Flow Field Forecasting: An Introduction

Michael R. Frey[1]
Kyle A. Caudle[2]

[1]Bucknell University, Lewisburg, PA, 17837
[2]South Dakota School of Mines and Technology, Rapid City, SD, 57701

Abstract
A statistical learning methodology, called flow field forecasting, is introduced for predicting the future of a univariate time series. Flow field forecasting draws information from the interpolated flow field of an observed time series to build a forecast step-by-step. Flow field forecasting is premised on the principle of inductivism. We present flow field forecasting, along with a discussion of its motivating principle, examples with comparisons to other major forecasting techniques and a statistical error analysis.

Key Words: times series, forecasting, inductivism, penalized spline regression, Gaussian process regression

1. Introduction

The theoretical challenge and practical risk involved in making predictions beyond the domain of available data are well recognized among statisticians. Time series forecasting, involving as it does extrapolation from past data to future times, is one of the most challenging and problematic of statistical learning tasks. Today many general-purpose forecasting methods [1, 2], exponential smoothing [3, 4], Box-Jenkins ARIMA modeling [5], and artificial neural networks [6, 7] among them, are available to address the essential challenge posed by forecasting. Each of these established techniques has significant limitations, and still other forecasting methods are being developed—for example, wavelet-based methods [8].

Established forecasting methods do not always effectively address the problems posed by rapidly accumulating or very long data records. Some of these established methods are too computationally intensive, others do not readily yield error estimates and, most importantly, none can automatically (i.e., without human guidance) select a model, set procedure parameters, and screen the results. Regression modeling, for example, is a highly interactive procedure in which different models are tried and assessed against various criteria to achieve an acceptable result. Box-Jenkins ARIMA forecasting succeeds best when judgments are made from the data about model order and the multi-dimensional numerical optimization involved in parameter estimation is monitored. Even neural networks, which are essentially a highly flexible class of nonlinear regression models, involve significant human guidance in their training phase. Generally, the neural network model is over-parameterized, and the optimization problem that sets the network weights is non-convex and prone to instability [9]. We note, too, that ARIMA modeling and spectral/wavelet methods are generally formulated for data uniformly spaced in time, and adapting these methods to non-uniform spacings is not straightforward. These and similar concerns lead us to propose a new framework for forecasting, called flow field forecasting. Flow field forecasting, while retaining important strengths of established forecasting methods, designedly meets the challenge posed by the rapidly growing sizes of observational data sets and the flow volumes seen on modern data networks.

The remainder of the paper is organized as follows. Section 2 presents the premise—the principle of inductivism—on which flow field forecasting is based. Section 3 details the mechanics of flow field forecasting, including the three basic steps for making a forecast, and section 4 addresses the statistical error that accompanies a flow field forecast. Section 5 demonstrates flow field forecasting in head-to-head comparisons with ARIMA forecasting, exponentially weighted averaging, and artificial neural network modeling. Some broader remarks about flow field forecasting are made in Section 6. The computational efficiency of flow field forecasting is treated in [10, 11].

2. Inductivism—a premise for forecasting

Flow field forecasting is premised on the principle of (naive) inductivism recognizable from the philosophy of science [12]. To understand what this means, suppose, for example, that the past record of a statistical process shows a particular dynamic (or history) H_1, after which the process is seen to drop precipitously (change d_1) as shown in Fig. 1a. Suppose, maybe, that there are no counterexamples to this association in the data record and even suppose that a similar history H_2 in the record is seen to be followed be a corresponding similar change d_2. Then, we might reasonably anticipate that, if currently the process is changing according to a history H' that is similar to H_1 and H_2, it will likely in the near future undergo a precipitous drop d' similar to d_1 and d_2. This inductive reasoning is the basic premise of flow field forecasting: if a process exhibited a history H in the past followed by a level change d, then if the process is presently on a course H' similar to H, it will likely next undergo a level change d' similar to d.

We depict histories in Fig. 1a as made up of short sequences of consecutive observations in the data record. The observations need not be consecutive; the histories might involve, as shown in Fig. 1b, disjoint subsequences of observations.

The choice of the form of the history is one way in which flow field forecasting is flexible. The essential point of flow field forecasting is to build an inductive interpolation of the immediate future change d' from the current history H' and the past observed history-change associations. It is important to point out, also, that flow field forecasting bases its histories on sequences of knots, not sequences of observations. This last point becomes clearer in the next section where the actual steps in a flow field forecast are presented.

It is our experience that inductivism as a principle for forecasting is often enough applicable and sufficiently flexible that it offers a useful basis for forecasting. A concrete illustration of the applicability of flow field forecasting to energy management in gasoline/electric hybrid automobiles is offered in [13].

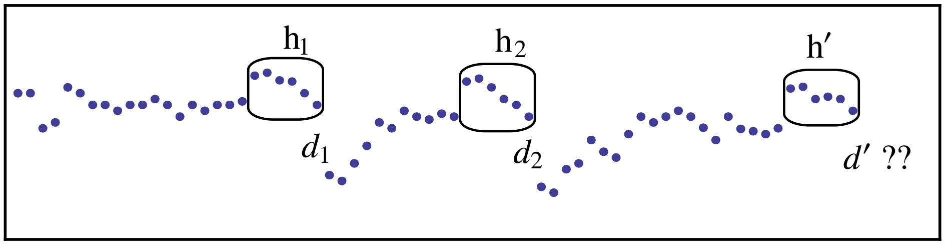

Figure 1a. Past histories $\mathbf{h}_1$, $\mathbf{h}_2$ and their associated changes d_1, d_2. These associations are used to interpolate the change d' associated with the current history $\mathbf{h}'$. The histories in this example have consecutive components.

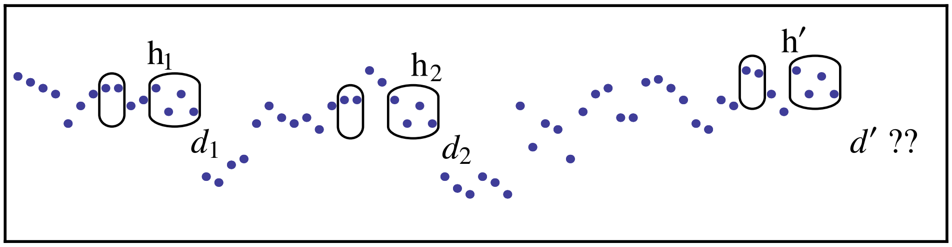

Figure 1b. Past histories $\mathbf{h}_1$, $\mathbf{h}_2$ and their associated changes d_1, d_2. These assocations are used to interpolate the change d' associated with the current history $\mathbf{h}'$. The histories in this example have non-consecutive components.

3. Flow field forecasting in three steps

Explicit, computationally efficient procedures have been identified [14] to translate the general premise of flow field forecasting into a method of making a forecast from past observations of a process. The inductivism paradigm requires: 1) from the data, a clear picture of the different histories that occurred in the process's past and the changes that followed in each there from, 2) an interpolation of process changes from histories seen in the past record to histories not available in the record, and 3) a mechanism for using interpolated change in the process to build a forecast from the present into the future. Flow field forecasting is a framework of statistical procedures that in three steps addresses each of these requirements.

Step I of flow field forecasting represents the time series record as a process with additive (homoscedastic) noise and estimates the underlying dynamic of the process via a penalized spline regression [15]. This spline smoother of the observed process is a set of process levels and level changes for each time in the set of spline knot times. Subsequent steps II and III of the flow field forecast are based solely on this spline smoother. The collection of information in the spline smoother is called the process data skeleton [14], and because the size of the skeleton is only loosely related to and much smaller than the size of the set of original data, this stage robustly scales even very large data sets to a more manageable size and supports computation in cases of very limited resources. Standard penalized spline regression involves a numerical search for the appropriate amount of smoothing. This numerical search is obviated by an asymptotic result that approximates the smoothing parameter λ by a simple calculation [16]. Also, by "stitching" the penalized spline regression can accommodate heteroscedasticity, be calculated still more efficiently and be updated in real time [11].

Step II of flow field forecasting applies Gaussian process regression [17] to the process skeleton to create an interpolator for future process change based on observed past process changes. Gaussian process regression derives this interpolator from the correlation of the process changes. This correlation must be estimated, usually within a parametric model such as the squared exponential model [17].

Step III uses the step II interpolator to step-by-step predict the process forward to the desired forecast horizon. We now describe in technical detail these three steps.

Step I: Extracting the skeleton

We are given N observations $\{Y_i, i = 1, \ldots, N\}$ of a process Y with their associated (not necessarily uniformly spaced) observation times $t_1 < t_2 < \ldots < t_N$, and we want to forecast Y at a time $t_F > t_N$. We assume that

$$Y_i = S(t_i) + \varepsilon_i \tag{1}$$

where $\{\varepsilon_i, i = 1, \ldots, N\}$ is a set of uncorrelated random variables with zero mean and common variance σ^2. The set $\{S_i = S(t_i), i = 1, \ldots, N\}$ constitutes the non-random, systematically determined component of the observation record. The method of penalized splines [15] is used to fit to the data a semiparametric regression model of the form

$$Y = \beta_0 + \beta_1 t + \sum_{k=1}^{K} \beta_{k+1}(t - \kappa_k)_+ + \varepsilon \tag{2}$$

where

$$(t - \kappa_k)_+ = \begin{cases} t - \kappa_k, & t \leq \kappa_k \\ 0, & t > \kappa_k \end{cases}$$

and the model parameters β_k are estimated from the data. The K knots κ_k are spaced uniformly within the range of observations t_i such that $t_1 < \kappa_1$ and $\kappa_K < t_N$. Without loss of generality the observation times t_i are assumed to be coded such that the knots are $\kappa_k = k\Delta$ where Δ is the chosen knot spacing and such that $t_1 > 0$. Add $\kappa_0 = 0$ to the set of knots (for a total of $K + 1$ knots), and let $b_k = \hat{\beta}_k$ be the penalized spline estimate of β_k for $k = 0, 1, \ldots, K + 1$. Then the smoothed responses $s_k = \hat{S}(\kappa_k)$ at the knot times $\kappa_k = k\Delta$ are

$$s_k = b_0 + \Delta \sum_{j=0}^{k-1} d_j, \quad k = 0, 1, \ldots, K$$

where the d_k are the estimated (forward) response derivatives

$$d_k = \sum_{j=1}^{k+1} b_j, \quad k = 0, 1, \ldots, K \tag{3}$$

at these times. Assembled with their knot times, these level and change estimates

$$(\kappa_k, s_k, d_k), \quad k = 0, 1, \ldots, K \tag{4}$$

constitute the skeleton of the original data. Only this skeleton is used subsequently in steps II and III to construct flow field forecasts. While we started with a data record whose size is of order N, we go forward with only the skeleton (4) with size of order $K < N$. This contributes to flow field forecasting's applicability to large data sets and voluminous data flows. Our choice in (2) of linear basis functions $(t - \kappa_k)_+$ with uniformly spaced knots is not only the simplest choice, it is the appropriate choice. Quadratic or cubic basis functions [15] would not have yielded the change estimates (3) so summarily, and non-uniformly spaced knots would have led to incommensurable estimates across the different knot times.

Step II: Interpolating the flow field

Step II of flow field forecasting builds an interpolator of the change $\dot{S}(t)$ in the systematically determined component $S(t)$ of the process Y, basing this interpolator on the information contained in the data skeleton. The point (s_k, d_k) in the data skeleton (4) estimates the derivative $\dot{S}(t)$ as the forward rate of change d_k at the knot time κ_k where the level was estimated to be $s_k = \hat{S}(\kappa_k)$. More particularly, the skeleton associates the estimated change d_k with the estimated process level s_k. Using the $K + 1$ points (s_k, d_k) in the skeleton, we can consider interpolating $\dot{S}(t)$ for other levels s not in the skeleton. In fact, the skeleton is providing estimates of

$\dot{S}(t)$ not just for different levels s_k, but with much greater specificity. For example, the skeleton can be interpreted to say that, when $S(t)$ was s_{k-1} with change d_{k-1} and subsequent level s_k at corresponding times κ_{k-1} and κ_k, the derivative at time κ_k was estimated to be d_k. Here we are treating the relevant history to be $\mathbf{h}_k = (s_{k-1}, s_k, d_{k-1})$. Flow field forecasting allows the history to be defined to include any number and combination of previous levels s_k and changes d_k. Consequently, the skeleton, for all its simplicity, is sufficient for interpolating a change $\dot{S}(t)$ from, even, a rather detailed history. The derivative $\dot{S}(t)$ viewed as a function of this multi-component history is called the flow field.

Gaussian process regression [17] is a natural choice for interpolating $\dot{S}(t)$; it is computationally efficient, it readily yields error estimates, it is very flexible about the size and content of the predictor history, and it has the desirable property that the default interpolated change is $\dot{S}(t) = 0$ in cases where the current history is far different from all the histories available in the skeleton. Let $\mathbf{h}_* = (\mathbf{s}_*, \mathbf{d}_*)$ be the current history involving present and past levels $\mathbf{s}_*$ and past changes $\mathbf{d}_*$ for which we want to interpolate the derivative d_* from the flow field. Let $\mathbf{k}(\mathbf{h}_*)$ be the column vector of kernels $k(\mathbf{h}_*, \mathbf{h}_k)$ where the $\mathbf{h}_k$ are the histories in the data skeleton. A kernel $k(\mathbf{h}, \mathbf{h}')$ in this context is a covariance model for the changes d and d' associated with the histories $\mathbf{h}$ and $\mathbf{h}'$. A standard covariance model for this purpose is the squared exponential covariance model [17]. Let $\mathbf{K}$ be the matrix with i, jth entry $k(\mathbf{h}_i, \mathbf{h}_j)$. Finally, let $\mathbf{y}$ be the column vector of derivatives d_k in the skeleton corresponding to the histories $\mathbf{h}_k = (\mathbf{s}_k, \mathbf{d}_k)$. Then the derivative d_* interpolated by Gaussian process regression corresponding to the history $\mathbf{h}_*$ is [17]

$$d_* = \mathbf{k}(\mathbf{h}_*)^\top \widetilde{\mathbf{y}} \tag{5}$$

where $\widetilde{\mathbf{y}} = \mathbf{K}^{-1}\mathbf{y}$. The matrix inversion of $\mathbf{K}$ required for (5) can be accomplished efficiently and in numerically stable fashion using Cholesky decomposition. Notice that, for successive uses of (5) for different histories $\mathbf{h}_*$, the vector $\widetilde{\mathbf{y}} = \mathbf{K}^{-1}\mathbf{y}$ must be calculated just once. This is a general feature of Gaussian process regression that contributes significantly to the overall efficiency of flow field forecasting.

Step III: Iterating to the future

Step III of flow field forecasting uses the flow field interpolator (5) prepared in step II to forecast the path of the process Y out to the desired future time t_F. The forecasting proceeds incrementally. We start by estimating the process level at time $\kappa_{K+1} = (K+1)\Delta$ (one knot increment Δ beyond the skeleton) by $s_{K+1} = s_K + \Delta d_K$. With this we assemble a new current history

$$\mathbf{h}_{K+1} = (\mathbf{s}_{K+1}, \mathbf{d}_{K+1}) \tag{6}$$

for the time κ_{K+1}. We return with this updated current history to the interpolated flow field and predict the next change d_{K+1} by

$$d_{K+1} = \mathbf{k}(\mathbf{h}_{K+1})^\top \widetilde{\mathbf{y}} \tag{7}$$

in accordance with (5). We use this change prediction to find

$$s_{K+2} = s_{K+1} + \Delta d_{K+1} \ . \tag{8}$$

We repeat (6), (7) and (8), cycling in general through the steps

$$\begin{aligned} \mathbf{h}_{K+m} &= (\mathbf{s}_{K+m}, \mathbf{d}_{K+m}) \\ d_{K+m} &= \mathbf{k}(\mathbf{h}_{K+m})^{\top}\widetilde{\mathbf{y}} \\ s_{K+m+1} &= s_{K+m} + \Delta d_{K+m} \\ \kappa_{K+m+1} &= (K+m+1)\Delta \end{aligned} \tag{9}$$

until the time $\kappa_{K+M} = \Delta(K+M)$ is reached that just exceeds the desired forecast time t_F. This construction yields not only a forecast for time t_F, it also predicts the path leading up to the forecast.

In the next section we estimate the statistical error that accompanies a flow field forecast, and in section 5 we present some applications of flow field forecasting. We close this section by mentioning three attractive properties of flow field forecasting that derive from the three steps described above:

- Flow field forecasting in step I reduces potentially very long data records with possibly non-uniformly spaced observations to a skeleton that may have only a relatively small number (50∼200) of uniformly spaced knots plus the process level and change at those knot times. Converting the original data record into the step I data skeleton achieves a very useful degree of data reduction and standardization.
- Penalized spline regression is computationally efficient in most respects. To optimize its efficiency, we replace the standard numerical search for the optimal smoothing by an approximation [16]. The step II Gaussian process regression and the step III extrapolation mechanism are also computationally efficient. This makes flow field forecasting attractive in settings with only limited computational resources or where data flow volume is an issue.
- Flow field forecasting is adaptable to a variety of forecasting settings by, for example, choice of history. However, once such choices are made and a few parameters are estimated or otherwise set, flow field forecasting performs autonomously, with no interactive supervision of a skilled analyst required. This is desirable in situations involving high data flows or where many forecasts must routinely be made.

4. Forecast error

Steps I and II of flow field forecasting are each a source of statistical error. The penalized spline regression in step I readily yields an estimate of the variance σ^2 of the noise ε in (1). Similarly, the Gaussian process regression in step II yields standard errors with its predictions. The errors from these two steps combine and are reflected in step III in the final forecast in a straightforward fashion, as we now show.

The variance of the error ε in (1) is estimated in step I by

$$\hat{\sigma}^2 = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{n - df_{\text{res}}} \tag{10}$$

where $\mathbf{Y}$ is the vector of observations, $\mathbf{Y}$ is the spline fit in (2) and df_{res} is the number of residual degrees of freedom [15]. This error makes an uncorrelated contribution to the total statistical error in the forecast.

The skeleton found in step I is a set of flow field values for different histories. The Gaussian process regression flow field estimates in step II are interpolations of these flow field values. Gaussian process regression readily yields standard errors for its estimates. The variance of the flow field estimate (5) is [17]

$$V[d_*] = \tau^2 - \tau^2 \mathbf{k}(\mathbf{h}_*)^\top \mathbf{K}\mathbf{k}(\mathbf{h}_*) \tag{11}$$

where τ^2 is the variance of the Gaussian process regression response. Here as in (5) a fast, numerically stable calculation of (11) is possible using the Cholesky decomposition of $\mathbf{K}$. The variance τ^2 in (11) is estimated by the sample variance S_d^2 of the observed changes d_k recorded in the skeleton.

We now combine (10) and (11) to identify the standard error of our flow field forecast constructed by successive applications of (9). The squared standard error is

$$\hat{\sigma}^2 + \Delta^2 S_d^2 \left(M - \sum_{m=1}^{M} \mathbf{k}(\mathbf{h}_{K+m})^\top \mathbf{K}\mathbf{k}(\mathbf{h}_{K+m}) \right) . \tag{12}$$

The number M of knot increments needed to reach t_F is approximately

$$M \approx \frac{t_F - t_N}{\Delta} .$$

Thus, according to (12), the standard error of the flow field forecast at time t_F is approximately

$$\sqrt{\hat{\sigma}^2 + \Delta(t_F - t_N) S_d^2 (1 - \bar{\omega})} , \tag{13}$$

where $\bar{\omega}$ is the average of is approximately

$$\omega_m = \mathbf{k}(\mathbf{h}_{K+m})^\top \mathbf{K}\mathbf{k}(\mathbf{h}_{K+m}) , \quad m = 1, 2, \ldots, M .$$

The ω_m, and their average $\bar{\omega}$, are bounded between 0 and 1. The average $\bar{\omega}$ reflects the overall usefulness of the information in the interpolated flow field for the desired forecast. For example, maybe the sequence of future forecasts traverses regions of the flow field not represented in the skeleton; in these cases the interpolated flow field provides little or no useful information, and $\bar{\omega} \approx 0$. Or maybe the future process always lands on or near histories where the flow field has been observed; in these cases the Gaussian process regression prediction has zero or near zero error, and $\bar{\omega} \approx 1$. The average $\bar{\omega}$ is a measure of the extent to which the flow field model explains the future; $\bar{\omega}$ is the counterpart for flow field forecasting of the coefficient of determination in least squares regression.

The square of the standard forecast error in (13) is roughly linear with time into the future. Since $0 \le \bar{\omega} \le 1$, we obtain from (13) the simple upper bound

$$\sqrt{\hat{\sigma}^2 + \Delta(t_F - t_N) S_d^2}$$

for the standard error. Notably, this error bound can be calculated from just the spline fit in step I.

5. Demonstrations of flow field forecasting

We begin with demonstrations of flow field forecasting based on two fabricated time series, each with $N = 1,000$ uniformly spaced observations. Each time series was fabricated according to model (1) with $\sigma = 50$. Fig. 2 shows each time series with its extracted skeleton based on $K = 40$ knots. In each case the spline fit (2) satisfactorily estimates σ, with respective estimates $\hat{\sigma} = 51.2$ and $\hat{\sigma} = 50.7$ for the two series. Fig. 2 also shows the incremental development of forecasts for the two series, going out $M = 15$ knot steps—against roughly $K = 40$ knot steps spanned by the data. The forecasts are given with 50% and 95% confidence error bounds. These bounds are based on the standard error in (13), and they assume a normal error distribution.

The forecasts and the growth of the forecasting error in the two cases of Fig. 2 are distinctly different. In the case of data set 1, the forecast is accessing a region of the interpolated flow field with significant information, and the forecast is effectively capturing that information, reflecting, for example, the somewhat periodic dynamic in the past data. The information drawn from the flow field allows the forecast to proceed with error that grows relatively slowly. In the case of data set 2 on the other hand, the present time is characterized by levels declining to around 400 and there is very little data for this history in the flow field. When the flow field is accessed to construct a forecast, it recommends only small changes, and it does so with large cautionary standard errors. Thus the prediction error grows much more rapidly in this latter case.

We next compare flow field forecasting with standard forecasting methods, using data sets commonly employed for this type of comparison. In particular, we look at head-to-head comparisons of flow field forecasting with Box-Jenkins ARIMA modeling [5], exponential smoothing [3, 4] and artificial neural networks [6, 7]. The four data sets we use for our comparisons are:

Births ($N = 365$) The daily total number of female births in California in 1959 [18]. These numbers vary between about 30 to 60 per day.

Pines ($N = 625$) Tree rings index data for Ponderosa pines in Bryce Water Canyon, Utah, from 1340 to 1964 [19].

S&P ($N = 388$) The Standard & Poor 500 stock index is employed as a measure of the general level of U.S. stock prices, as it includes both “growth” stocks and less volatile “value” stocks. The components of the S&P 500 index are selected by committee and change over time [20].

Spots ($N = 288$) Wolf’s relative sunspot numbers from 1700 to 1987 [21]. The relative sunspot number counts a combination of sunspots and groups of sunspots observed on the Sun. These data exhibit a 10.5-year cycle. The attribution of these data is discussed in [22].

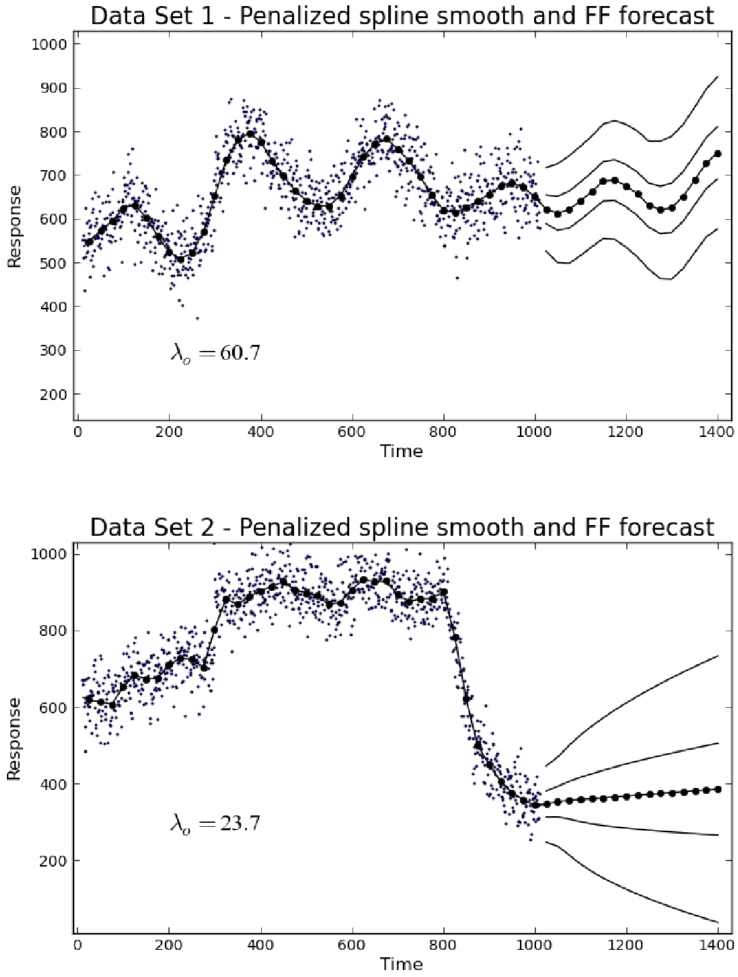

Figure 2. Two fabricated time series with flow field forecasts.

For each data set the number K of knots used for flow field forecasting was the smaller of $N/4$ and 35, and the amount of smoothing for the penalized spline regression in step 1 was obtained using Wand's aymptotic expression [16]. Also, we estimated the characteristic length (a key parameter in the squared exponential covariance model used in step II for the Gaussian process regression) by $20/K$. These crude rules-of-thumb somewhat limited the relative performance of flow field forecasting in our comparisons, but their strict application also protected against possibly unfairly "tweaking" the flow field forecast. We made similar straightforward choices for the other three competing forecasting methods, in an attempt to duplicate typical performances that might be seen in practice.

For our comparisons we set aside the ten most recent observations from each data set, used each of the four methods to forecast these values and then scored the four methods based on their mean absolute errors

$$\mathrm{MAE} = \frac{1}{10}\sum_{i=1}^{10} |Y_i - \hat{Y}_i| \, , \tag{14}$$

where Y_i is the known value and $\hat{Y}_i$ is the corresponding forecast value. Since flow field forecasting only gives forecasts at knot times, for these comparisons forecasts for times between knots were derived by simple quadratic interpolation. Shown in Fig. 3 is the relative performance of the four forecasting techniques with each of the data sets above. Fig. 3 gives the MAE of each technique as a precent of the worst performing technique for each data set; so for each data set the worst performer had 100% relative MAE and the best performer had the lowest relative MAE. Flow field forecasting was best or second best for each data set. Of course, no comparisons based on four—or four hundred—data sets can be determinative. The present comparison, though, does suggest that flow field forecasting can perform as well as, or better than, the most popular statistical forecasting techniques.

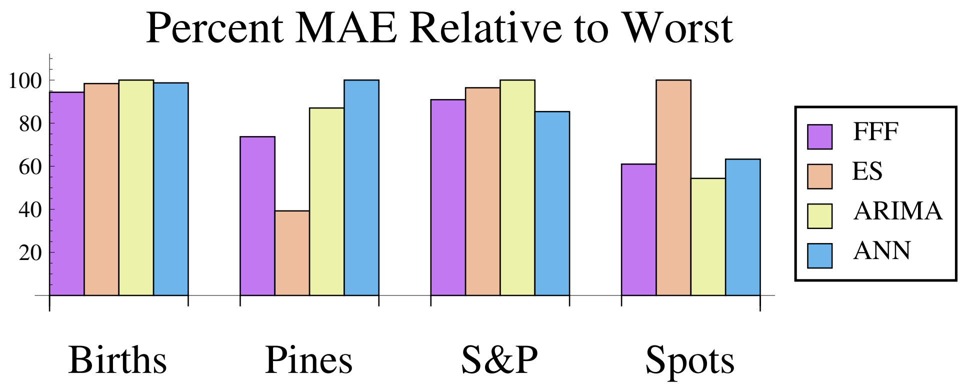

Figure 3. Relative performance of flow field forecasting (FFF), exponential smoothing (ES), ARIMA modeling and artificial neural network (ANN) modeling with four common data sets.

6. Closing remarks

Flow field forecasting is perhaps best viewed first and foremost as an (inductivist) framework for forecasting, in which choices of techniques are available for steps I and II. For example, kernel averaging might be used instead of penalized spline regression to build the step I data skeleton. Or an artificial neural network might be used in place of Gaussian process regression to interpolate the flow field in step II. Preprocessing prior to steps I or II might be applied, also, within the framework of flow field forecasting. Deseasonalization [23], detrending, differencing, log transformation, and multiple averaging, in particular, have all been found in one or more contexts with other forecasting methods to have an impact on forecasting performance [24]. The relative merits of these and other implementations and modifications of flow field forecasting are attractive subjects for future study.

Histories in flow field forecasting are short sequences of (not necessarily consecutive) knots developed from the penalized spline regression in step I of the flow field forecast. Our idea of history in flow field forecasting is notionally similar to that of a motif used in data mining. The purposes of motifs and histories are very different, though. Data mining seeks, as a task in *unsupervised* learning, to extract motifs [25] and, for example, evaluate their significance [26]. Flow field forecasting, by contrast, is performing *supervised* learning to interpolate a new change for a current history based on a set of past observed associations between history and change. We prefer the distinctive term "history" partly because of this difference.

We noted in the Introduction the essential challenge that time series forecasting presents. Extrapolation into the future is problematic (see Fig. 4) even in the best of circumstances. A remarkable feature of flow field forecasting is that it translates this extrapolation into a cumulative series of interpolations; with flow field forecasting, instead of forecasting into the future, we interpolate changes in the flow field. This in no way reduces the risk inherent in forecasting, but it does offer a new way to understand and address that risk.

Acknowledgment

This work was supported by the Office of Science of the U.S. Department of Energy, grant no. 128382.

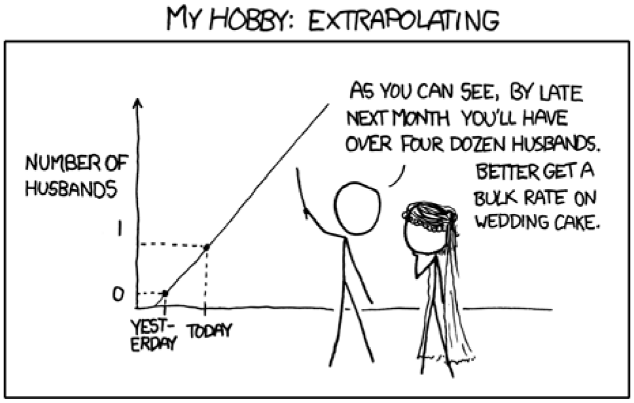
MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Figure 4. Forecasting is problematic [27].

References

[1] [1] J.G. De Gooijer and R.J. Hyndman, “25 Years of time series forecasting,” *International Journal of Forecasting*, **22**, 443–473 (2006).

[2] D.C. Montgomery, C.L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, Hoboken, NJ: Wiley-Interscience (2009).

[3] R.G. Brown, *Statistical Forecasting for Inventory Control*, New York: McGraw-Hill (1959).

[4] R.G. Brown, *Smoothing, Forecasting, and Prediction*, Englewood Cliffs, NJ: Prentice Hall (1963).

[5] G.E.P. Box, G.M. Jenkins and G.C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th Ed., Hoboken, NJ: Wiley (2008).

[6] S. Haykin, *Neural Networks and Learning Machines*, 3rd Ed., New York, NY: Pearson (2009).

[7] T. Hill, L. Marquez, M. O’Connor and W. Remus, “Artificial neural network models for forecasting and decision making,” *International Journal of Forecasting*, **10**, 5–15, (1994).

[8] P. Fryzlewicz, S. Bellegem, and R.Sachs, “Forecasting non-stationary time series by wavelet process modeling,” *Annals of Statistical Mathematics*, **55**, (2003).

[9] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed., New York, NY: Springer, (2009).

[10] M.R. Frey and K.A. Caudle, “Characteristic length estimation for flow field forecasting,” *Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, June 24–27, (2012).

[11] K.A. Caudle and M.R. Frey, “Continuous updates of penalized spline regression for flow field forecasting,” *Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, June 24–27, (2012).

[12] D. Gillies, *Philosophy of Science in the Twentieth Century: Four Central Themes*, Oxford, UK: Blackwell Publishers (1993).

[13] M.R. Frey and K.A. Caudle, “A new premise for forecasting,” *Proceedings of the 17th Annual Army Conference on Applied Statistics*, Annapolis, Maryland, October 19–21, (2011).

[14] M.R. Frey and K.A. Caudle, “Introducing flow field forecasting,” *10th International Conference on Machine Learning and Applications*, Honolulu, Hawaii, December 18–21, (2011).

[15] D. Ruppert, M. P. Wand and R. J. Carroll, *Semiparametric Regression*, New York, NY: Cambridge University Press (2003).

[16] M.P. Wand, “On the optimal amount of smoothing in penalized spline regression,” *Biometrika*, **86**, 936–940, (1999).

[17] C.E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press (2006).

[18] Newton, "Daily Female Births in California in 1959," http://datamarket.com

[19] K.W. Hipel and A.I. McLeod, "Preservation of the Rescaled Adjusted Range," *Water Resources Research*, **14**, 491–517, (1978). Ponderosa pine trees, Bryce Water Canyon, Utah, 1340 to 1964," http://datamarket.com

[20] Quarterly Standard & Poor 500 Numbers, http://datamarket.com

[21] Tong, "Wolf's sunspot numbers, 1700–1988," http://datamarket.com

[22] A.J. Izenman, "An Historical Note on the Zurich Sunspot Relative Numbers," Journal of the Royal Society, Series A, **146**(3), 311–318, (1983).

[23] D.M. Miller and D. Williams, "Damping seasonal factors: shrinkage estimators for the X-12-ARIMA program," *International Journal of Forecasting*, **20**, 529–549, (2004).

[24] N.K. Ahmed, A.F. Atiya, N.E. Gayar and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, **29**(5–6), 594–621, (2010).

[25] P. Ferreira, P. Azevedo, C. Silva, and R. Brito, "Mining approximate motifs in time series," *Discovery Science*, Secaucus, New Jersey, Springer, 89–101, (2006).

[26] N.C. Castro and P.J. Azevedo, "Significant Motifs in Time Series," *Statistical Analysis and Data Mining*, **5**, 35–53, (2012).

[27] xkcd, *A Webcomic of Romance, Sarcasm, Math and Language*, http://xkcd.com/605/, (2012).