The overall improvement of the forecasts when using \( \hat{\alpha}_{\text{opt}} \) versus \( \alpha = 1 \) can be assessed when comparing panels B and C from Figure 2, where the long-range forecasting at a selected vector element from a single realization is shown. Indeed, the choice of \( \hat{\alpha}_{\text{opt}} \) allows the ESN to better retain the long-range information in data, thus producing more accurate forecasts than \( \alpha = 1 \), which only relies on the parameter \( m \) (i.e., the horizon of the input vector) in capturing the information. Additionally, in panels B and C of Figure 2 the 95% prediction intervals were generated using the forecast ensemble standard deviations without any adjustment. These intervals fail to properly account for the variability in the data, with a median coverage of 75.5% across all of the vector elements and realizations. The uncertainty surrounding these forecasts can be properly quantified using the method presented in Section 3.4, as will be shown in Section 4.3.

In order to compare the ESN with traditional time series modelling approaches, we also fit an auto-regressive fractionally integrated moving average model (ARFIMA; Granger and Joyeux (1980); Hosking (1981)). ARFIMA(\( p, d, q \)) can be written in operator form as

\[
\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) (1 - B)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) \epsilon_t, \tag{8}
\]

where \( B \) is the backshift operator: \( BX_t = X_{t-1} \) and \( \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2) \), i.e., independent and identically distributed zero mean Gaussian random variables. ARFIMA models are a generalized version of the popular auto-regressive integrated moving average (ARIMA; Brockwell and Davis (2016)) model with non-integer difference parameter \( d \).
We also fit a simplified state-space model ([Durbin and Koopman, 2012]). The state-space model assumes the ‘state variables’ at time $t$ are derived from the state at time $t - 1$ and that the output variables are dependent upon the value of the state variables at time $t$. In this work, we treat the ESN as a special case of the state-space model by assuming that $\alpha = 1$ in (1b) and $f_h(\cdot)$ is the identity function in (1c). The ESN in equation (1) would then take the form:

$$Y_t = B h_t + \epsilon_t,$$

$$h_t = \frac{\nu}{|\lambda_W|} W h_{t-1} + W^\text{in} x_t.$$  

The forecasts from the ARFIMA and the state-space model are shown in Figure 3 for the same vector element as in Figure 2 and both models clearly produce inadequate forecasts compared to the ESN. The sub-optimal performance can be attributed to the underlying nonlinear dynamics of the [Lorenz (1996)] model. ARIMA models are designed to only capture linear dependence, and ARFIMA models are not sufficiently flexible to capture a highly nonlinear behavior such as that of [Lorenz (1996)]. We quantify the forecasting performance of the different methods using the MSE and the continuous ranked probability score (CPRS;
Gneiting et al. (2007), a proper scoring rule for probabilistic forecasts which would be equal to zero with a perfect forecast. The results, shown in Table 1, highlight a median MSE across vector elements and realizations of 2.92 with interquartile range (IQR) of 2.16 for the ARFIMA model, whereas the state-space model returns an MSE of 4.58 (2.75). The ESN drastically outperforms these models in the case of an optimal leaking rate \( \hat{\alpha}_{\text{opt}} \), with MSE of 1.26 (0.90), while a short-range choice of \( \alpha = 1 \) would lead to the worst MSE of 6.91 (4.03). The same relative ordering of forecasting performance was found for the CRPS, with the ESN(\( \hat{\alpha}_{\text{opt}} \)) model showing the best performance.

![Figure 3: Comparison of different methods for the same vector element from Figure 2 with long-range forecasts for \( h = 20 \) time points](image)

| Forecasting Method | MSE     | CRPS    |
|--------------------|---------|---------|
| ESN (\( \hat{\alpha}_{\text{opt}} \)) | 1.26 (0.90) | 0.72 (0.20) |
| ESN (\( \alpha = 1 \)) | 6.91 (4.03) | 1.53 (0.46) |
| State-Space        | 4.58 (2.75) | 1.23 (0.37) |
| ARFIMA             | 2.92 (2.16) | 0.99 (0.35) |

Table 1: Comparison of the forecasting methods in terms of MSE and CRPS across all vector elements from each of the 10 realizations of simulated Lorenz (1996) data in equation (7). A median across all vector elements and realizations is shown, and in parenthesis the IQR is reported.
4.3 Prediction Uncertainty

The calibration method discussed in Section 3.4 was implemented with the ESN for \( n_w = 20, n_f = 20 \) for all 10 realizations from the Lorenz 96 equation (7). The first row of Table 2 shows that, when using the uncalibrated ensemble forecasts, the marginal uncertainty is incorrectly quantified and leads to a considerably smaller coverage for the nominal 95%, 80%, and 60% prediction intervals (PIs), with an average under-coverage of 21.5% across the three confidence levels. When using the adjusted standard deviation for the calibration approach, however, the marginal forecast uncertainty is properly quantified and does capture the appropriate amount of data within each of the specified PIs. Indeed, the calibrated 95% prediction interval on row 2 captures a median of 94.7% of the data across all vector elements with an IQR of 2.0%, a figure comparable to the 0.9% average under-coverage across all three intervals. Figure 4A shows the calibrated 95% PIs for a sample vector element where the increase in the uncertainty through time is readily apparent, due to the use of the monotonic spline on the standard deviations. Figure 4B shows instead how the calibrated PIT is closer to uniform than the uncalibrated PIT thereby underscoring an improved quantification of the uncertainty, as expected from the discussion in Section 3.4.

| Coverage   | Method     | 95%      | 80%      | 60%      |
|------------|------------|----------|----------|----------|
| Marginal   | Uncalibrated | 75.5 (3.3) | 56.0 (3.6) | 38.9 (3.1) |
|            | Calibrated  | 94.7 (2.0) | 79.0 (3.3) | 58.7 (3.7) |
| Difference | Independent | 96.8 (4.4) | 87.4 (10.7) | 71.1 (12.8) |
|            | Dependent   | 94.8 (1.2) | 81.5 (2.5) | 62.6 (2.8) |
|            | Dependent & Sparse | 96.4 (2.5) | 86.1 (7.0) | 68.5 (8.3) |
| Mean       | Independent | 99.8 (0.3) | 95.5 (0.7) | 81.4 (2.4) |
|            | Dependent   | 91.0 (1.8) | 73.5 (2.4) | 53.8 (1.9) |
|            | Dependent & Sparse | 98.6 (0.4) | 89.2 (1.6) | 71.4 (2.6) |

Table 2: Coverage from: marginal prediction intervals (rows 1-2), the difference between neighboring elements (rows 3-5), and the grand mean of forecasts (rows 6-8) across each of the 10 realizations of simulated Lorenz (1996) data in equation (7). The median is shown and in parenthesis the IQR of the coverage is also reported.

In order to estimate the correlation matrix \( \mathbf{C} \) of the standardized residuals in equation (6), we propose a penalized non-parametric sparse method. We use the majorize-minimize algorithm proposed by Bien and Tibshirani (2011) to estimate a sparse correlation matrix as determined by some pre-specified penalty parameter \( \lambda_s \). The minimization problem among all symmetric non-negative matrices \( \mathbf{C} \geq 0 \) can be formulated as follows:

\[
\hat{\mathbf{C}}_{\text{sparse}}(\lambda_s) = \arg \min_{\mathbf{C} \geq 0} \left[ \log \{ \det (\mathbf{C}) \} + \text{tr} (\mathbf{C}^{-1} \hat{\mathbf{C}}) + \lambda_s ||\mathbf{P} \ast \mathbf{C}||_1 \right],
\]
where $\hat{C}$ is the sample correlation matrix of the standardized residuals of the adjusted forecasts in equation (5), $P$ is a matrix including zeros on the diagonal and ones elsewhere, and $*$ indicates element-wise multiplication which ensures that only the off-diagonal elements are penalized. This method is especially suitable in cases where the vector elements have no spatial information, as is the case here with the Lorenz (1996) data. The sparsity can be controlled by evaluating the trade-off between the proportion of zeros in the matrix and loss of information in the form of changing variation of the standardized residuals. Figure 4C shows the change of the variance of the average standardized residuals, $\sigma^2_{\hat{R}}(\lambda_s) = \frac{1}{n_f} \mathbf{1}^T \hat{C}_{\text{sparse}}(\lambda_s) \mathbf{1}$ ($\mathbf{1}$ is an $n_f$-dimensional column vector of ones), as well as the proportion of non-zero elements in $\hat{C}_{\text{sparse}}(\lambda_s)$ with respect to the penalty parameter $\lambda_s$. Here, the red line represents the baseline variance assuming $\lambda_s = 0$, the dashed black line represents the percentage change of $\sigma^2_{\hat{R}}(\lambda_s)$ against $\sigma^2_{\hat{R}}(0)$ as a function of $\lambda_s$, and the dotted blue line represents the proportion of non-zero elements. Heatmaps of $\hat{C}$ and $\hat{C}_{\text{sparse}}(\lambda_s = 0.1)$ (represented as the solid green line in Figure 4C) can be seen in panels A and B of Figure S6.

$\hat{C}_{\text{sparse}}(\lambda_s = 0.1)$ was then used to calculate the nominal coverage of PIs for $R_j^{(\ell)} - R_j^{(\ell+1)}$, $j = 1, \ldots, n_f$, the difference in residuals between neighboring vector elements. More specifically, their variance was calculated assuming both independence and dependence (from $\hat{C}$) among the elements. For this coverage calculation, only vector elements with mild absolute correlation between 0.4 and 0.6 were considered and this constraint was applied across all 10 realizations of the simulated data. It is shown in rows 3-5 of Table 2 that the nominal coverage improves in the dependent case where the adjusted uncertainty is used to determine the dependence among the vector elements. Even when considering $\hat{C}_{\text{sparse}}(\lambda_s = 0.1)$, the coverage is still improved versus the independent case. In these dependent and sparse dependent cases the average over-coverage across the specified intervals is 1.3% and 5.3%, respectively, whereas the average over-coverage in the independent case is 6.8%.

The median coverage of the grand mean, i.e., the average across all vector elements, is shown in rows 6-8 in Table 2. When the vector elements are treated independently for each realization, there is an average over-coverage of 13.9% across the chosen intervals. The nominal coverage improves when $\hat{C}$ and $\hat{C}_{\text{sparse}}(\lambda_s = 0.1)$ are used to inform dependence with an average under-coverage of 5.6% and an average over-coverage of 8.1% across the intervals, respectively. By sparsifying the empirical correlation matrix, the overall variance is reduced, therefore explaining the improved coverage versus the independent case.
Figure 4: Uncertainty quantification for the same vector element from Figures 2 and 3. (A) Forecasts with adjusted PIs generated using standard deviations from equation (4), (B) original and adjusted PIT for the standardized residuals from the forecasts and (C) percentage ratio between the new and baseline variance of the mean standardized residuals and proportion of non-zero elements in correlation matrix as the penalty parameter $\lambda_s$ increases in equation (10). The green line represents the $\lambda_s = 0.1$ value chosen to generate the sparse correlation matrix heatmap in Figure S6B.

5 Application

We now use the proposed approach to produce long-range forecasts with calibrated uncertainty quantification for the San Francisco air pollution data presented in Section 2. The first $T = 840$ points, consisting of measurements from 2020/01/01 to 2020/05/19, were used as a training set, while the remaining $h = 30$ points, representing 2020/05/20 through 2020/05/24, were used as the testing set. As in the simulation study, the ESN is insensitive to the choice of hyper-parameters ($\pi_w$, $\pi_w^{in}$) (see Figure S4), so they were both fixed at 0.1. Additionally, $g(\cdot)$ in equations (1d) and (1e) is here a standard normal distribution, therefore there are no hyper-parameters ($\eta_w$, $\eta_w^{in}$). Thus, the only hyper-parameters selected via validation for the ESN are $(n_h, m, \nu, \lambda_r, \alpha)$. The grids used for selection are as follows: $n_h = \{30 + 30 \times s_h : s_h = 0, \ldots, 5\}$, $m = \{2, \ldots, 6\}$, $\nu = \{0.05 \times s_\nu : s_\nu = 1, \ldots, 20\}$, $\lambda_r = \{0.001, 0.005, 0.01\}$ and the grids for $\alpha$ were the same as those used for the Lorenz (1996) simulated data in Section 4.2. All of the hyper-parameter grids were optimized and evaluated using MSE across all time points for all locations using a forecast ensemble of 300 elements. The hyper-parameters which produced the minimum validation MSE were:
\( n_h = 180, m = 5, \nu = 0.25, \lambda_r = 0.01, \text{ and } \alpha_{\text{opt}} = 3 \times 10^{-4}. \) A main assumption of the ESN model is that \( Y_t \) follows a normal distribution as per equation (1a). The logarithm of the data was used in order to achieve Gaussianity, thus it is assumed that the original data follows a log-normal distribution. We compare the forecasting skill of the ESN against a ARFIMA and the state-space model in Section 5.1 quantify the forecasting uncertainty and model the spatial dependence in Section 5.2. Finally, in Section 5.3, we interpolate forecasts, estimate the population exposure to excess levels of PM\(_{2.5}\) and compare the PurpleAir results with the EPA station. The model diagnostics are detailed in the supplementary material.

### 5.1 Forecasting Air Pollution

The forecasts from ARFIMA and state-space models for the same location from Figure 5 can be seen in Figure S7. Similarly to the simulation study, both models fail to produce accurate forecasts, hence underscoring their inadequacy in capturing the nonlinear dynamics of urban air pollution. Further evidence of the lack of fit can be assessed by computing the MSE and CRPS as in Section 4 with results in Table 3. The ARFIMA model returns a median MSE value of 0.28 (IQR across sites 0.19), the state-space model of 0.46 (0.14), whereas for the ESN of 0.18 (0.06), which exhibits the improvement in forecasting of the ESN against the other models. The ESN(\( \hat{\alpha}_{\text{opt}} \)) also outperforms the forecasts from an ESN using \( \alpha = 1 \), where this model returns a MSE of 3.13 (1.16). An identical ranking, with ESN(\( \hat{\alpha}_{\text{opt}} \)) outperforming the other models, can be seen with CRPS.

| Forecasting Method | MSE    | CRPS   |
|--------------------|--------|--------|
| ESN (\( \hat{\alpha}_{\text{opt}} \)) | 0.18 (0.06) | 0.24 (0.04) |
| ESN (\( \alpha = 1 \))      | 3.13 (1.16) | 1.29 (0.25) |
| State-Space          | 0.46 (0.14) | 0.42 (0.07) |
| ARFIMA               | 0.28 (0.19) | 0.31 (0.10) |

Table 3: Comparison of the forecasting methods in terms of median MSE and CRPS across all locations from the San Francisco air pollution data. In parenthesis the IQR across locations is reported.

### 5.2 Assessing the Forecast Uncertainty

Row 1 of Table 4 shows that, without any calibration of the forecasts, the original marginal 95% PIs are only able to capture a median of 16.7% of the data across all 44 locations with an IQR 2.5%. After calibration, the median 95% PIs coverage across all locations is 94.0% (0.9%), as shown in row 2, an improvement consistent with the results presented in the simulation study in Section 4. Additionally, Figure 5A shows how the adjustment...
method from Section 3.4 calibrates the forecasts, as the PIT is closer to resembling a standard uniform distribution that the original uncalibrated forecast. Figure 5B shows the test data forecasts and adjusted 95% prediction intervals for a sample location indicated in Figure 1. From this plot, it can be seen that the post-adjustment PIs are able to capture the uncertainty in the data and become wider as we forecast further into the future, as we would expect.

Table 4: Comparison of log PM$_{2.5}$ concentration forecast methods in terms of discrepancy between nominal and actual coverage. Rows 1 and 2 show the marginal coverage for uncalibrated and calibrated PIs. Rows 3 and 4 focus on the difference between forecasts in neighboring locations assuming independence and dependence (similarly to the simulation study in Section 4 only locations with absolute correlation between 0.4 and 0.6 were considered). Rows 5 and 6 compare the mean of the forecasts assuming independence and dependence. The median coverage is reported and in parenthesis the IQR is shown.

In order to capture spatial dependence of the forecasts, we apply a convolution-based non-stationary model to the standardized residuals calculated from equation (5). This model allows the spatial structure to vary in space (Paciorek and Schervish, 2006; Risser and...
More specifically, the correlation is generated as a result of the convolution at some fixed knots over the spatial domain. The correlation structure is formulated as in Risser and Calder (2017):

$$
\hat{C}_{\text{spatial}}(s, s') = \frac{\left| \frac{V(s)^{1/4}V(s')^{1/4}}{\sqrt{V(s)+V(s')}} \right|^{1/2}}{k \left( \sqrt{Q(s, s')} \right)},
$$

$$
Q(s, s') = (s - s')^\top \left( \frac{V(s) + V(s')}{2} \right)^{-1} (s - s'),
$$

where $k(\cdot)$ is a correlation function, $V(\cdot)$ is a matrix representing the local anisotropy and $Q(s, s')$ is the Mahalanobis distance. In this work, $k(\cdot)$ is specified as the exponential function with a spatially-varying range and a fixed, unknown nugget. In order to specify $V(s)$, we assume a mixture at some knots for $Z$ fixed locations $\{b_z; z = 1, \ldots, Z\}$. More specifically, we define

$$
V(s) = \sum_{z=1}^{Z} w_z(s)V_z,
$$

where the weights, $w_z(s)$, are normalized. In this work $\lambda_z$ was fixed to be half the minimum distance between the knots. A similar approach is used to estimate the range parameter for the exponential correlation function. As discussed in Risser and Calder (2017), inference in performed through local likelihood.

In order to further improve our estimation of the spatial correlation structure, a new estimate is generated as a convex combination of the non-stationary correlation estimate, $\hat{C}_{\text{spatial}}$, and the empirical correlation, $\hat{C}$, a process known as generalized shrinkage (Friedman et al., 2007) which has been applied in previous works as a means to improve the estimated correlation structure (Castruccio et al., 2018). More specifically, the new correlation $\hat{C}^*(\delta_c)$ is formulated as:

$$
\hat{C}^*(\delta_c) = (1 - \delta_c)\hat{C}_{\text{spatial}} + \delta_c \hat{C},
$$

where $\delta_c \in (0, 1)$ is a parameter to be estimated. In this work, $\delta_c$ was chosen such that the coverage of the grand mean of the data, in the log-scale, is as close to the specified PIs as possible. This correlation matrix $\hat{C}^*(\delta_c)$ therefore specifies the spatial dependence of the standardized residuals in equation (6) and is used to calculate the PIs for the difference in
neighboring locations and mean forecasts in Table 4 rows 3-6. Accounting for the spatial dependence among locations improves the coverage of the PIs for both the forecast difference of neighboring locations and forecast grand mean. Indeed, assuming independence among the locations results in an average over-coverage for the difference of 11.5% versus an average over-coverage of 0.3% when assuming dependence. When considering the coverage of the mean forecast, the spatial model dramatically improves the coverage. More specifically, the under-coverage of the data, when assuming independence, is 60.5% versus 3.8% when assuming dependence. Even though the PIs are reported in the logarithmic scale, they are still interpretable in the original scale. In fact, the marginal coverage is unchanged since the logarithmic transformation is monotonic. The difference between neighboring elements and the mean in the logarithmic scale corresponds in the PM$_{2.5}$ scale to the ratio and the geometric mean, respectively.

5.3 Interpolation, Exposure and Comparison with EPA Data

After the forecasts at the locations in Figure 1 were calibrated, maps of San Francisco were produced via interpolation using the spatial model in the previous section. The interpolated forecasts were then partitioned into 194 census tracts or districts, as determined by the United States Census Bureau, and the PM$_{2.5}$ value for each district was computed as the average of the interpolated forecasts within each district’s borders. These forecasts were generated for all $h = 30$ future 4-hour time points, which represents 2020/05/20 through 2020/05/24. Panel A in Figure 6 shows the total number of citizens exposed to ‘moderate’ levels of air pollution (PM$_{2.5} > 12.1$µg·m$^{-3}$, Environmental Protection Agency (2021)) across all $h = 30$ forecast points with the corresponding PIs. We consider only susceptible citizens or those between the ages of 30 and 70 (World Health Organization [2014], see supplementary material for the population data preprocessing). As seen in panel A, the number of individuals exposed increases before the early morning of 2020/05/24. Panel B in Figure 6 shows the population districts at forecast point $T + 27$ (the time point with the maximum exposure) with PM$_{2.5}$ levels greater than 12.1 µg·m$^{-3}$. Here, the southwestern most districts of the city are exposed to ‘moderate’ levels of air pollution, and this result coincides with the results presented in panel C where the interpolated forecasts are shown. Additionally, Panel D in Figure 6 shows the standard deviation of the interpolated forecasts.

For comparison with EPA data, the ESN was implemented with the EPA monitoring station, with training and testing data on the same date range as the PurpleAir data. Since EPA data are collected at the daily level, the ESN was trained using only $T = 140$ points and tested on the next $h = 5$ points. The MSE of the EPA forecasts was 0.29, worse than the 0.18 with the PurpleAir data (row 1 of Table 3). The inability of the ESN to produce as
Figure 6: (A) Time series of susceptible population exposed to levels of PM$_{2.5}$ concentration $> 12.1 \mu g \cdot m^{-3}$ across all San Francisco with PIs. (B) Population districts exposed to levels of PM$_{2.5}$ concentration $> 12.1 \mu g \cdot m^{-3}$ at forecast time $T + 27$, or 2020/05/24 at 8:00:00. (C) Mean interpolated forecasts for all of San Francisco at time point $T + 27$. (D) Standard deviation of interpolated forecasts for all of San Francisco at time point $T + 27$.

Accurate forecasts with EPA data can be attributed to a smaller training set, but also to the lack of any spatial information. Indeed, for PurpleAir multiple locations are used to inform the forecasts, thereby borrowing spatial information. Additionally, none of the forecasts with the EPA data for all $h = 5$ future points exceeded the threshold of $12.1 \mu g \cdot m^{-3}$. Since this is the only air quality monitoring station within San Francisco, a map would trivially show the same value everywhere, thus losing information on local areas of high pollution as the ones presented in panels C and D of Figure 6.

6 Conclusion

In this work we have introduced a new approach for forecasting air pollution in an urban environment and quantifying the associated uncertainty with a recurrent neural net-
work. The large parametric space of NN-based models is not practical in applications with high-frequency data such as sub-daily air pollution. Therefore, we rely on an Echo-State Network (ESN), a faster alternative to standard dynamical NN models with sparse stochastic networks informed by a spike-and-slab prior and controlled by a low dimensional parameter space, thereby dramatically reducing the computational burden. The stochastic nature of the ESN allows for real time forecasts, but simultaneously introduces additional uncertainty that needs to be quantified. In this work, we propose a fast approach to calibrate the marginal uncertainty propagated by the stochastic network. Additionally, in order to produce multivariate (possibly spatial) forecasts, instead of embedding the dependence within the ESN in either the network or in the error structure (a task which would add computational complexity and hamper its practicality), we propose a post hoc adjustment by modeling the dependence of the vector of marginally calibrated forecasts, either with penalized sparse correlation matrix estimation in a multivariate setting, or with a non-stationary spatial model in the case of urban air pollution. The proposed approach shows an appreciable improvement in the predictive abilities compared to standard forecasting strategies such as ARIMA and state-space models, and shows better-calibrated marginal and joint forecasts versus uncorrected or independent approaches.

Another important contribution of this work is as a demonstration of the value of citizen science networks. This very recent source of data provides spatially-resolved information at urban scale, overcoming a major limitation of the sparse EPA network, satellite data and numerical simulations. This work represents one of the early efforts in assessing the added value of this new data in early warning systems on urban air quality, and some of our current work is expanding the methods presented here in assessing mortality from extreme events, such as wildfires (Shen et al., 2021).

While representing an opportunity, air pollution from citizen science data needs to be used with the awareness of potentially strong sampling bias due to income inequalities within the urban areas. While in less populous, relatively homogeneous cities, such as San Francisco, this may not be a substantial issue, considerably larger and more diverse cities would inevitably suffer from undersampling in areas of low income, where acquisition of such monitoring devices could not be as widespread. The sampling bias in citizen science projects due to underrepresentation of certain socio-economic groups is an emerging issue (Sorensen et al., 2019), and may be at least partially mitigated by federally funded, coordinated urban sensing projects such as the recent Array of Things (Catlett et al., 2017).
Araujo, L. N., J. T. Belotti, T. A. Alves, Y. de Souza Tadano, F. Trojan, and H. Siqueira (2020). Analysis of Regularized Echo State Networks on the Impact of Air Pollutants on Human Health, pp. 357–364. Springer International Publishing.

Ardon-Dryer, K., Y. Dryer, J. N. Williams, and N. Moghimi (2020). Measurements of PM2.5 with PurpleAir under atmospheric conditions. Atmospheric Measurement Techniques 13(10), 5441–5458.

Bien, J. and R. J. Tibshirani (2011). Sparse estimation of a covariance matrix. Biometrika 98(4), 807–820.

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra (2015, 07–09 Jul). Weight uncertainty in neural network. In F. Bach and D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, Volume 37 of Proceedings of Machine Learning Research, Lille, France, pp. 1613–1622. PMLR.

Briggs, D. J., C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone (2000). A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Science of The Total Environment 253(1), 151–167.

Brockwell, P. J. and R. A. Davis (2016). Introduction to Time Series and Forecasting. New York: Springer.

CAL FIRE (2021). List of wildfires on May 27th. https://www.fire.ca.gov/incidents/2020/5/27/range-fire/ last accessed 2021/04/16.

Carlsten, C., S. Salvi, G. W. Wong, and K. F. Chung (2020). Personal strategies to minimise effects of air pollution on respiratory health: advice for providers, patients and the public. European Respiratory Journal 55(6), 1902056.

Castruccio, S., H. Ombao, and M. G. Genton (2018). A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fmri data. Biometrics 74(3), 823–833.

Catlett, C. E., P. H. Beckman, R. Sankaran, and K. K. Galvin (2017). Array of things: a scientific research instrument in the public way: platform design and early lessons learned. In SCOPE ’17: Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering, pp. 26–33.
de Boor, C. (1978). *A Practical Guide to Splines*, Volume 27. Springer.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Environmental Protection Agency (2021). NAAQS table. [https://www.epa.gov/criteria-air-pollutants/naaqs-table](https://www.epa.gov/criteria-air-pollutants/naaqs-table) last accessed 2021/04/18.

Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432–441.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2), 243–268.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. [http://www.deeplearningbook.org](http://www.deeplearningbook.org).

Goodkind, A. L., C. W. Tessum, J. S. Coggins, J. D. Hill, and J. D. Marshall (2019). Fine-scale damage estimates of particulate matter air pollution reveal opportunities for location-specific mitigation of emissions. *Proceedings of the National Academy of Sciences* 116(18), 8775–8780.

Granger, C. W. J. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1(1), 15–29.

Graves, A. (2011). Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 24. Curran Associates, Inc.

Grell, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. C. Skamarock, and B. Eder (2005). Fully coupled “online” chemistry within the WRF model. *Atmospheric Environment* 39(37), 6957–6975.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika* 68(1), 165–176.

Huang, H., S. Castruccio, and M. G. Genton (2021). Forecasting high-frequency spatio-temporal wind power with dimensionally reduced echo state networks. *arXiv:2102.01141*. 

25
Ishwaran, H. and J. Rao (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics 33*(2), 730–773.

Jaeger, H. (2001). The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148.

Jaeger, H. (2007). Echo state network. *Scholarpedia 2*(9), 2330.

Karimi, A. and M. R. Paul (2010). Extensive chaos in the lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science 20*(4), 043105.

Kelly, K. E., W. W. Xing, T. Sayahi, L. Mitchell, T. Becnel, P.-E. Gaillardon, M. Meyer, and R. T. Whitaker (2021). Community-based measurements reveal unseen differences during air pollution episodes. *Environmental Science & Technology 55*(1), 120–128.

Levy, R. C., L. A. Remer, R. G. Kleidman, S. Mattoo, C. Ichoku, R. Kahn, and T. F. Eck (2010). Global evaluation of the collection 5 modis dark-target aerosol products over land. *Atmospheric Chemistry and Physics 10*(21), 10399–10420.

Lorenz, E. (1996). Predictability: a problem partly solved. In *Proceedings Seminar on Predictability*, Reading Berkshire, UK, pp. 1–18. ECMWF.

Lukosevicius, M. (2012). A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade*, pp. 659–686. Springer.

Malsiner-Walli, G. and H. Wagner (2011). Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics 40*(4), 241–264.

McDermott, P. L. and C. K. Wikle (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat 6*(1), 315–330.

McDermott, P. L. and C. K. Wikle (2018). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics 30*(3), e2553.

McDermott, P. L. and C. K. Wikle (2019). Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy 21*(2), 184.

Miller, J. and H. Safford (2012). Trends in wildfire severity: 1984 to 2010 in the sierra nevada, modoc plateau, and southern cascades, california, usa. *Fire Ecology 8*, 41–57.

Monteiro, A., M. Lopes, A. I. Miranda, C. Borrego, and R. Vautard (2005). Air pollution forecast in portugal: a demand from the new air quality framework directive. *International Journal of Environment and Pollution 25*(1/2/3/4), 1–9.
Neto, J. H. V., A. M. Schmidt, and P. Guttorp (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63(1), 103–122.

Paciorek, C. J. and M. J. Schervish (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 17(5), 483–506.

Risser, M. D. and C. A. Calder (2017). Local likelihood estimation for covariance functions with spatially-varying parameters: The convospat package for r. *Journal of Statistical Software, Articles* 81(14), 1–32.

Shen, P., P. Crippa, and S. Castruccio (2021). Assessing urban mortality from wildfires with a citizen science network. under review.

Sorensen, A., R. Jordan, S. LaDeau, D. Biehler, S. Wilson, J.-H. Pitas, and P. Leisnham (2019). Reflecting on efforts to design an inclusive citizen science project in west baltimore. *Citizen Science: Theory and Practice, 4*(1), 1–13.

South Coast Air Quality Management District (2021). Air quality sensor performance evaluation center. [http://www.aqmd.gov/docs/default-source/aq-spec/summary/purpleair-pa-ii---summary-report.pdf?sfvrsn=16](http://www.aqmd.gov/docs/default-source/aq-spec/summary/purpleair-pa-ii---summary-report.pdf?sfvrsn=16). last accessed 2021/04/14.

Wolberg, G. and I. Alfy (1999). Monotonic cubic spline interpolation. In *Proceedings of Computer Graphics International*, pp. 188–195.

World Health Organization (2014). *Global status report on noncommunicable diseases 2014*. World Health Organization.

Xu, X. and W. Ren (2019). Prediction of air pollution concentration based on mrmr and echo state network. *Applied Sciences* 9(9), 1811.

Zhang, Y., J. J. West, R. Mathur, J. Xing, C. Hogrefe, S. J. Roselle, J. O. Bash, J. E. Pleim, C.-M. Gan, and D. C. Wong (2018). Long-term trends in the ambient pm$_{2.5}$- and o$_3$-related mortality burdens in the united states under emission reductions from 1990 to 2010. *Atmospheric Chemistry and Physics* 18(20), 15003–15016.