Paper SA06

Automated Real-Time Forecasting of Stream Conditions with SAS®

Samuel T. Croker, Independent Consultant
Shane L. Hornibrook, Independent Consultant
Tomonori Ishikawa, USC Department of Statistics, Columbia, SC

ABSTRACT

When you are a SAS® programmer planning a canoe trip down the Edisto River in South Carolina, you want to know what river conditions to expect. One easy-to-run program, incorporating three critical features of SAS, can generate reasonably-accurate, short-term forecasts of river level using real-time data provided online by the United States Geological Survey. First, using the URL access method of the FILENAME statement and basic DATA step techniques, you can import data from the online USGS water data repository. Second, you can analyze the resulting data set and produce forecasts with SAS High-Performance Forecasting procedures by choosing among models as diverse as ARIMA, ESM and UCM. Third, you can display effective graphs of the resulting forecasts with SAS/GRAPH. You can easily apply these techniques to other data and analyses to provide a complete SAS-based solution from web-enabled ETL to automatic generation of forecasts to web publication of the results. Keywords: Base SAS, SAS High-Performance Forecasting, PROC HPFENGINE, PROC HPFARIMASPEC, SAS/GRAPH.

INTRODUCTION

As a skilled SAS programmer, you can easily generate accurate forecasts of short-term stream conditions by analyzing river data provided online by the U.S. Geological Survey (USGS). Beginning with web-oriented DATA step techniques, you can import data from online USGS water data repositories. Next, you can produce forecasts using High-Performance Forecasting (HPF) procedures to choose the best model from among a wide range of candidates. Possible models include, but are not limited to, Auto-Regressive Integrated Moving Average (ARIMA) models, Exponential-Smoothing Models (ESM) and Unobserved Components Models (UCM). Finally, you can create attractive figures using SAS/GRAPH, which, to come full-circle, you can post to a website for easy dissemination.

Forecasting stream conditions can be quite challenging given the dizzying array of potential man-made, geophysical and meteorological factors which can influence the behavior of a river's characteristics such as stage and flow. Confronting limited resources of time and effort, you can leverage the power of HPF to evaluate many models for many river data sets in a large-scale forecasting project. However, as any experienced statistician knows, there is simply no substitute for detailed knowledge about the data, and you should not use HPF as a "black box" replacement for familiarity with the data.

This paper begins by describing the river data which you can obtain on the web from the USGS. You will see how you can capture and prepare the data. Next, you will read about how to select appropriate forecast models using HPF procedures. Then, you will read about how to use SAS/GRAPH to visualize the forecasts. At the end of the paper, by comparing data from two rivers with contrasting characteristics, you can assess the potential strengths and weaknesses of HPF and review the importance of detailed data knowledge when you select a forecast model.

OBTAINING AND PREPARING THE RIVER DATA

AVAILABILITY OF THE DATA

The United States Geological Survey provides a wealth of data online. The USGS reporting system is extensive, and the data layout is fairly complex. Different monitoring stations collect different measurements, but almost all stations collect *stream stage*, the height of the water surface over a fixed point at the station. *Stream flow* is also observed by the USGS at many stations. Fewer stations collect other river characteristics such as water temperature, dissolved oxygen concentration or meteorological data.

Online, the USGS posts river data observations from the most recent 31 days. You would not want to base

long-term forecasts on data from such a limited time frame, but you could construct a larger data set which would be more appropriate for long-term forecasting by storing USGS data extracts and appending new observations to the archived data every 31 days.

SAS METHODS TO RETRIEVE THE DATA

You can access the USGS database by pointing-and-clicking from a web browser. However, to extract a large amount of data for a wide variety of stations on demand, you would want to write some type of automating script. While some programmers familiar with web application techniques may use Perl, it is not only possible, but also easy to use SAS to acquire data form the web. This avoids the complexity of marrying the output and input of two scripts and reduces another potential source of maintenance issues.

SAS facilitates hypertext retrieval through LIBNAME statements using the URL library type, also known as the "LIBNAME URL" method. Other SAS options include the FILENAME URL method or reading sockets. By reverse engineering the parameters in the USGS URL from a small sample of data requests, you can write a small macro loop to generate DATA step code and import the required data for all USGS stations used in this forecasting project.

GUIDING PRINCIPLES FOR DATA EXTRACTION CODE DESIGN

While a detailed examination of the code used to extract the USGS data is beyond the scope of this paper, a guiding principle is readily apparent: design the extraction to gather the data that is available instead of the data that you expect to find. Designing flexibility into your code is the key to success.

Ideally, you would design a data import scheme to examine the river characteristics available at each station and only attempt to import the data that is available at that particular station. Writing such data-driven import code affords flexibility and avoids the necessity of recoding when data availability changes in the source system. You can use the SAS macro language judiciously to generate code based on the data that you actually find instead of simply what you expect to find.

As it happens, the USGS provides reliable data on their websites, both in terms of *data longevity* (it is *where* you expect it to be) and *data quality* (it is *what* you expect it to be). The USGS site provides two core files that facilitate data extraction. First, the USGS lists the locations of the stations where it gathers data. Second, the USGS lists the particular river characteristics (such as temperature, dissolved oxygen, river stage, river flow, etc.) that it gathers at each station.

However, this is not true of many websites. Unfortunately, when "screen-scraping" or otherwise pulling data from internet sites, changes in structure, data gaps, and/or periods of unavailability are often the norm, not the exception. Thus, as a general rule, when you design a mechanism to gather data sourced from internet sites, it must be robust enough to gracefully degrade when presented with missing or unavailable data sources.

DATA PREPARATION

The defining characteristic of a *time series* is that it has observations at equally-spaced time intervals. In contrast, *transactional data* occurs at irregular intervals. In this case, in order to use time series techniques, the data must be pre-processed. SAS provides several utilities to accomplish this task: the EXPAND, TIMESERIES and HPFENGINE procedures.

PROC EXPAND is the oldest of these procedures and affords the most flexibility in accumulating transactional data to a time series. It can also perform other imputations and transformations that are useful. PROC TIMESERIES can be a little more intuitive for this task. But in the end, PROC HPFENGINE, a component of HPF, can eliminate the need for running multiple procedures. You will note that the analysis in this paper uses the HPFENGINE procedure for data preparation.

An important example of knowing as much as possible about the data is the ability to make distinctions between *structural zeros* and missing values. A structural zero occurs when zero is an actual value for the observation.

To examine this issue in the context of a river stage example, suppose a time series for stream stage has the following observations:

```
3.4 3.6 3.8 . 4.2
```

In order to proceed with time series analysis, you need to impute the missing fourth observation using the other observations. What value should you use for the missing value?

For stream stage data, if you set the missing value to zero, you would probably be making a mistake. Given the other observations, consider what that would imply about the behavior of the river. Setting the missing value to zero would imply a drop of 3.8 feet from the third observation to the fourth observation of 0 and a rise of 4.2 feet back to the fifth observation. Clearly, you would be wise to use an average interpolation in this situation, perhaps by using a value of 4, the average of the surrounding values of 3.8 and 4.2.

In contrast, suppose you were working with count data or rarely-occurring transactional data. In that case, you could potentially record many missing values if you did not record any observations during the accumulation period. Clearly, you would be wise to use a zero value in this situation since you actually observed no events.

Once again, there is no substitute for familiarity with the data you are analyzing.

FORECAST MODEL SELECTION AND THE HIGH PERFORMANCE FORECASTING SYSTEM

The SAS High-Performance Forecasting System is an easy-to-use system for evaluating forecast models. With a relatively small amount of code, it is possible to generate forecasts. In fact, it can automatically analyze sample data, chose a holdout-error-minimizing forecast model and output forecasts.

However, you would be foolish to use HPF as "black box" forecast generator, substituting its ease-of-use for explicit knowledge about the data and what models might be appropriate. Instead of misusing HPF to make model choices for you in the blind, you should use the power of HPF to augment the number of models that you can evaluate. This way, you can choose a better model by increasing the choice set for your optimal model.

MODEL CLASS IDENTIFICATION ISSUES

There are many books and papers written about time series identification and model fitting, and a thorough discussion of this material is well beyond the scope of this paper, but you should ask yourself several questions when examining data to forecast:

- Is the data time series data or is it transactional data?
- Are short-term or long-term forecasts needed?
- Does the data exhibit periodicity or seasonality?
- Are there exogenous variables that may help the forecast if they are included as covariates?

Your answers to these questions will help determine which models you should evaluate. As with any data-driven forecasting method, the quality of the forecast depends, not only on the data itself, but also on how well you understand and model the data. Auto-Regressive Integrated Moving Average (ARIMA) models are a very flexible class of models that can be very effective at long-term as well as short-term forecasting. However, they can require lots of data for accurate estimation. Exponential Smoothing Models (ESM) can provide very good in-sample, or one-step-ahead, forecasts but often do not perform well in the out-of-sample forecasting region. Unobserved Component Models (UCM) are useful when you know a great deal *a priori* about the structure of the data in terms of seasonalities, cycles and exogenous inputs.

MODELING CHALLENGES POSED BY RIVER DATA

River data presents unique forecasting challenges. River characteristics can be influenced by a wide range of man-made, geophysical and meteorological factors:

- The demand for electricity can influence river stage data recorded downstream from hydroelectric facilities. For example, during hot weather, hydroelectric facilities may release large quantities of water to generate power in order to meet increased demands for air conditioning. In this case, you can observe a square-toothed appearance if you plot the time series for river stage. You can see a strongly seasonal weekday pattern (Monday–Friday has the highest and longest peaks.) due to strong

power demand.

- River stage levels can also affect river flow rates. During times of extremely high or extremely low water levels, river flow can be very different compared to river flow rates when water levels are near median levels.
- Increasing or decreasing water levels can change the delay parameter when upstream station data is used as an inputs to transfer function models. The amount of this change depends greatly on the shape of the riverbed. Again, if the river spills over its banks, the relationship embodied in a forecasting model can change.

HOW YOU SHOULD USE HPF IN FORECASTING RIVER DATA

By now, you should understand how *not* to use HPF, but how is it useful? When can or should you use it? HPF can be an excellent forecasting tool as long as you are using it to augment your knowledge of a time series. If the typical models that forecast the data well for a particular station are known *a priori*, HPF can pick the best one and provide forecasts with little additional input. Thus, HPF can make the chore of forecasting many time series easier by automatically choosing a separate model with the best holdout forecast for each time series.

For example, if a river station has characteristics that make its data fit several models that are well-known to you, then HPF can easily choose the one that minimizes the holdout error among them. Or, if a river's condition doesn't change much, but you need to update forecasts quite often, HPF can ease the chore of maintaining forecasts.

VISUALIZING THE RESULTS WITH SAS/GRAPH

Once you have generated forecasts, you will want to see your results. Visualizing your forecasts is a much more effective than examining a single number metric like Mean Absolute Percentage Error (MAPE). The SAS code below will plot your data and forecasts in three regions: the training sample used to fit the model, the holdout sample used to evaluate an error measure for observed data that you did not use to fit the data and the pure forecasts themselves. You can learn more about Sam Croker's `threeregionforecast` macro in "Effective Forecast Visualization Using SAS®."

```
%macro plotforecast;
  proc sql noprint;
    select count(distinct site_no) into :numsites from &dataset;
    %let numsites=&numsites;
    select site_no
       , max(dtstamp) format=30.
       , min(dtstamp) format=30.
      into
         :siteno1-:siteno&numsites
        ,:maxdt1-:maxdt&numsites
        ,:mindt1-:mindt&numsites
      from &dataset
      group by site_no;
  quit;
  goptions device=jpeg;
  ods html;
  %do i=1 %to &numsites;
    %threeregionforecast(outfor(where=(site_no="&&siteno&i"))
      ,&&maxdt&i
      ,&&mindt&
      ,168
      ,168
      ,grtitle="sitename"
      ,dtm=dtstamp
      ,lciname=lower
      ,uciname=upper
```

```
      ,fcname=predict
      ,varname=actual
      ,dtformat=dthour.
      ,xinterval=hour.
      ,xminorticks=0
      ,ymajnum=8
      ,dtdisplay=tod5.
      ,acth=.5
      ,hatitle=""
      ,vatitle="Stage");
  %end;
  ods html close;
%mend;

%plotforecast;
```

SAS HIGH-PERFORMANCE FORECASTING FOR STREAM CONDITIONS

FORECASTING THE ASHLEY RIVER WITH HPF: A LITTLE KNOWLEDGE GOES A LONG WAY

The historic city of Charleston, SC is situated on a peninsula that lies between the Ashley River to the west and the Cooper River to the east. In fact, Charlestonians have a saying that the Ashley and Cooper Rivers meet in Charleston to form the Atlantic Ocean.

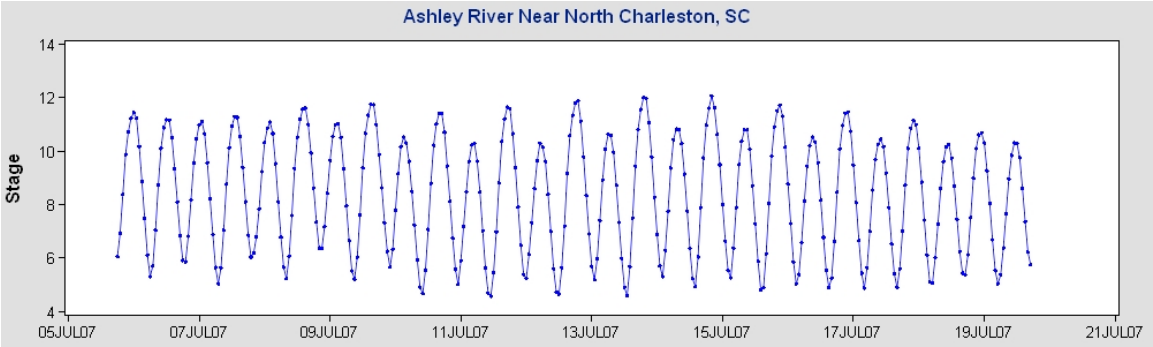

Figure 1: River Stage for USGS Station 02172081: Ashley River Near North Charleston, SC

As you can see in Figure 1, the river stage of the Ashley River at USGS station 02172081 is greatly influenced by tides. Tidal cycles are composed of a set of well-known, very regular cycles. If you incorporate them into a structural model, you can dramatically improve the forecast. Figure 1 displays hourly observations of river stage from July 5th, 2007 to July 19th, 2007. You can see a clear sinusoidal pattern in the river stage time series driven by the tides.

But, suppose you ignore the tidal information and simply feed the series into HPF, misusing HPFDIAGNOSE by giving it no other information. How good a forecast do you think HPF will produce? The offending code below is deceptively compact:

```
proc hpfdiag
  data=WORK.SC_rivers
  print=all
  repository=WORK.arima
  criterion=aic
  outest=diagest;
    by site_no;
    id dtstamp interval=hour accumulate=avg;
    forecast stage;
    ucm component=(all);
    arimax outlier=(detect=maybe) method=minic;
    esm;
    idm;
```

```
    trend dif=auto;
    transform type=auto;
run;
```

You can see the result in Figure 2. HPF chose a relatively simple ARIMA model which has poorer results than you might expect: the amplitude of the forecast is quickly attenuated to the mean in the forecast period.

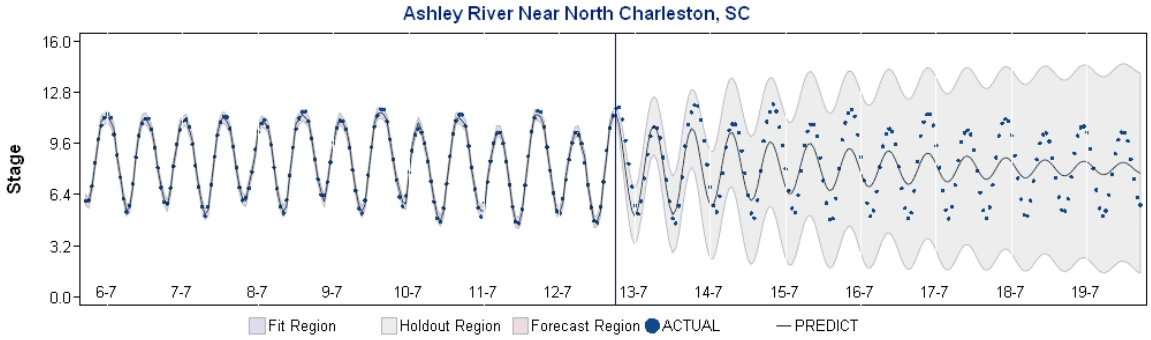

Figure 2: Naive HPFDIAGNOSE Forecast (ARIMA) for the River Stage of the Ashley River

But HPF can do considerably better if you use a bit of research on tidal cycles. You can examine the period of the primary tidal component cycles at a website maintained by the Oak Ridge National Laboratory:

```
    http://www.phy.ornl.gov/csep/CSEP/OM/NODE31.html
```

Using this little bit of information about the tides dramatically improves the forecast. Once again, the SAS HPF code to exploit this information is very compact:

```
proc hpfucmspec
    repository=WORK.arima
    label="Tidal Structural Model"
    name=UCMTIDAL;
  forecast symbol=stage;
  irregular;
  level;
  cycle period=12.42;
  cycle period=12.00;
  cycle period=24.06;
run;
```

Figure 3 illustrates the improvements.

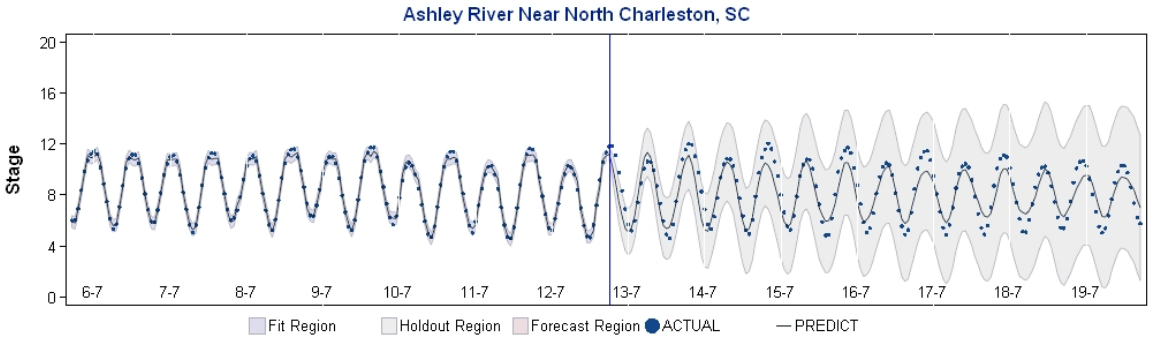

Figure 3: River Stage of the Ashley River: Forced UCM Tidal Model

USING HPF TO DETERMINE WHETHER OR NOT TO CANOE ON THE EDISTO RIVER

As you can see, depending upon the river, forecasting conditions can be tricky. The Edisto River in South Carolina is the longest free-flowing blackwater river in the South. Fortunately, it has no man-made, geophysical or meteorological factors that cause it to be a particularly difficult river to forecast.

Some smaller tributaries of the Edisto (un-navigably small for canoeing purposes) are subject to spikes in river stage following heavy rains, but the river stage of the main Edisto River exhibits the characteristics of a regular, long-memory time series. It is an ideal data set to feed to HPF to generate automated real-time forecasts.

For a canoe trip on the Edisto, you would be interested in the river stage on the main branch of the Edisto River near Givhans Ferry, downstream from the North Fork and South Fork branches of the Edisto River.

In this example, you would properly use the HPFDIAGNOSE procedure to suggest alternative forecast models, but not rely upon it exclusively for model selection. If you have developed a familiarity with the Edisto River, you would have performed some preliminary time series analysis and identified classes of ARIMA and cyclic UCM models as appropriate models. For example, using your knowledge of the Edisto River, you might generate ARIMA models using the PROC HPFARIMASPEC with P(0..3) Q(0..3) and D(0,1,7). By adding the models suggested by HPFDIAGNOSE to your set of model candidates, you are wisely using HPFDIAGNOSE as a check to see if you have missed any simple models that provide reasonably good forecasts.

You would make the identified models available to HPF using several procedures. The HPFARIMASPEC, HPFESMSPEC and HPFUCMSPEC procedures build the specifications for different models, and the HPFSELECT procedure gathers them into a model selection list. Here, the HPFDIAGNOSE procedure attempts to find the best model without regard to known series information.

You would accomplish the forecasting itself, as well as accumulation of the time-stamped data, using the HPFENGINE procedure.

```
proc hpfengine
    repository=work.arima
    globalselection=myselect
    data=EDISTO
    outfor=outfor
    outest=outest
    back=24
    lead=48
    print=(select);
  by site_no;
  id dtstamp
    interval=hour
    accumulate=avg;
  forecast stage;
run;
```

You can see the results for forecasting river stage on the Edisto River near Givhans, SC in Figure 4. This graph shows the fit, holdout and forecast periods.

These graphs show both the holdout and the forecast period for all of the stations in the Edisto River Basin. It is clear that you can probably expect no dramatic changes within the next 48 hours so plans can be made accordingly. All that is left to do is pack up and head to the water!

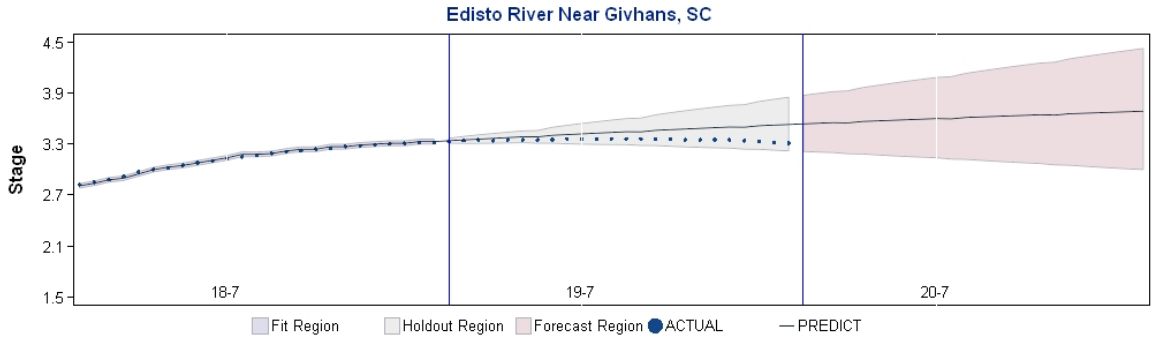

Figure 4: Edisto River Near Givhans, SC: July 18th–20th, 2007

CONCLUSION

The Edisto River is a very stable river. Observations of its river stage reveal characteristics of a regular, long-memory time series. This makes it an excellent candidate for forecasting using the SAS High-Performance Forecasting procedures. In fact, owing to the stability of the river stage, you can even generate relatively good forecasts with very little detailed knowledge of the data. However, you would be unwise to attempt a similar blind forecast of the Ashley River. The river stage of the Ashley exhibits strong tidal influences. You can obtain superior forecasts by incorporating some knowledge of tidal cycles. Thus, the statistician's exhortation to know as much as possible about the data holds true: you can leverage your knowledge to generate better forecasts.

SAS HPF provides excellent tools that can leverage detailed knowledge of the data to aid in maintaining or updating forecasts for a large-scale project when the domain of possible models is well-understood. It is also possible to use HPF in an exploratory data analysis role.

Any forecasting system subordinates model fitting because forecasts are typically judged only on holdout error performance. When you need forecasts of rivers that are strongly-influenced by man-made, geophysical or meteorological factors, a model fitting approach that incorporates such exogenous factors may be a superior approach. SAS provides many good options for model fitting as well as forecasting.

If you need an automated forecasting system relying on data from an online data repository, you can easily construct an all-SAS-based solution from data extraction to forecast generation to forecast visualization using the SAS DATA step, SAS High-Performance Forecasting and SAS/GRAPH.

REFERENCES

Box, George E.P., Gwilym M. Jenkins and Gregory C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Brocklebank, John and David A. Dickey. 2003. *SAS® for Forecasting Time Series*, 2nd ed. Cary, NC: SAS Institute Inc.

Cartier, Jeff. "The Power of the Graphics Template Language." *Proceedings of the 30th Annual SAS® Users Group International Conference*. April 2004.
<http://support.sas.com/rnd/datavisualization/papers/sugi30/GTL.pdf>
(Accessed July 18, 2007).

Croker, Samuel T. "Effective Forecast Visualization with SAS/GRAPH." *SAS Global Forum 2007 Proceedings*. April 2007.
<http:
//www8.sas.com/scholars/Proceedings/2006/DataPresentation/DP01_06.PDF>

Shumway, Robert H. and David S. Stoffer. 2006. *Time Series Analysis and Its Applications with R Examples*, 2nd ed. New York: Springer Science+Business Media, LLC.

CONTACT INFORMATION

We value and encourage your comments and questions! You can find the latest version of the SAS code for this paper at: `http://www.scoyote.net/forecasting/`. Please note that we may update this code for use in other papers.

You can contact the authors at:

Name: Samuel T. Croker
E-Mail: `scoyote at scoyote.net`
Web: `http://www.scoyote.net/forecasting/`

Name: Shane L. Hornibrook
E-Mail: `sesug_paper at shanehornibrook.com`

Name: Tomonori Ishikawa
E-Mail: `ish at alum.mit.edu`
Web: `http://www.stat.sc.edu/~ishikawa/`