Bloomberg

A Forecast for Global CO_2 Levels

Working Paper

Jan W. Dash, PhD and Yan Zhang, PhD

Quant Analytics Group

Blacki Migliozzi and Eric Roston

Bloomberg News

Revised, September 2016

Table of Contents

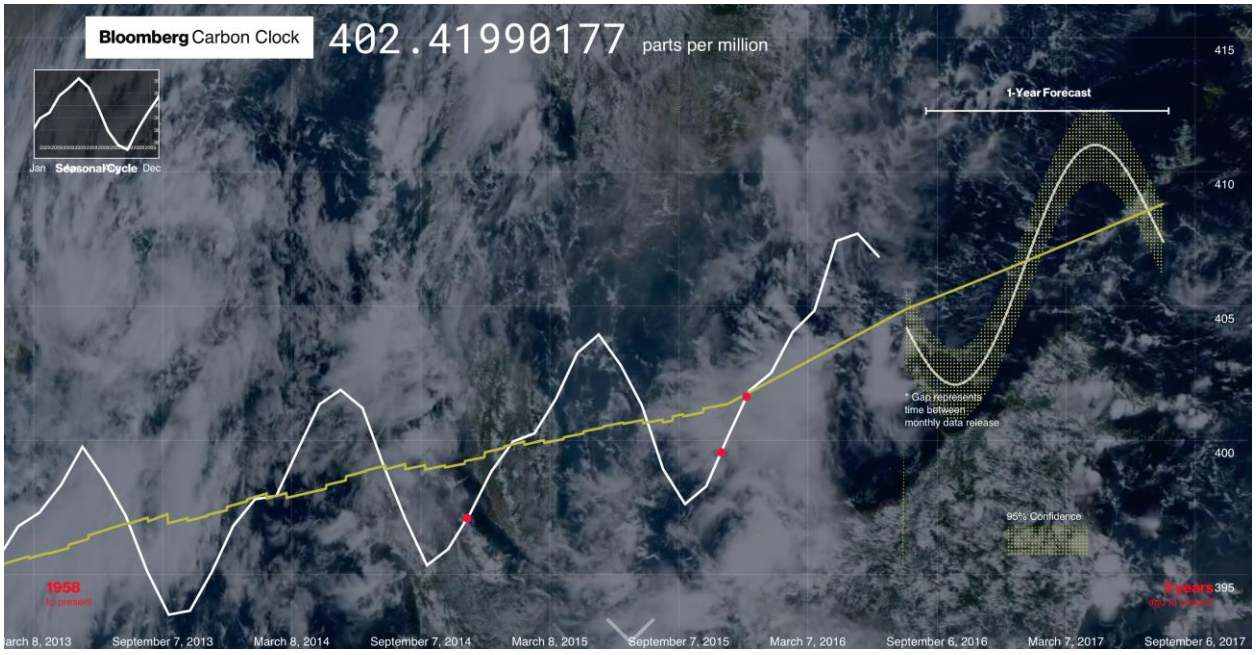
I. Abstract

The Bloomberg Carbon Clock is a web-based graphic that estimates in real time the global monthly average atmospheric concentration of carbon dioxide. This paper describes the statistical tools used to determine the Clock's values, which are extrapolated from a rolling monthly average of National Oceanic and Atmospheric Administration weekly data. We use two independent techniques: Singular Spectrum Analysis (SSA), which identifies subtle trends in data and provides missing values, and a "wavelet" function that combines the trend, seasonality, and annual variability.

II. A Different Way to Tell Time

Carbon dioxide (CO_2) is the main driver of global warming, making up more than 75 percent of climate pollution annually. Human activity is "extremely likely to have been the dominant cause of the observed warming since the mid-20th century," the Intergovernmental Panel on Climate Change concluded in its most recent assessment.[i]

The Bloomberg Carbon Clock is a graphic display that estimates in real time the global monthly average atmospheric concentration of carbon dioxide. It can be found at the website http://www.bloomberg.com/carbonclock. The real time display of the changing CO_2 is the "different way to tell time". Here is one view of the Clock taken on 9/6/16; it is described in detail in the paper:



A little bit of CO_2, just 0.04 percent, goes a long way when it comes to trapping the sun's heat. If you divide a volume of air from the atmosphere into a million parts, you can expect about 400 parts to be CO_2. That's the unit scientists use to measure its concentration—parts per million, or ppm.

Before industrialization, the level stood at approximately 280 ppm, near the upper bound for at least 800,000 years. The first Mauna Loa weekly average stood at 316 ppm on March 29, 1958. It is now beyond the 400 ppm threshold[1]. The upward trend has driven scientific and policy debates about climate change and what to do about it for a half century

Month-on-month changes in the CO_2 data are relatively insignificant compared with the long-term trend over years. The threshold "400 ppm" itself is largely symbolic. To detect *significant* differences in the climate response to CO_2, scientists and policymakers study scenarios that may differ by 50 ppm, 100 ppm, or more. For example, the real danger zone, according to peer-reviewed research, is beyond approximately 450ppm, with the likelier safe zone below 350ppm, an average level passed around 1988.[ii]

It's difficult to say with desirable precision exactly how much warming a given CO_2 level may bring. The concentration is already more than 40 percent higher than the preindustrial average. Scientists know that even more is even worse, and provide a range of estimates of how destructive climate change may become. Research summarized by the IPCC projects that a doubling of the pre-industrial CO_2 level may bring between 1.5C and 4.5C of warming. The year 2015 is likely to be the first with global average temperature a full degree Celsius above the preindustrial level. The atmosphere is expected to smash the annual global heat record in 2016 for the third consecutive year.

Scientists have watched the CO_2 level rise since late 1950s, when Charles David Keeling, a Scripps Institution geochemist, initiated monitoring at the Mauna Loa Observatory, in Hawaii, and elsewhere. Over six decades, the plotted CO_2 data—Keeling Curve—have become iconic in Earth science

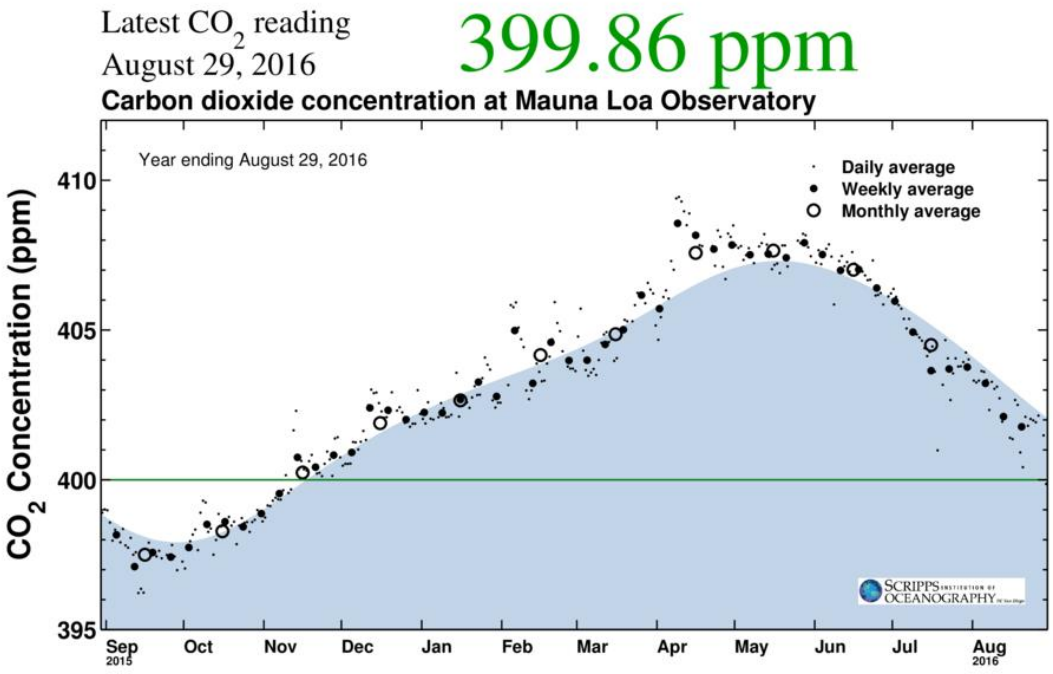
Like Scripps, the National Oceanic and Atmospheric Administration measures CO_2 at the Mauna Loa Observatory. It also collects data from a global network of observing stations, towers, flights, and flasks.

The procedure described in this paper relies on a rolling monthly average based on NOAA weekly data.

CO_2 emitted from a power plant or other source joins the atmosphere relatively quickly, its plume dissolving within a day or so, depending on winds and weather. On a larger scale, CO_2 pollution from the Northern Hemisphere, where most of it is produced, takes longer to

[1] Measurements in the Southern Hemisphere also (consistently) show CO_2 concentration over 400 ppm.

diffuse into the Southern Hemisphere. Averaging the data over time helps to eliminate local weather or observational biases and provides a closer estimate of the underlying background trend. Here is a graph of the Mauna Loa data as of August 29, 2016:



This graph from the Scripps Institution of Oceanography website[2] shows how CO_2 averages fluctuate depending on the observational averaging time frame. Daily and weekly averages contain "noise" that make the trend look less clear. The picture comes into sharper focus when looking at monthly and annual data.

The Bloomberg Carbon Clock is projected from the average of the four most recent NOAA weekly estimates, and therefore may be slightly lower or higher than other measures at any given moment.

III. Bloomberg Carbon Clock: How It Works

1. Two Techniques – SSA and the Wavelet

Scientists rely on statistical tools that ensure the data they study are as valid and complete as possible. When a time-series has gaps, or ends prematurely, geophysicists and others typically apply any of several mathematical techniques that can interpolate the gaps and extrapolate past the end of data.

[2] See https://scripps.ucsd.edu/programs/keelingcurve/

We use two independent techniques to project CO_2 levels for the Clock.[3]

Singular Spectrum Analysis[iii] (SSA) is a statistical technique that identifies subtle trends in data and estimates any missing values. Scientists use SSA and similar tools to project a trend past the end of a time series (a "cliff") or interpolate data missing from the time series (a "hole").[4] Holes exist because some data are not reported between two existing data points. A "cliff" exists either because no data are reported past a given date in the past up to now, or because we want to forecast the data into the future. Our procedure[5] is consistent with the global SSA analysis of historical CO_2 data performed by Dettinger and Ghil in 1998,[iv] although that study did not forecast CO_2.

The second statistical technique is a formula derived from the Mauna Loa data itself. We call this function a "wavelet" for its superficial resemblance to a wave, and functions known elsewhere in mathematics. The wavelet is made up of three terms:

- *Trend line*: A "LOESS" quadratic function in time captures the slope of the CO_2 trend month by month, which has risen on average every year since monitoring began in 1958.
- *Seasonal change*: The second term is a sine wave, which mimics the seasonal ups and downs of data.
- *Variability*: The third term, a modulating function, multiplies the sine wave. It modulates the amplitude and positions of CO_2's seasonal peaks and dips.

2. Short-Term CO_2 Forecasting: Pedagogical Example

This section describes how we forecast the CO_2 level and apply an algorithm that reduces noise, or randomness, that comes mostly from observational sampling effects.[6]

CO_2 is measured continuously. Its daily, weekly, monthly, and annual averages tend not to match because of short-term variation—noise—that masks the observational CO_2 signal. Weather conditions around the Mauna Loa Observatory push the short-term averages randomly above or below the longer term average values.

---

[3] Acknowledgments: We thank Gavin Schmidt for helpful suggestions motivating our choice of "wavelet" phenomenological functions and for suggesting the long-term forecast simulation test. We also thank Michael Ghil for helpful conversations.

[4] SSA and its multivariate extension MSSA is being used at Bloomberg LP to improve data quality by filling holes for missing data and removing erroneous data spikes in time series. SSA/MSSA is a common geophysics technique (see refs.)

[5] Some other standard forecast methods (ARIMA, Prony analysis, wavelets without SSA) were also examined. Tests indicate that the SSA-wavelet procedure makes more reliable projections.

[6] An introduction to SSA is in Appendix IV.

NOAA's Earth System Research Laboratory publishes CO_2 data weekly.[7] We average the four most recent weekly updates to arrive at single monthly updates. Hence, one forecast month in the model may not correspond to a calendar month.

3. How The Model Works[8]

1. Training the model: The SSA forecast starts with a fitting procedure. Suppose that we are generating the projections for January.[9] The four weekly data points in December are averaged to a monthly data point, which is assigned to Dec. 15. Three years of previous monthly data are used for obtaining the wavelet function by a standard least-squares fitting routine. The data are approximated using the smoothing function $f^{(\text{fit})}(t)$ —the quadratic trend term plus the sine wave term. We then apply a modulating Gaussian, or bell-shaped, function $f_G(t)$, determined separately for each half year cycle, so that the tops and bottoms of the seasonal curves better resemble the last three years of data. That produces $f^{(\text{modulated})}(t)$, the modulated wavelet function.

2. Using the Model: The three years of monthly data described in (1) are interpolated into daily estimates of the modulated wavelet function, with one change. The forecast wavelet is obtained by extrapolating the last historical wavelet without modulation. This tentative forecast is successively modified by the SSA algorithm until successive steps of the SSA algorithm become smaller than a convergence tolerance specified in advance. The SSA algorithm, which is described in detail in appendix IV, generates the forecast for January. Linear interpolation from the forecast provides the second-by-second readings displayed on the Clock.

3. The date moves to Feb. 1 and the steps above are repeated. The algorithm updates monthly. On Feb. 1, the original forecast is redone, using the January average. The new data point also allows us to refine or update the January prediction. The differences between the forecast and updates (measured historically) allows us to provide a statistical measure of uncertainty for a given forecast in advance.

---

[7] See http://www.esrl.noaa.gov/gmd/ccgg/trends/weekly.html

[8] In practice, we update the Clock weekly, not monthly, as in this example. For the Clock's accompanying graphic display, we produce an annual update.

[9] The numerical results in this paper approximate the monthly average CO_2 for a month with the average of the four most recent weekly data points.

4. Formulas for the Wavelet

The above description defines $f^{(\text{modulated})}(t)$, the final function for describing the data using our smoothing procedure, where *t* is the time. We call this the "wavelet". We use daily interpolation of these functions. Here are the formulas for the components of the wavelet:

- Quadratic, or trend line function: $f_Q(t) = A + Bt + Ct^2$
- Sine wave, or seasonal variation function: $f_S(t) = M \sin\left[(2\pi + \nu)t + \phi\right]$
- Modulating Gaussian function, which compensates for variability in annual peaks and troughs in the data: $f_G(t) = \left(2\pi\sigma_G^2\right)^{-1/2} \exp\left[-\frac{\left(t - t^{(\text{Max data})}\right)^2}{2\sigma_G^2}\right]$

The monthly fit $f^{(\text{fit})}(t)$ and modulated function $f^{(\text{modulated})}(t)$ are defined as[10]

$$f^{(\text{fit})}(t) = f_Q(t) + f_S(t) \tag{1.1}$$

$$f^{(\text{modulated})}(t) = f_Q(t) + f_S(t)\, f_G(t) \tag{1.2}$$

We can refine the model with a second sine wave (the "leaf term") added to the seasonal variation function $f_S(t)$. The leaf term accounts for an extra observed increase of CO_2 in autumn[11]. Defining $\Psi = (2\pi + \nu)t + \phi$, the new definition of $f_S(t)$ is[12],

$$f_S(t) = M_1 \sin(\Psi) + M_2 \sin(2\Psi) \tag{1.3}$$

[10] The normalization is determined by matching the maximum (or minimum) data value and position in time using the two parameters in the Gaussian function.

[11] We think it is physically reasonable that this "leaf" term corresponds to leaves falling off trees and the decay of flora in the Northern Hemisphere in autumn. The results we obtain do not depend on this interpretation.

[12] At the end of the paper, we give some numerical results using this refined model with the leaf term.

IV. Numerical Results and Figures

1. A decade of data

Figure 1 shows weekly NOAA CO_2 data in ppm from Mauna Loa for 2005-2015.

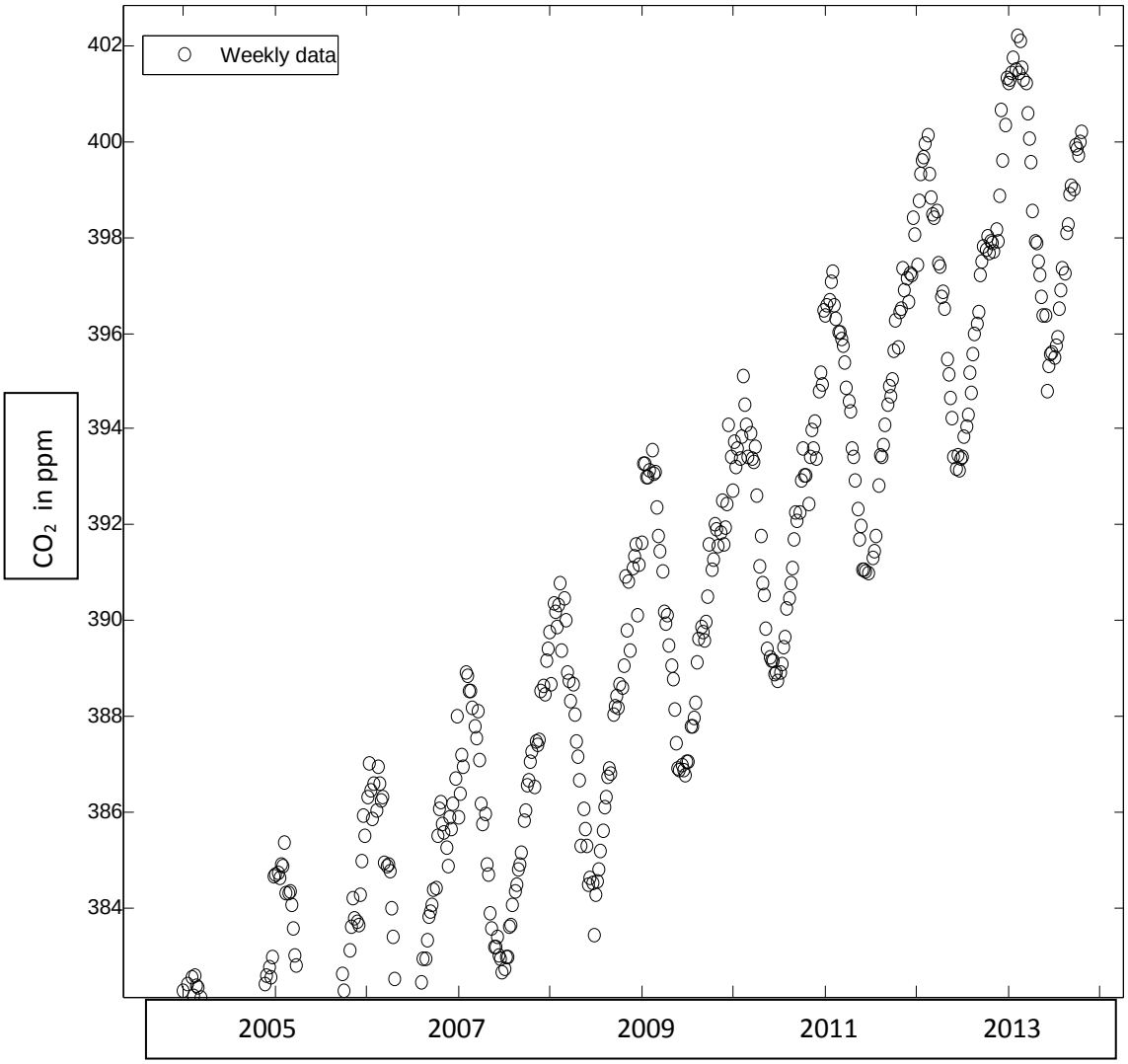


Figure 1: Weekly NOAA CO_2 data in ppm from Mauna Loa

The CO_2 concentration has risen by roughly 85 ppm, or 27 percent, since 1958. In the last 10 years, it has risen from about 376 ppm past 400 ppm. We average the most recent four weekly data points to generate a monthly average, a standard practice in research that reduces noise.

2. Forecast trend, extrapolated from last three years

The model "learns" the CO_2 trend by looking at three years of data. The one-month projection in Figure 2 is shown in red (at the far right end).

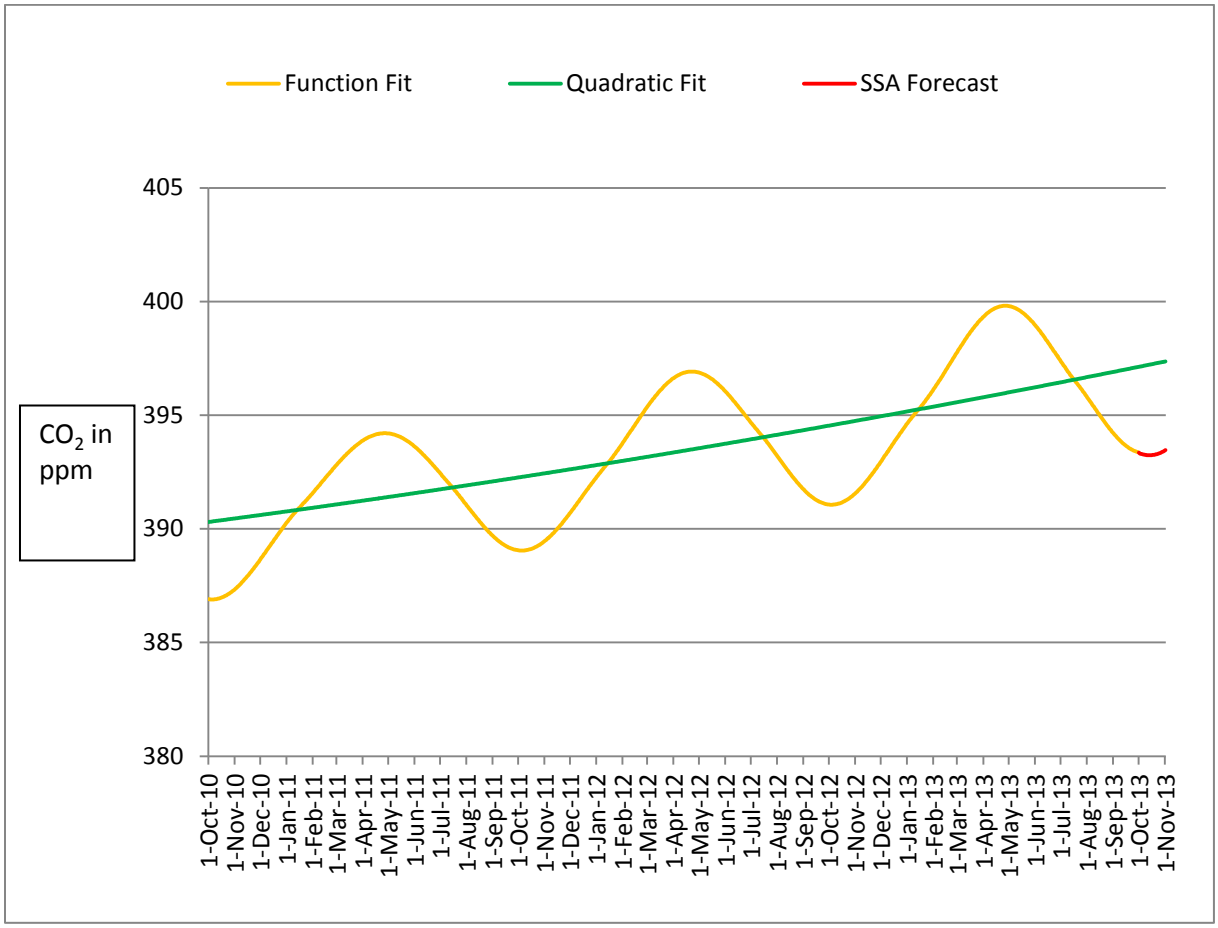


Figure 2: One SSA forecast and one function fit—the input to SSA—using previous 3 years of data. Note that this is one fit and not a concatenation of monthly updates. The forecast (**Red line at right end**) was made on Oct. 1, 2013 for the October period Oct. 1-31. The **Yellow oscillating line** shows the function fit up to Oct. 1. The relatively straight **Green line** shows the quadratic fit, or trend line, up to Oct. 31.

3. A one-month forecast

Figure 3 below shows the one-month forecast (red line at the far right of the oscillating curve), including the updating in the procedure as described above:

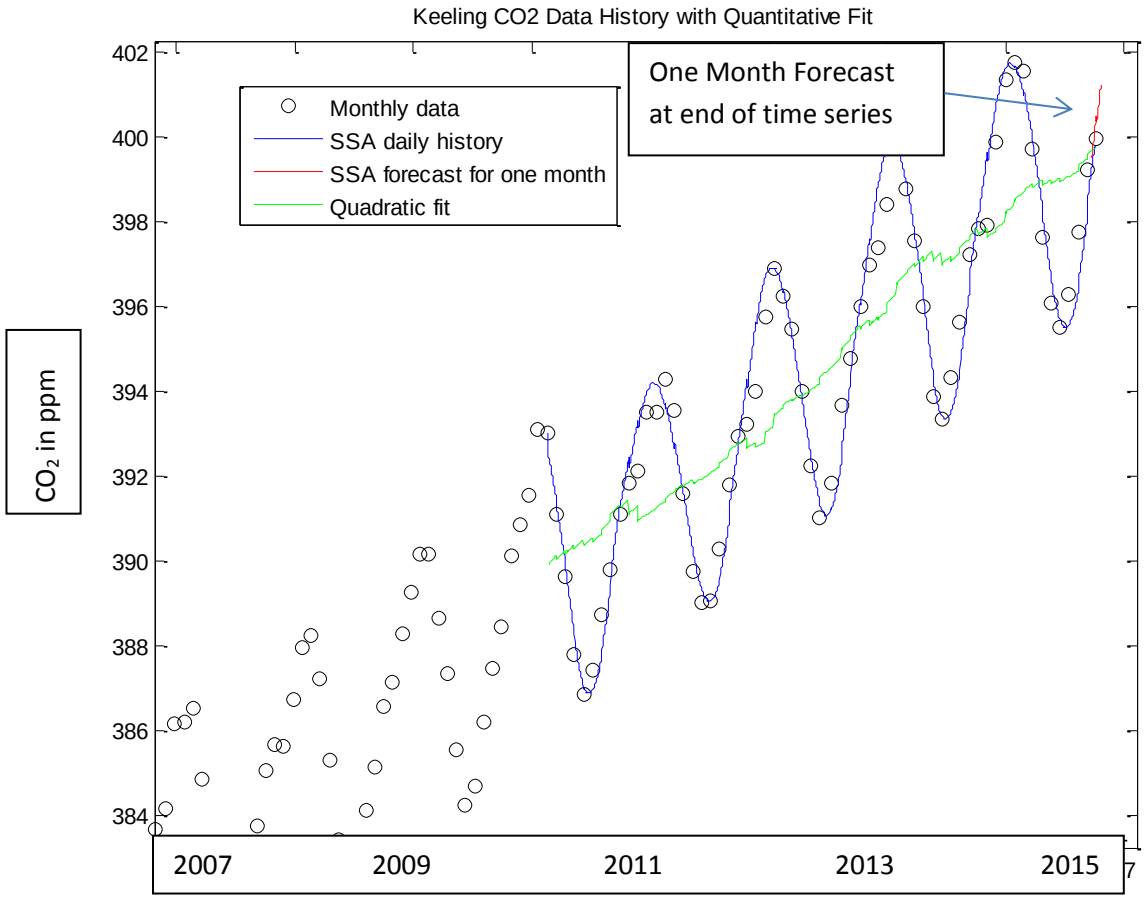


Figure 3: SSA five-year update and one-month forecast. The short **Red line** far to the right in the graph shows SSA one-month forecast (statistically this is "out of sample"). The **Blue line** is the SSA update with new information included. The **Green line** is the quadratic function trend from the fit, updated monthly. Open circles are monthly CO_2 data.

4. Checking the forecast with a back test

The data serve as the ultimate check on the Clock's forecast. Each month's forecast can also be updated with a new SSA fit once that month's data are in[13]. This provides a backtest of the forecast. The results are shown in Figure 4. .

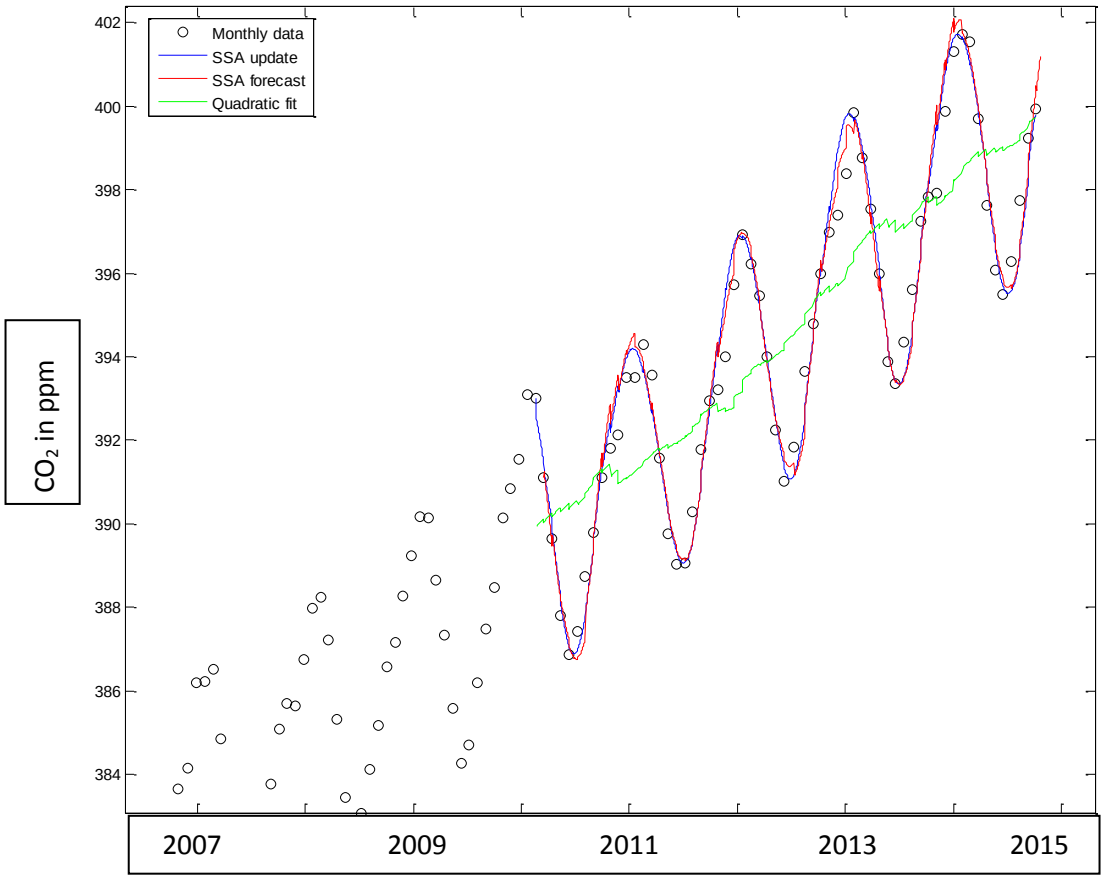


Figure 4: SSA five-year forecast and back test result. **Red line**: SSA forecast (statistically "out of sample"). **Blue Line**: SSA update with new information. **Green line**: quadratic fit to trend. Open circles are monthly data, each obtained by averaging weekly data.

Every time a new data point comes in, we re-run the SSA algorithm. That's the blue line, an update that checks the original forecast. The green line in Figure 3 is the main trend, the quadratic function, with fluctuations that reflect monthly variability[14].

[13] The forecast takes place in a "cliff" period that contains no data. The updated information has a new data point at the end of the cliff, turning the cliff into a "hole". An SSA updated fit is done with this extra information, and the result is compared with the SSA forecast over the same (cliff/hole) time period.

Figure 4 shows that the forecasts are quite robust. When the SSA algorithm is re-run, or backtested, with each new data point, the update closely matches the original forecast. The red line shows the original forecast, made on the first day of every month. The newest forecast is visible in the top-right of the graph, beyond the last data point.

5. Comparison of the updated SSA and wavelet curves

In Figure 5 we show the wavelet with modulation, compared to the final SSA updated curve. Each month we run the two methods, including the last data point (there is no forecasting here). Both methods produce daily interpolation. The results are similar. This is a robustness result since we get similar results with either method.

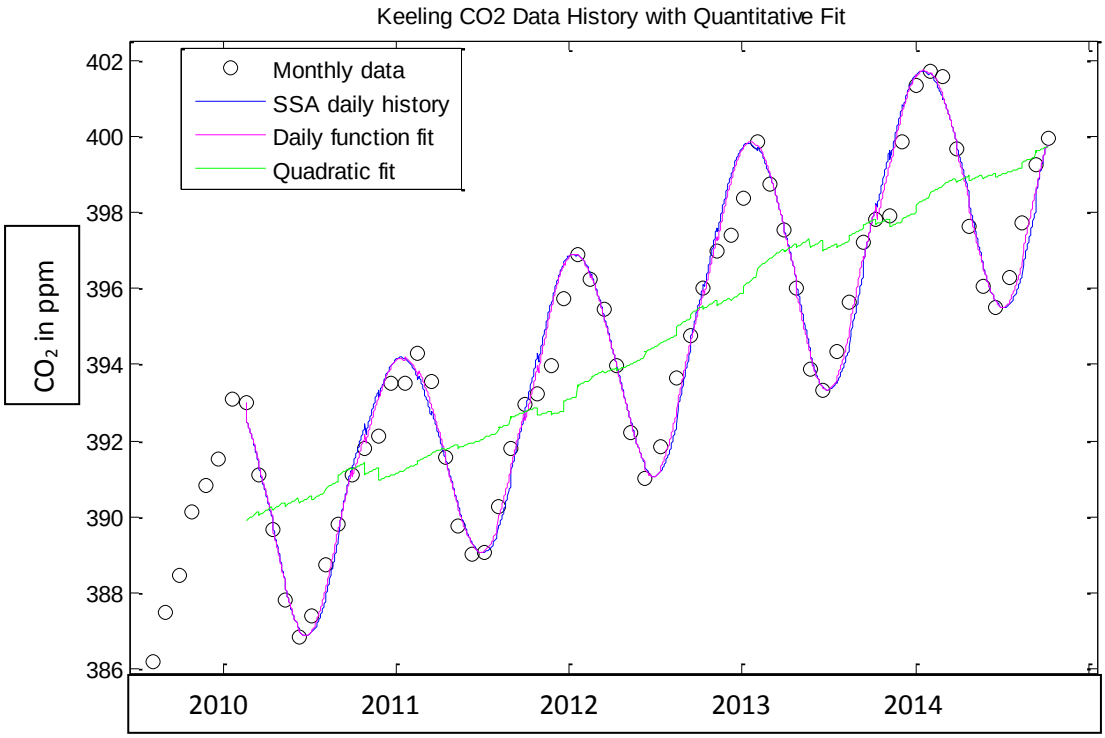


Figure 5: SSA 5-year update and function fit**. Pink line:** wavelet function fit. **BLUE line**: SSA update. **Green line**: quadratic fit to trend. Open circles are monthly data.

[14] Note for quadratic fit components: A given fit has 36 points. However for a given month, the graph only shows the quadratic component that is obtained on the date just after the new data point is received after this given month. That is, previously updated quadratic fit components are not re-updated.

6. Forecast Procedure Illustration

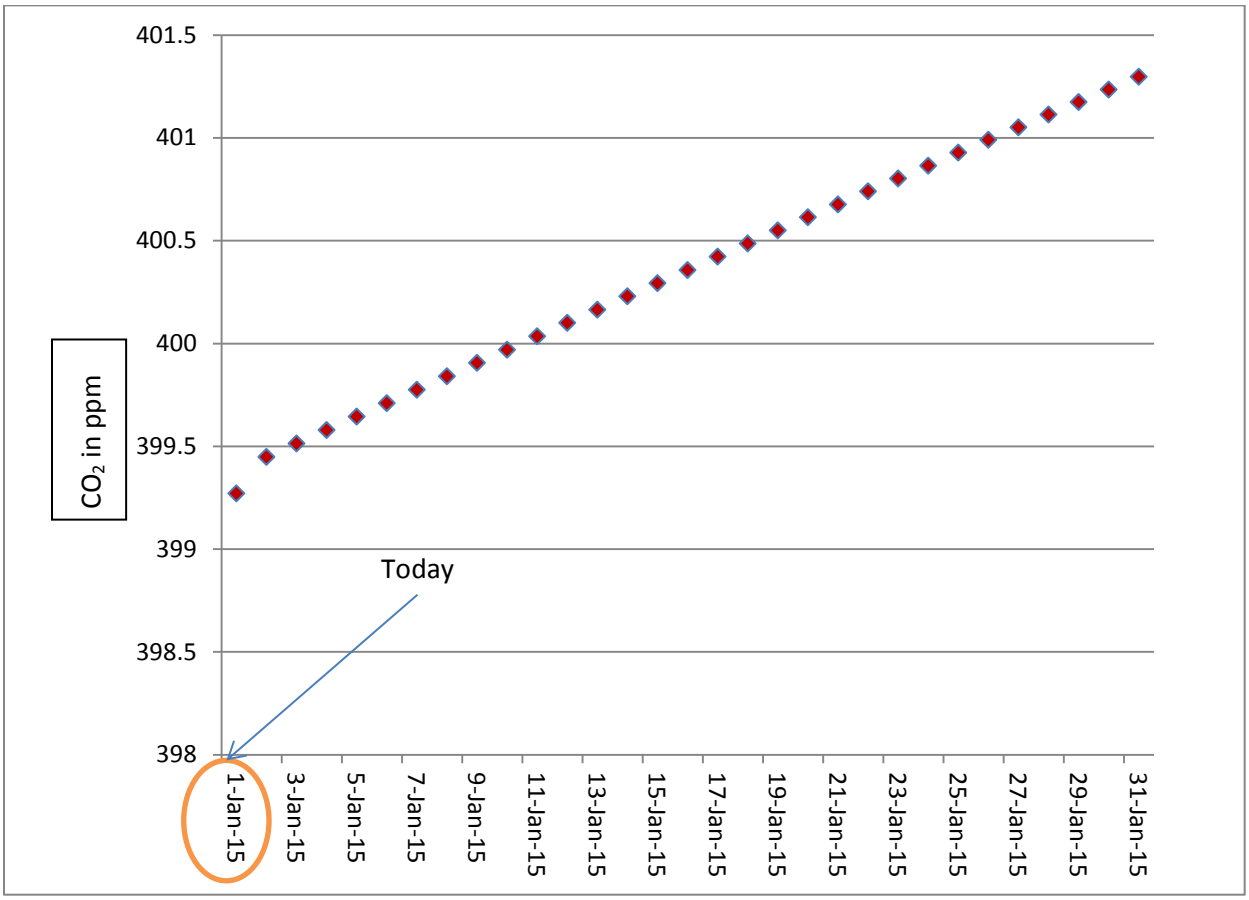


Figure 6: Illustration of the forecast procedure. The SSA forecast is made on Jan. 1, 2015, when the December monthly data point became available, for the period Jan. 1-31.

7. The trickiest time of the year

The CO_2 curve bends down in the spring and up in the fall. That's when projections may be most likely to diverge from the data. The curve below shows summer yielding to fall, and, consequently, extra CO_2 returning to the atmosphere. The figure below gives an example. Note that the scale is magnified relative to the other graphs.

By the end of the month, the forecast lags the updated reading by about 0.1 or 0.2 parts per million.

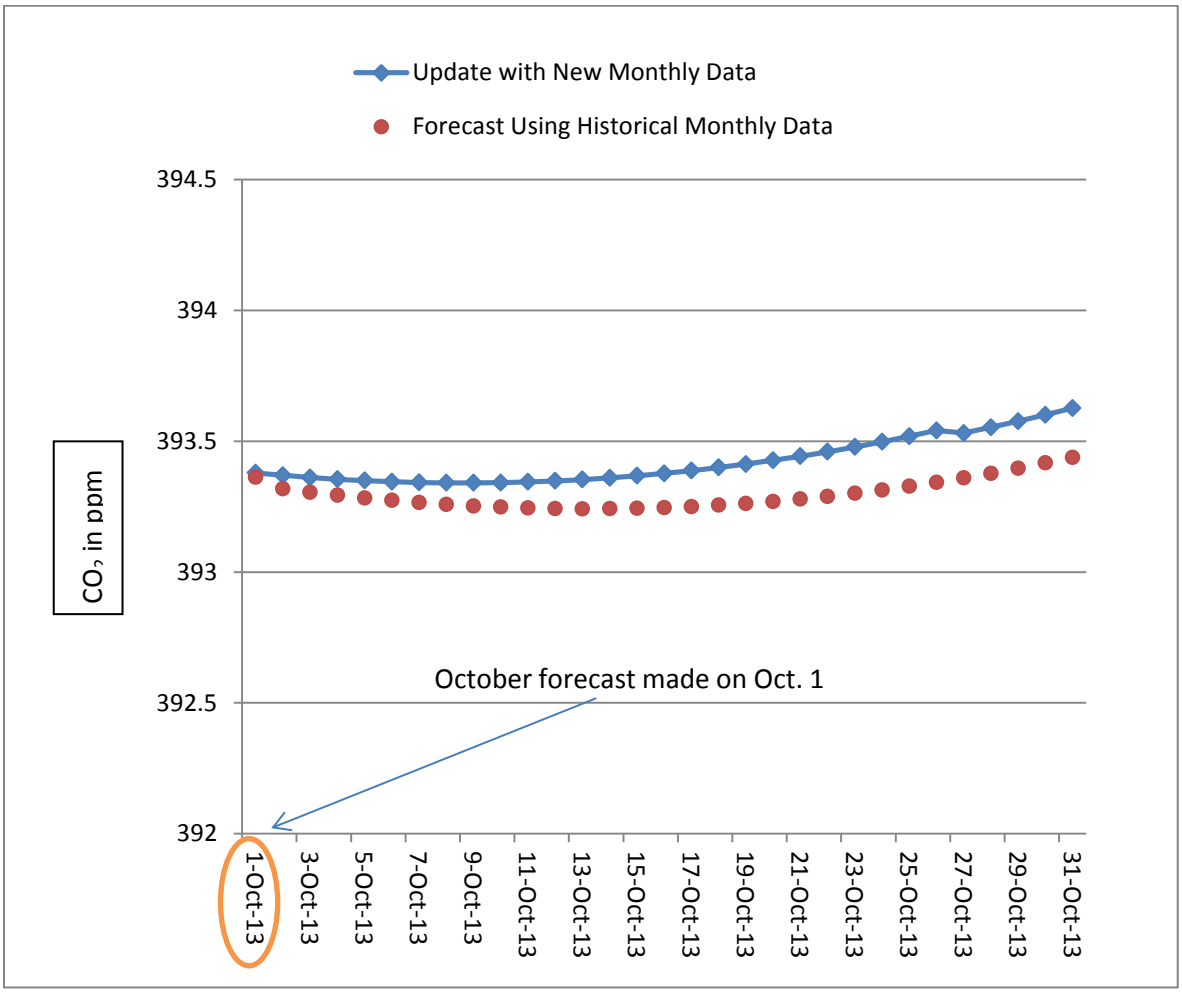


Figure 7: The forecast and update hit a turning point. The forecast (**red dots**) was made on Oct. 1, 2013 for the October period Oct. 1-31. The last monthly data point before the forecast was on Sept. 28. The forecast was subsequently updated when a new data point became available, resulting in the **blue line**.

8. 30-Year Description and the Cycle Standard Deviation

Figure 8 is an application of the model for over 30 years. The results show that the model operates in a stable fashion over rather long time periods.

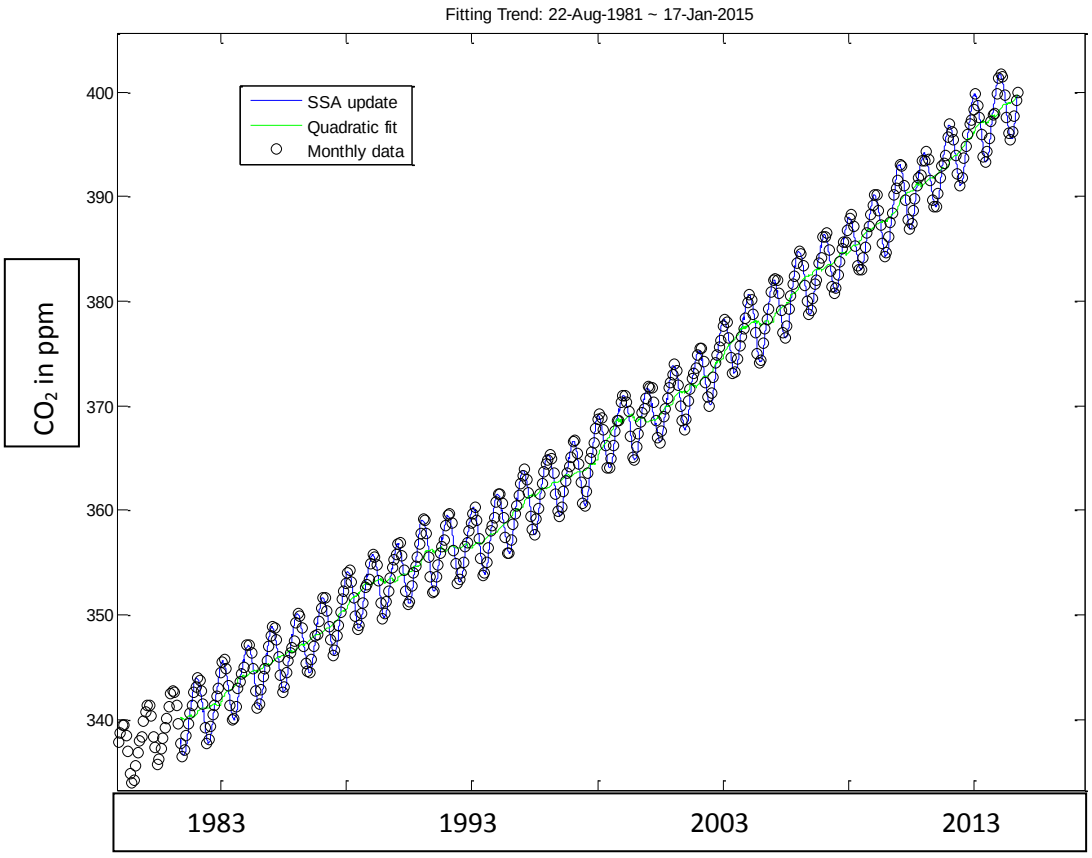


Figure 8: **Blue Line**: SSA monthly updates with new information, closely following the open circles (monthly data). **Green line**: the model fitted quadratic trend over 1981-2015, updated monthly.

Difference between the forecast and updates for 30-year simulation

The difference between the forecast and updates in the Blue Line in Figure 8 for the 30-year simulation can be characterized by the "cycle standard deviation" $\sigma^{(\text{Cycle})}$; it has the value

$$\sigma^{(\text{Cycle})} = 0.69\,ppm \tag{1.4}$$

The cycle standard deviation is used to measure the uncertainty in the forecast cycle, and is used for this purpose in the Bloomberg Carbon Clock.

Trend behavior over 30 years

Since the quadratic fitted trend is adjusted every month with new data, the trend reveals more than CO_2 level's long-term upward behavior. It also shows occasional flat or even locally downward behaviors over short time periods.

9. Two Definitions of Trend and Uncertainty Quantification

One of the most important features of the CO_2 data is the generally upward trend. Our methodology allows a sophisticated treatment for three reasons: 1. We can isolate the trend from the seasonal oscillations, 2. We have two ways to measure the trend, and 3. Our methodology has updating that gives a time-dependence to the trend, providing a rich structure.

Consider the first SSA reconstruction component denoted as $R_k(t)$ with $k=1$ (see Appendix IV). This $R_1(t)$ constitutes the SSA-specified trend.

Hence we have two different reasonable candidates for the CO_2 trend: (1) the successive monthly wavelet quadratic forecasts $f_Q^{\text{Monthly Adjusted}}(t)$, and (2) the monthly-adjusted first SSA reconstruction component $R_1^{\text{Monthly Adjusted}}(t)$. We can combine these by taking the average. The absolute difference $D(t)=\left|f_Q^{\text{Monthly Adjusted}}(t)-R_1^{\text{Long Term Average}}(t)\right|$ gives a measure of the uncertainty in the trends. Figure 9 has the results:

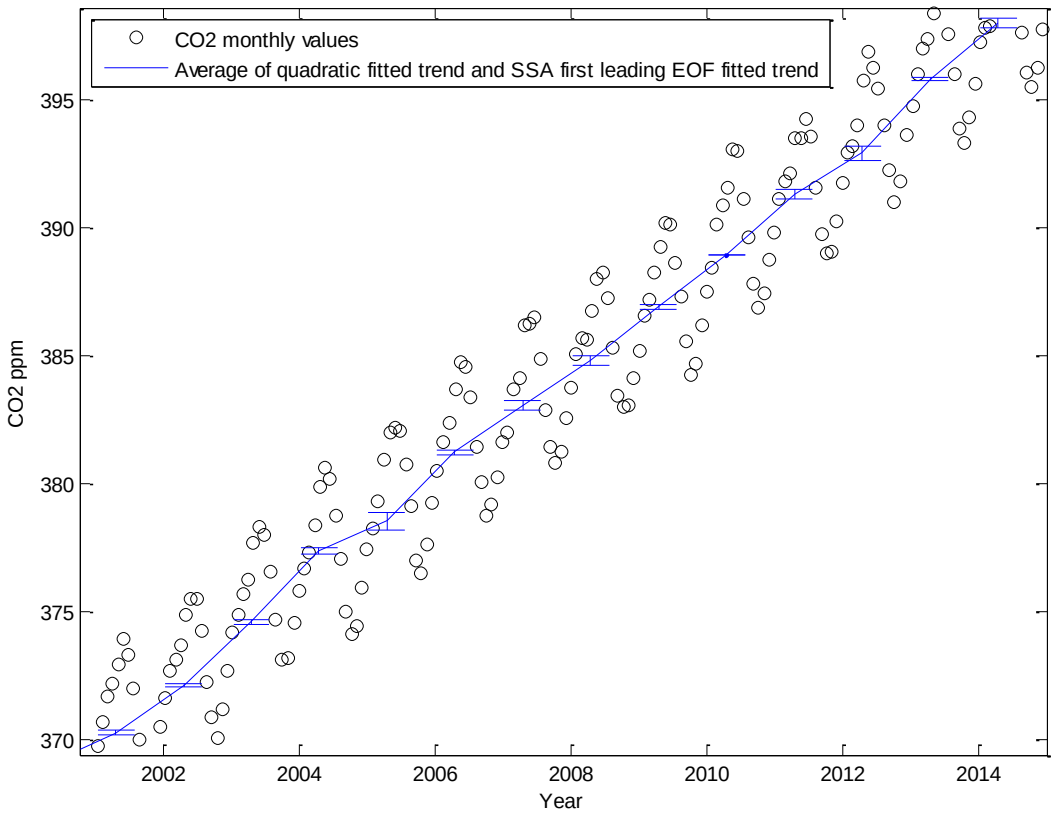


Figure 9: The average (**blue** line) of (1) the monthly-updated wavelet quadratic fitted trends and (2) the SSA monthly-updated first reconstruction component trends. The error bars illustrated by the I symbols are the absolute differences between the trends.

Forecast Trend Uncertainty

The forecast trend uncertainty is defined in the same way using the absolute difference between (1) the successive monthly wavelet quadratic forecasts $f_Q^{\text{Monthly Adjusted}}(t)$, and (2) the monthly-adjusted first SSA reconstruction component $R_1^{\text{Monthly Adjusted}}(t)$ starting at the forecast beginning time point. This forecast trend uncertainty is in the final Bloomberg Carbon Clock.

10. Comparison of monthly updated trends

Next we consider the trends both with monthly adjustments, viz $f_Q^{\text{Monthly Adjusted}}(t)$ and $R_1^{\text{Monthly Adjusted}}(t)$. These two monthly adjusted trends are close, as shown in Figure 10.

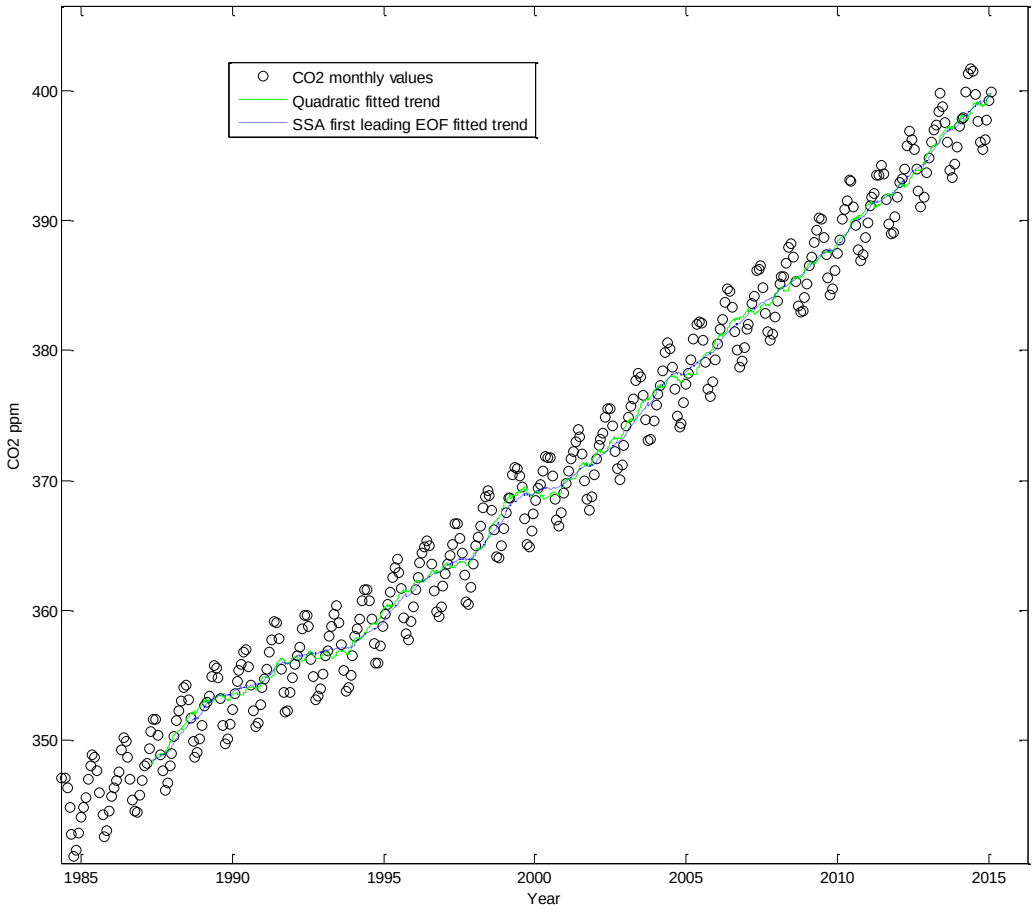


Figure 10: The model fitted monthly-adjusted trends, over 30 years (1985-2015). **Green line**: monthly-adjusted quadratic fit to trend. Open circles are monthly data. **Blue dotted line**: Monthly-adjusted trend derived from SSA first reconstruction component.

11. The differences of monthly quadratic and SSA trends

Next we consider the trend differences both with monthly adjustments, viz $f_Q^{\text{Monthly Adjusted}}(t) - R_1^{\text{Monthly Adjusted}}(t)$. Figure 13 is the histogram of two monthly adjusted trend differences and the graph of a normal distribution $N(0.07, 0.35)\text{ppm}$. Again this means that the monthly-updated wavelet-defined trend and the monthly-updated SSA-defined trend are statistically indistinguishable, and so produce essentially equivalent local-trend definitions. Note that the difference with updates has a smaller standard deviation than the difference without monthly adjustments in Figure 11.

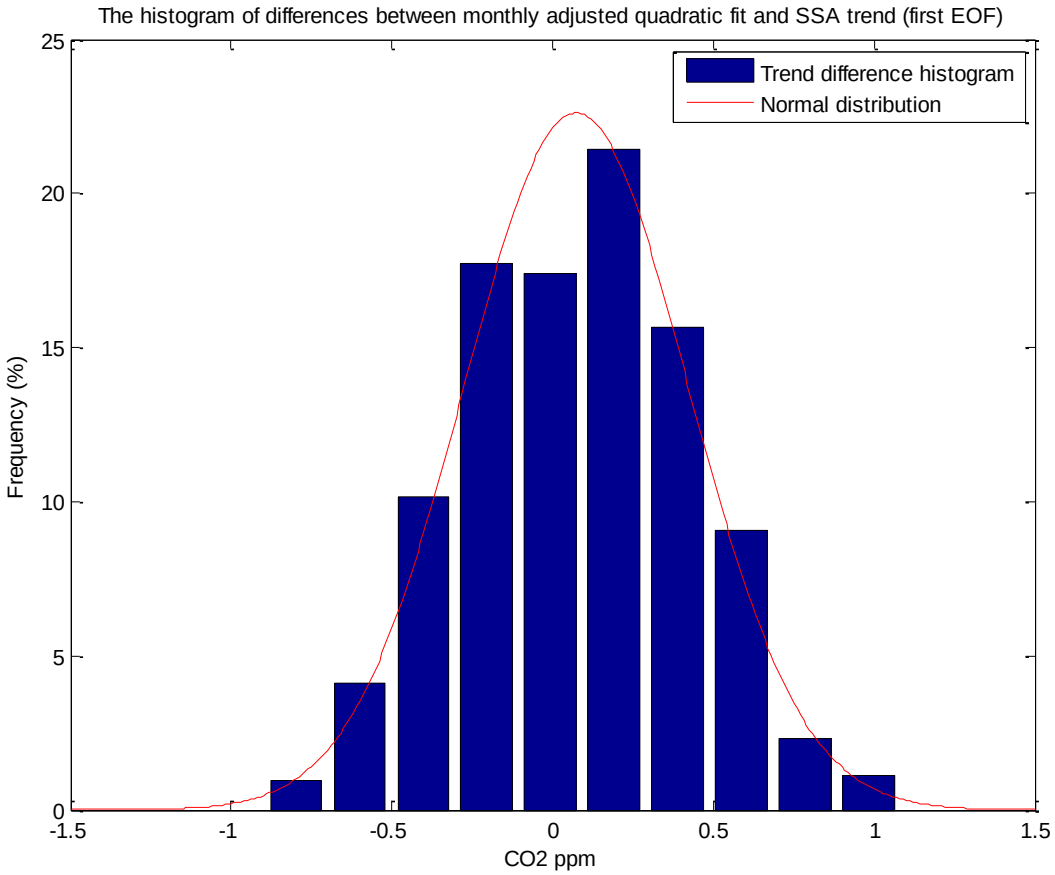


Figure 11: The distribution of the difference of monthly trends: monthly-adjusted quadratic fitted trends minus the monthly-updated first reconstruction component of SSA (not the long-term average). This difference is pure noise as shown by the good fit of the Gaussian $N(0.07, 0.35)\text{ppm}$ to the histogram.

12. The long-term trend from SSA

We next compute $R_1^{\text{Long Term Average}}(t)$ for a long term (30 years) average, and compare it to the wavelet monthly adjusted trend $f_Q^{\text{Monthly Adjusted}}(t)$. The monthly-adjusted wavelet trends $f_Q^{\text{Monthly Adjusted}}(t)$ fluctuate around the long term average trend $R_1^{\text{Long Term Average}}(t)$. The results matching the ascent of the overall trend up to these fluctuations, showing stability of the results.

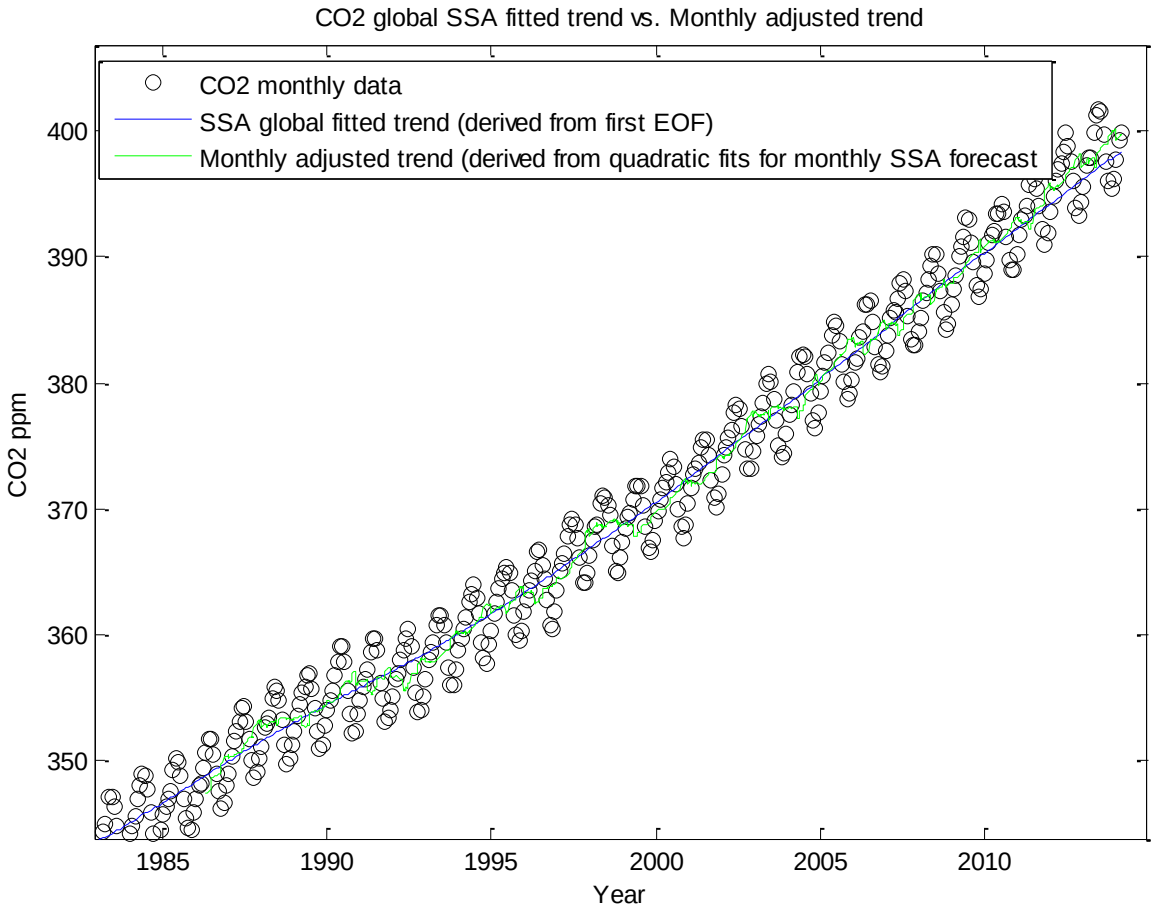


Figure 12: The Model fitted trend for 30 years (1985-2015). **Blue Line**: long-term average trend derived from the SSA first reconstruction component (or EOF – see Appendix IV). **Green line**: Monthly adjusted quadratic fit to trend. Open circles are monthly data.

13. Difference between trends: long-term SSA and wavelet quadratic

Figure 13 is the histogram of the differences $f_Q^{\text{Monthly Adjusted}}(t) - R_1^{\text{Long Term Average}}(t)$ between the two trends. It shows how frequently, and by what margin, the SSA trend deviates from the quadratic wavelet trend. The results closely follow a normal distribution N(0.08, 0.69) with mean = 0.08 ppm and standard deviation = 0.69 ppm. This is similar to the pattern of random noise, again showing that the two methods are (up to noise) reasonably equivalent.

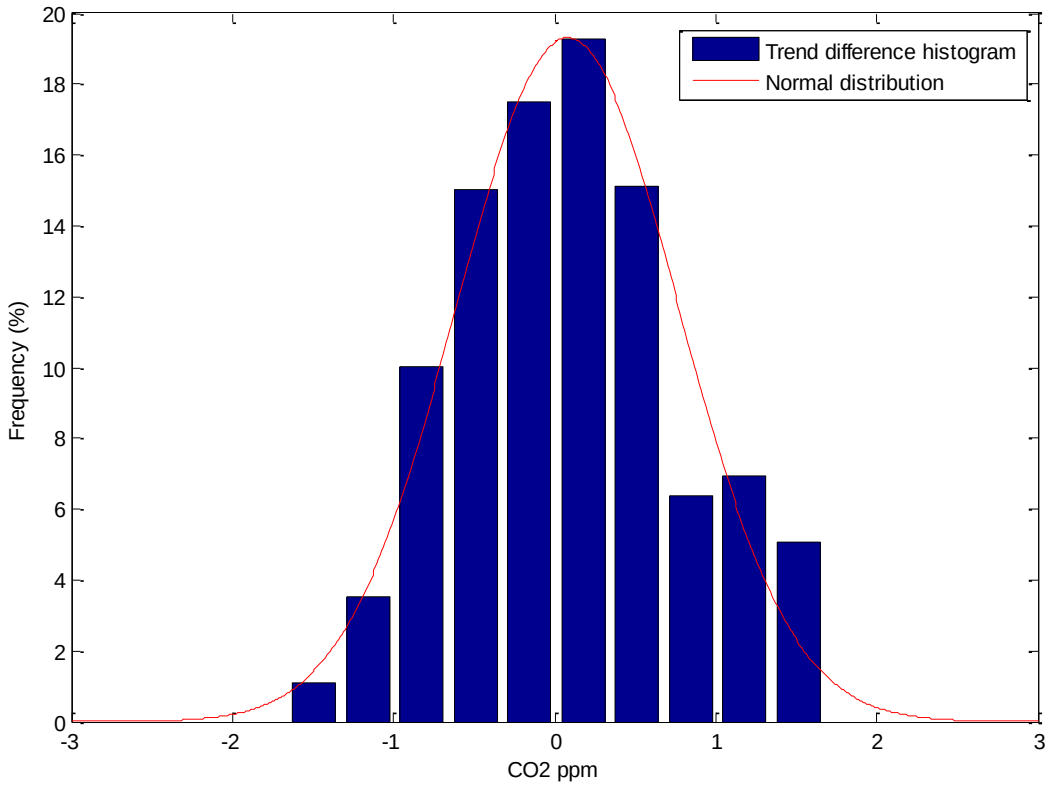


Figure 13: The distribution of the difference of the long-term SSA trend and the quadratic wavelet trend. The histogram is the monthly adjusted quadratic fitted trends minus the long-term average (not adjusted monthly) from the first reconstruction component of SSA. This difference is close to pure noise as shown by the reasonable fit of the Gaussian $N(0.08, 0.69)\text{ppm}$ to the histogram.

V. The Model Refined with a "Leaf Term"

Here we consider the extra "leaf term" in the wavelet described in Section II above. There is a slight extra increase of CO_2 in the autumn months when leaves fall in the northern hemisphere. This term accounts for this CO_2 miniburst. (Note the southern hemisphere has less land and the trees in the Amazon rainforest are green year-round).

For comparison, Figure 14 has the results without the leaf term.

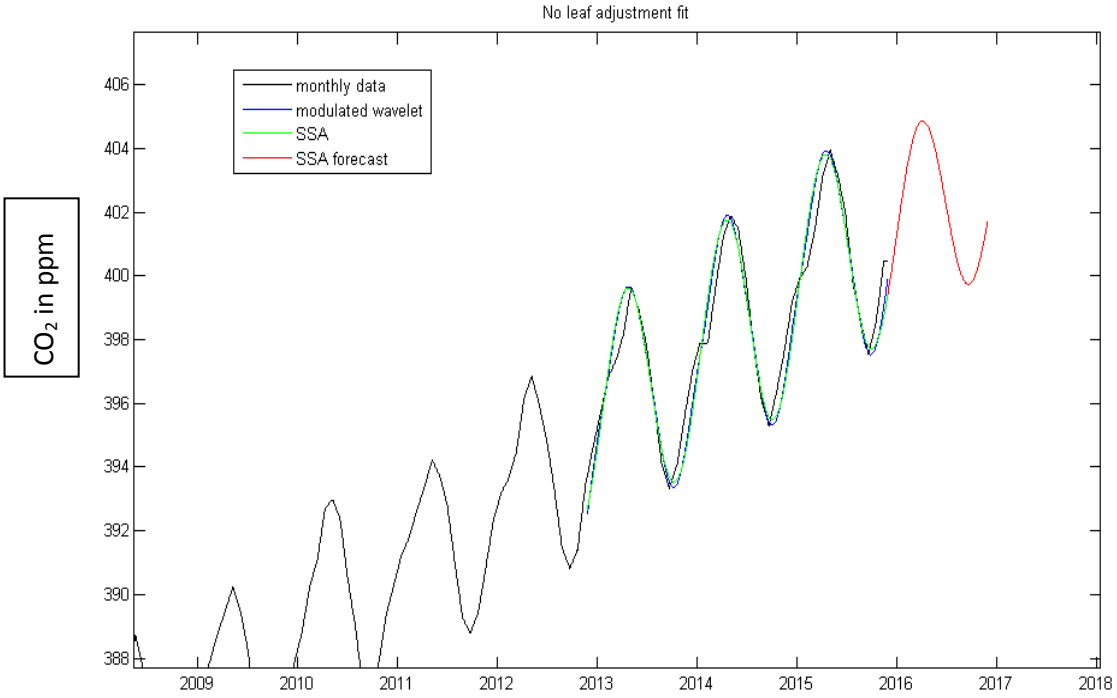


Figure 14: Wavelet (**blue**) and SSA (**green**) fit to historical data (**black**), without the extra "leaf term". Note that the result is a bit low during fall and high during winter, each year. The forecast is in **red**.

Figure 15 has the results with the leaf term. The description is more accurate with the leaf term.

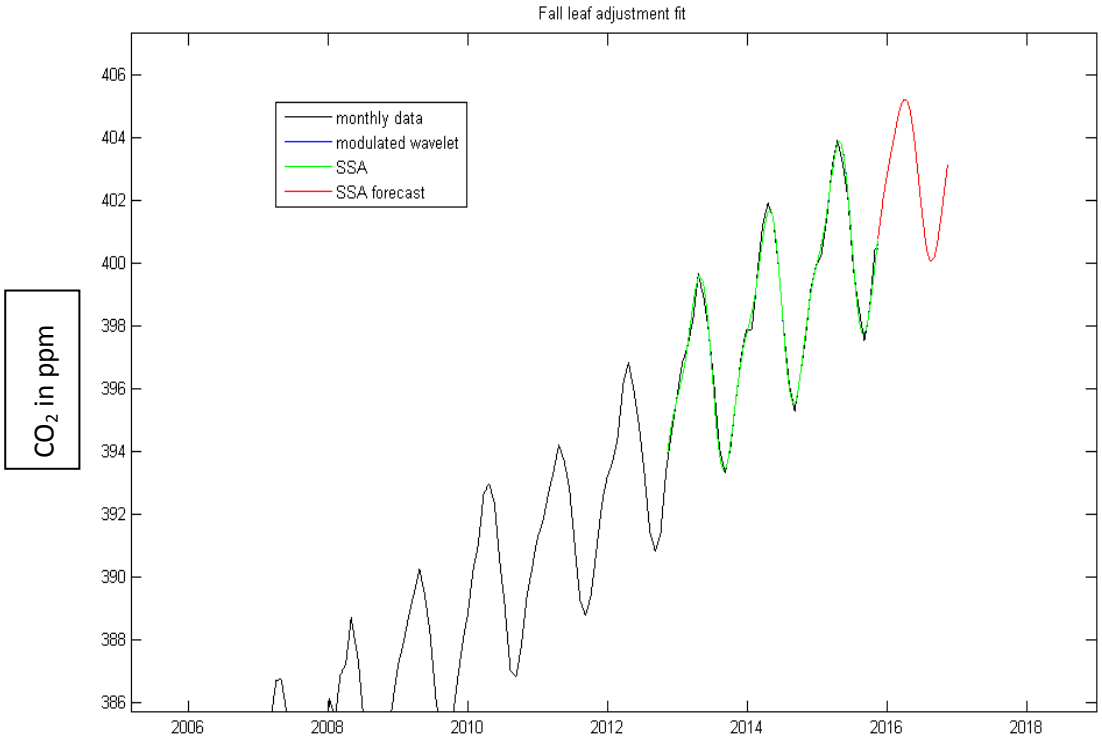


Figure 15: Wavelet with the extra "leaf term" (**blue**), SSA (**green**) fit to data (**black**). The forecast is in red. The fit is better during fall and winter each year than without the leaf term.

VI. Final Result – The Bloomberg Carbon Clock

Figure 16 shows the final model result for the forecast (red) including the forecast uncertainties given by the cycle standard deviation (Section IV.8). The figure shows SSA fit using the wavelet with the leaf term included (**green**), the historical trend (**pink**), and the forecast trend (**blue**) terms. The monthly data with rolling dates are also shown (circles).

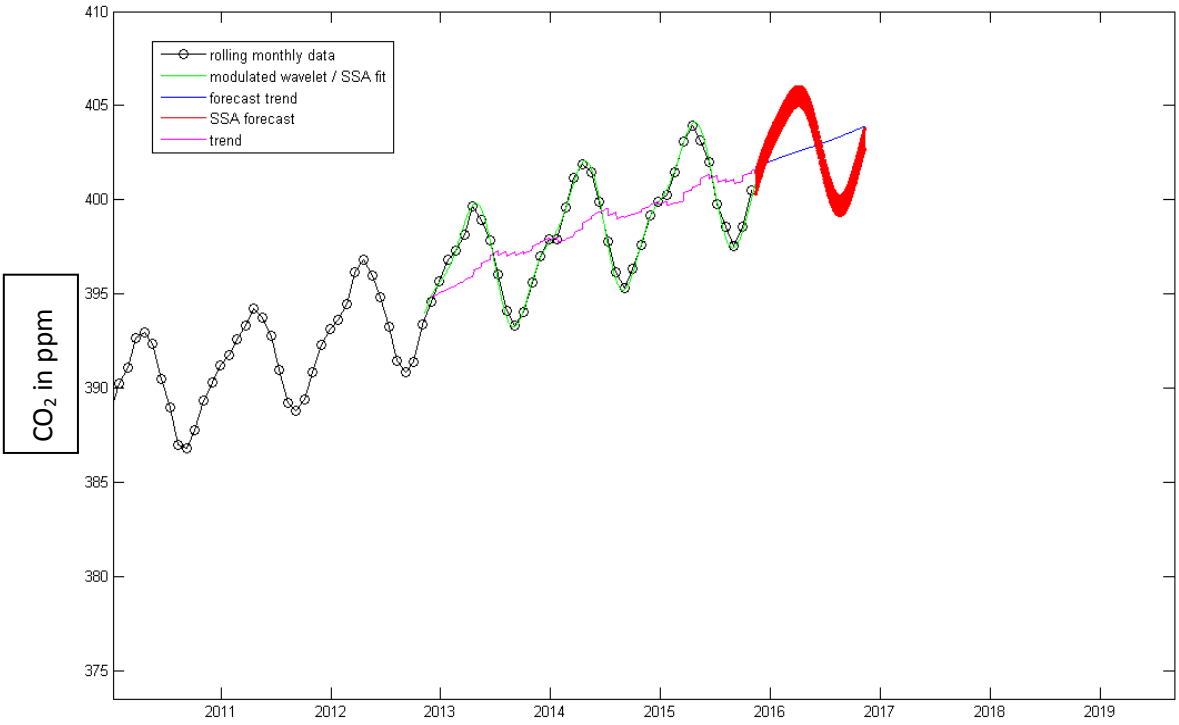


Figure 16: The final result for the one-year SSA forecast in **red**, historical trend (**pink**), the forecast trend (**blue**), and monthly data (circles).

Figure 17a below shows a similar view of the Bloomberg Carbon Clock on the website[15] taken on 9/6/16, at http://www.bloomberg.com/carbonclock .

[15] The leaf term refinement is planned but is not yet deployed on the website as of September 2016.

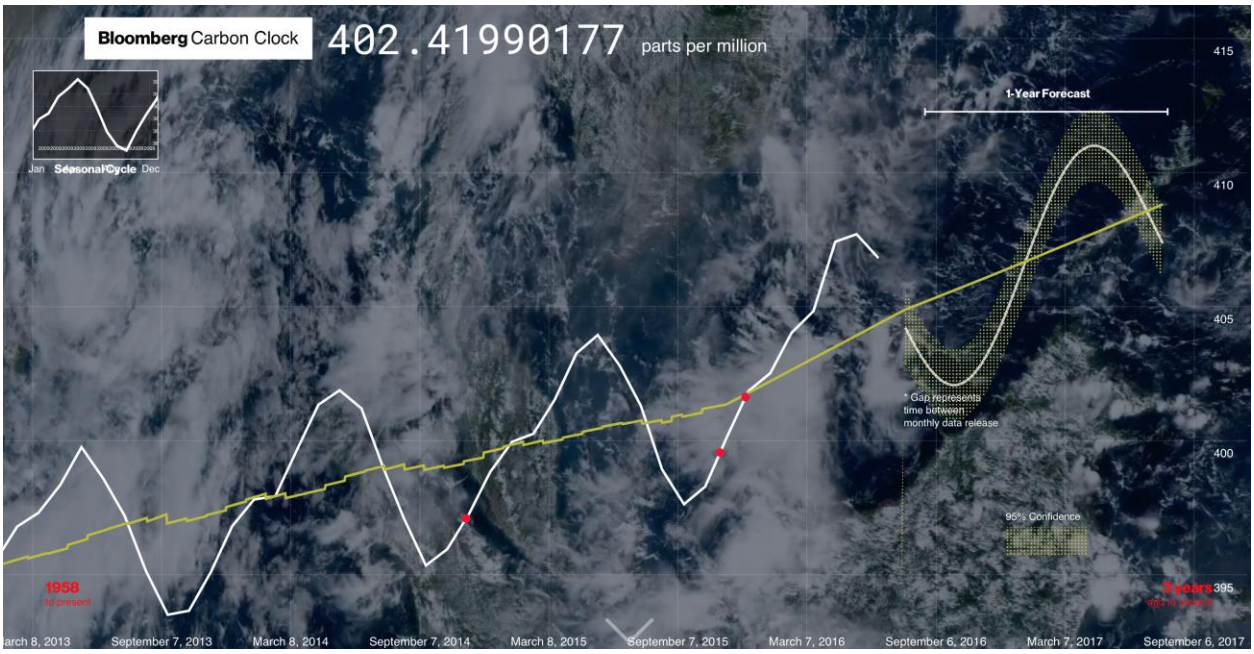


Figure 17a: The Bloomberg Carbon Clock showing the 1-year projected SSA model forecast including uncertainties, along with the historical and projected trends (**yellow**). The **red dots** are events (e.g. the last **red dot** is at the Paris UNFCCC Climate Conference, 2015). The white line segments are drawn between the historical monthly-averaged data at fixed times[16]. The Carbon Clock URL is http://www.bloomberg.com/carbonclock

The alternative view, including the uncertainties on the trend described in Section IV.9, is shown in Figure 17b:

[16] In Figures 17a, 17b, for convenience in plotting, the monthly data dates are fixed in advanced, and straight lines are drawn between neighboring monthly averaged data at these fixed times. The last segment stops at the last data point. Therefore there is a visual gap in Figures 17a, 17b between the last monthly fixed-time data point and the start of the model forecast. This gap is not material for the model. The model uses rolling monthly dates for averaging as described in the text, and the model forecast is continuous with the model historical result, without any gap. See Figure 16.

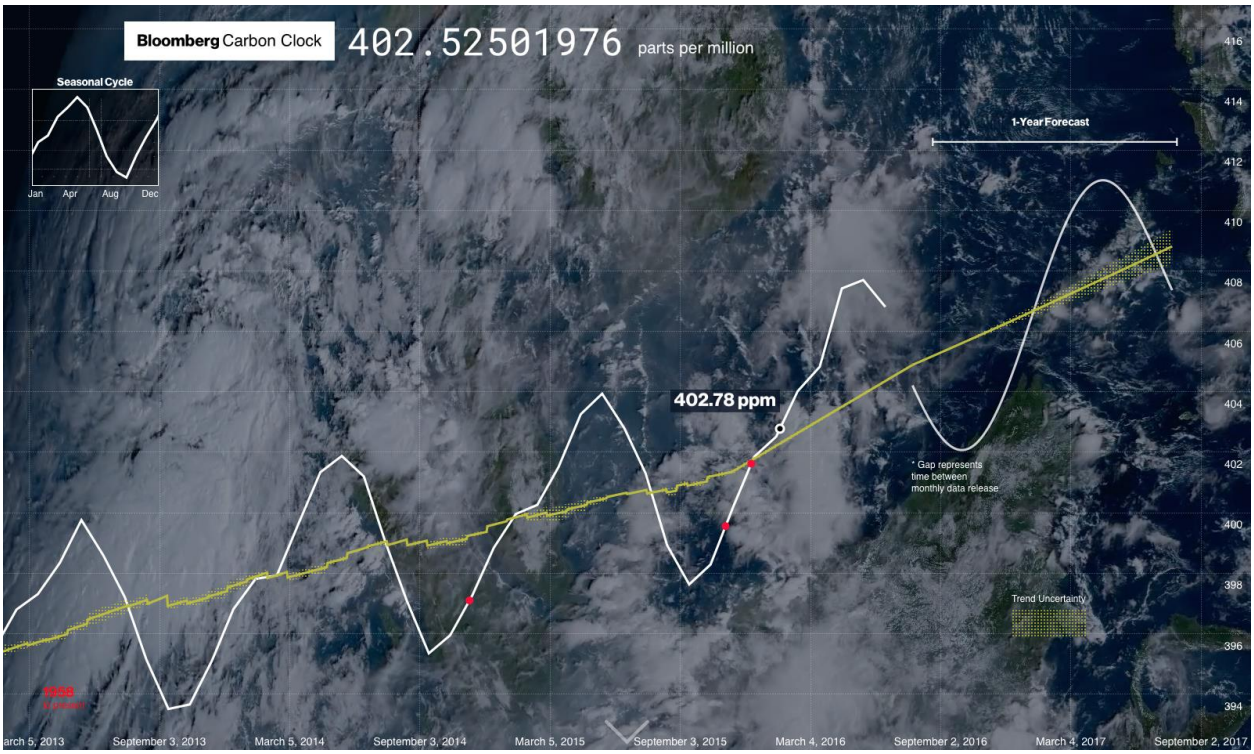


Figure 17b: The Bloomberg Carbon Clock showing the 1-year projected SSA model forecast, along with the historical and projected trends (**yellow**), with trend uncertainties. The **red dots** are events (e.g. the last **red dot** is at the Paris UNFCCC Climate Conference, 2015). The white line segments are drawn between the historical monthly-averaged data at fixed times. The Carbon Clock URL is http://www.bloomberg.com/carbonclock

Appendix I. Yearly forecasts

The procedure for yearly forecasting is the same as for monthly forecasting, except that the SSA-based extrapolation to future time is projected for one year.

Figure A1 has the yearly forecast past the last data point, joined to the previous monthly simulation using data:

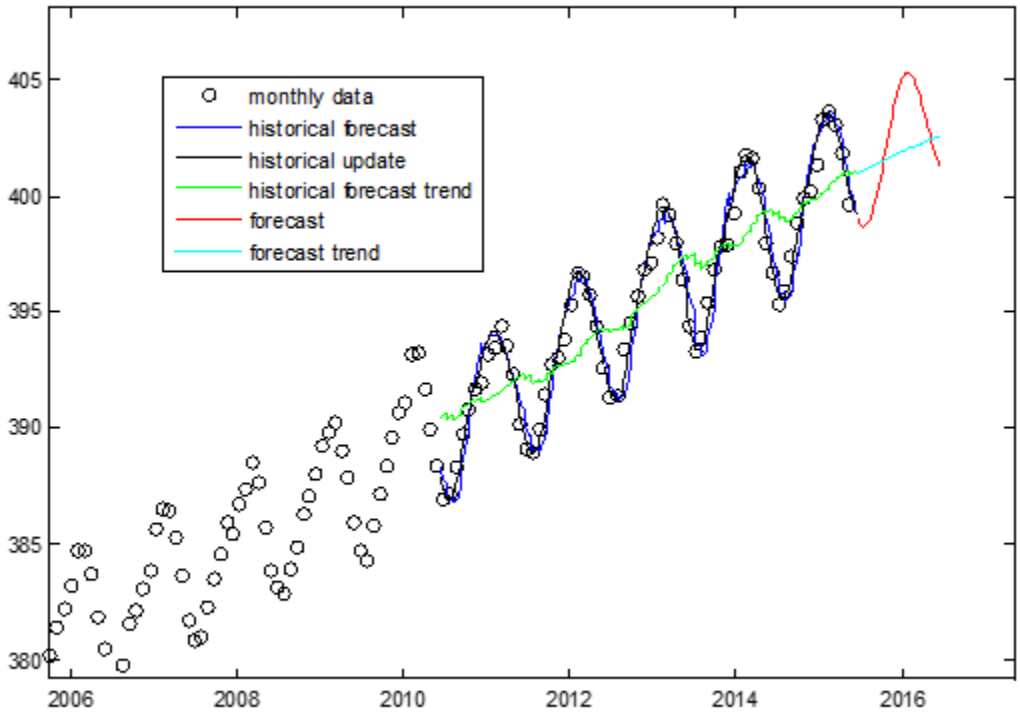


Figure A1: Red Curve: One-year forecast SSA forecast, at far right. Light Blue Line: One year forecast trend, at far right. **Blue Curve**: SSA monthly forecasts from simulation (statistically "out of sample"). **Black Curve**: SSA monthly updates with new information. **Green line**: monthly forecast quadratic trend. Open circles are monthly data.

Figure A2 shows the yearly forecast past the last data point, joined onto a yearly simulation (not monthly) using data:

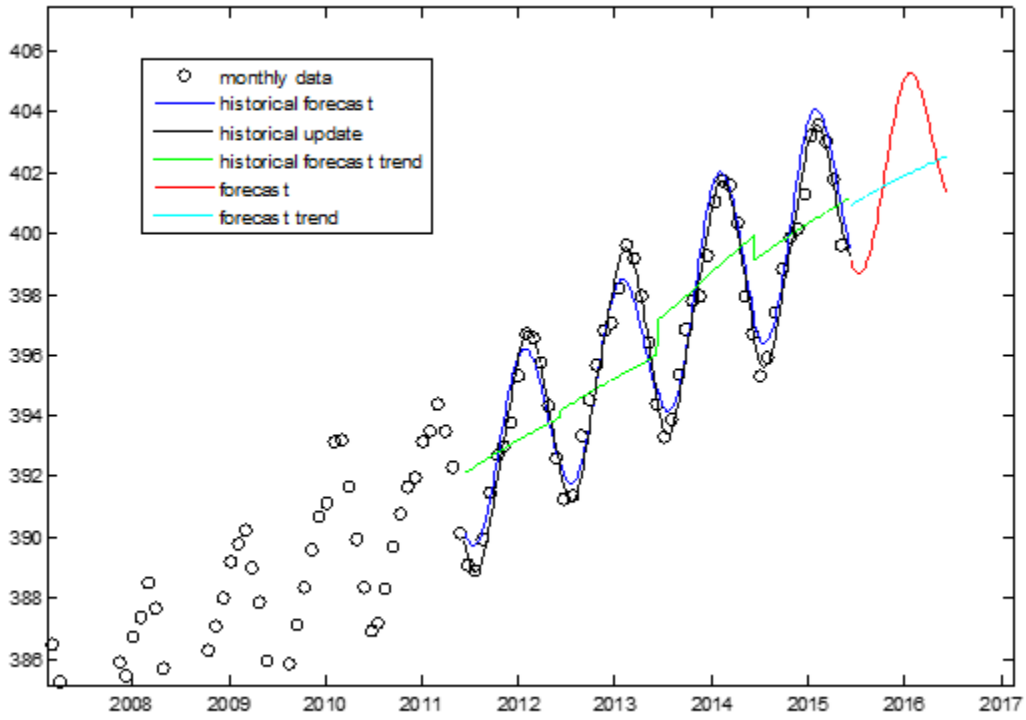


Figure A2: Red Curve: SSA one year forecast, at far right. Light Blue Line: One year forecast trend, at far right. **Blue Curve**: SSA yearly forecast (statistically "out of sample"). **Black Curve**: yearly update with new information. **Green line**: yearly forecast quadratic trend. Open circles are monthly data.

Figure A3 is the histogram of yearly differences in trends (forecast – update) from the yearly simulation.

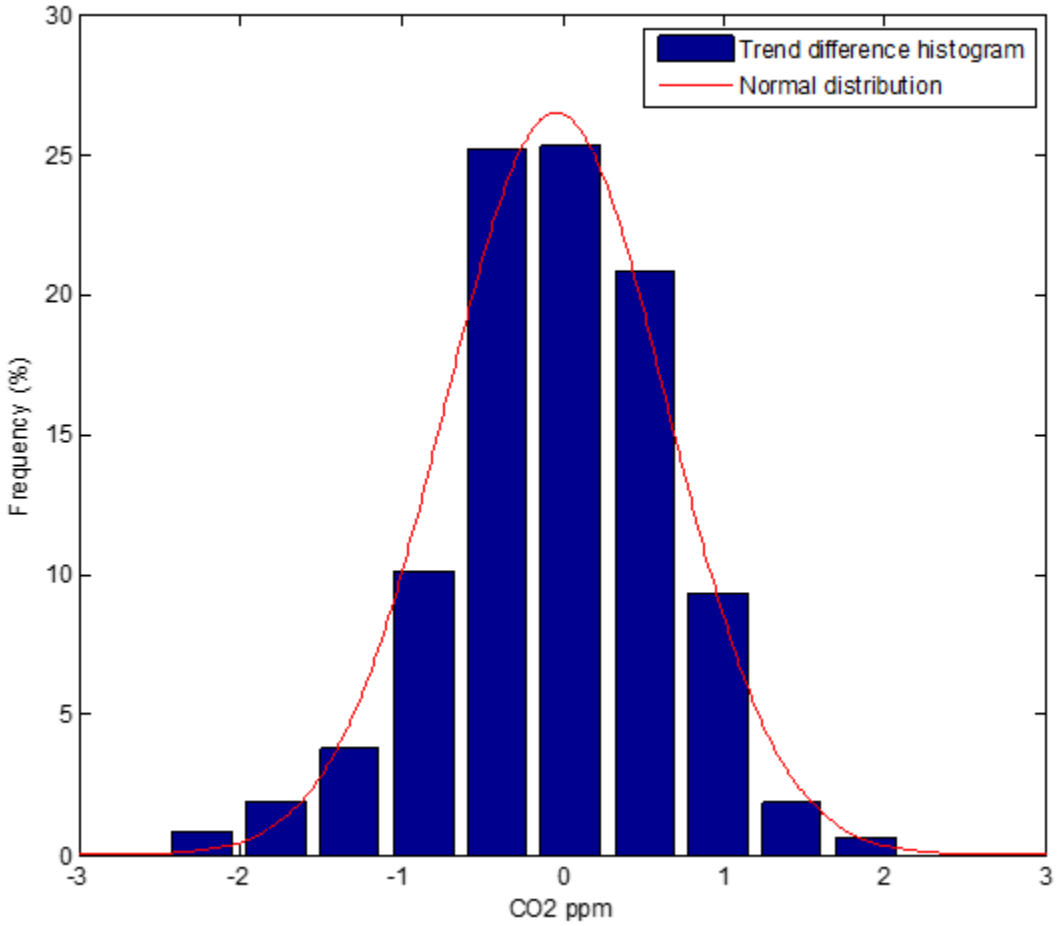


Figure A3: Histogram of yearly trend differences (forecast minus update). The red curve is a Gaussian fit (mean = -0.04 and standard deviation = 0.69). This fit shows that the difference is statistically indistinguishable from noise, so the forecasts are reliable.

Figure A4 is a pedagogical illustration of the procedure using the wavelets as input to the SSA algorithm using yearly timescales. There are 3 years of wavelets with modulation in the region where data exist, so that the amplitudes and positions of the peaks and valleys are known. There is one year of forecasting using the unmodulated wavelet, because without data the modulation cannot be predetermined.

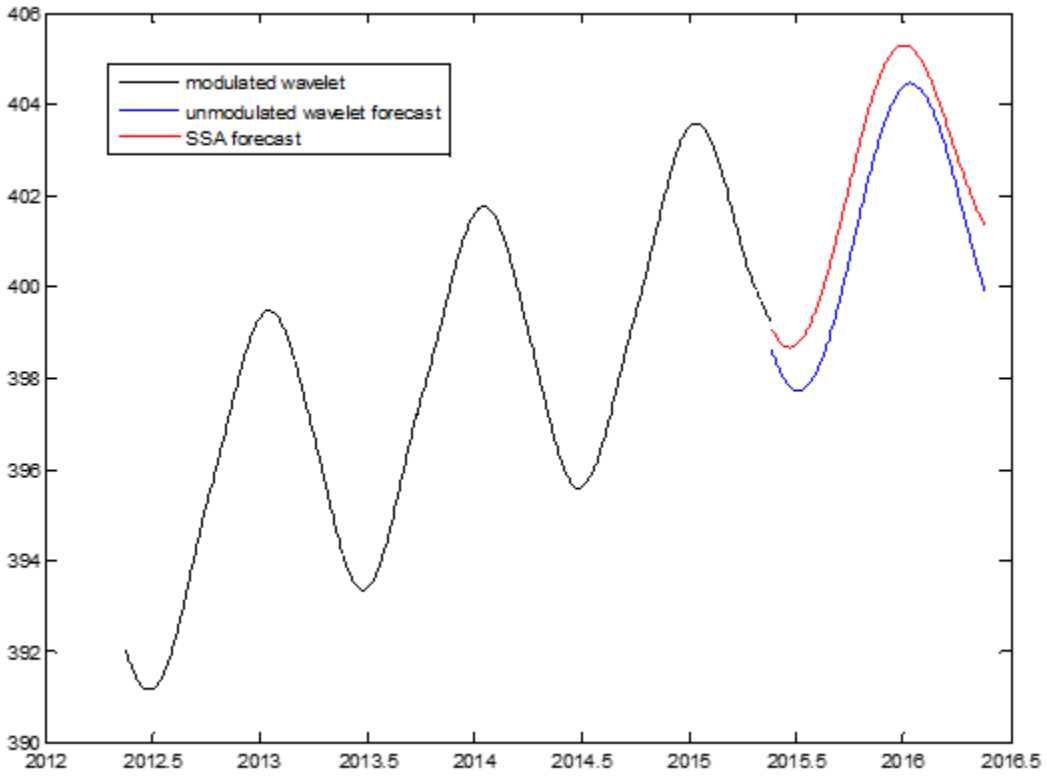


Figure A4: Pedagogical illustration for modulation. **Black Curve**: modulated wavelet fit for 3-year historical data. **Red Curve**: SSA one-year forecast. The SSA forecast takes the historical modulations into account. **Blue Curve**: unmodulated wavelet forecast, which is not as accurate as the SSA forecast.

Appendix II. Using only SSA (no wavelet)

As an exercise, we also compare with SSA run for the complete analysis, without the wavelet. We find comparable results. We actually prefer using the wavelet, since physical meaning can be assigned to its various terms.

It is possible to extract the trend and the components of the wavelet from various SSA eigenvectors. Figures A5–A8 have the results:

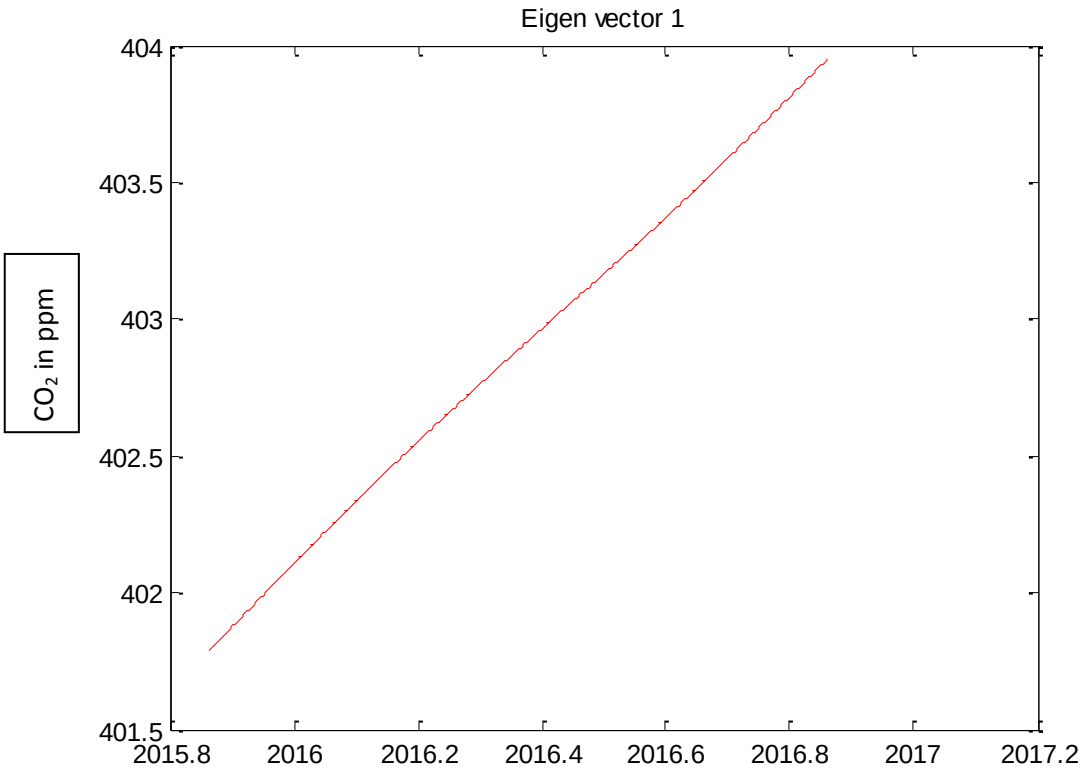


Figure A5: One year SSA forecast trend using first eigenvector.

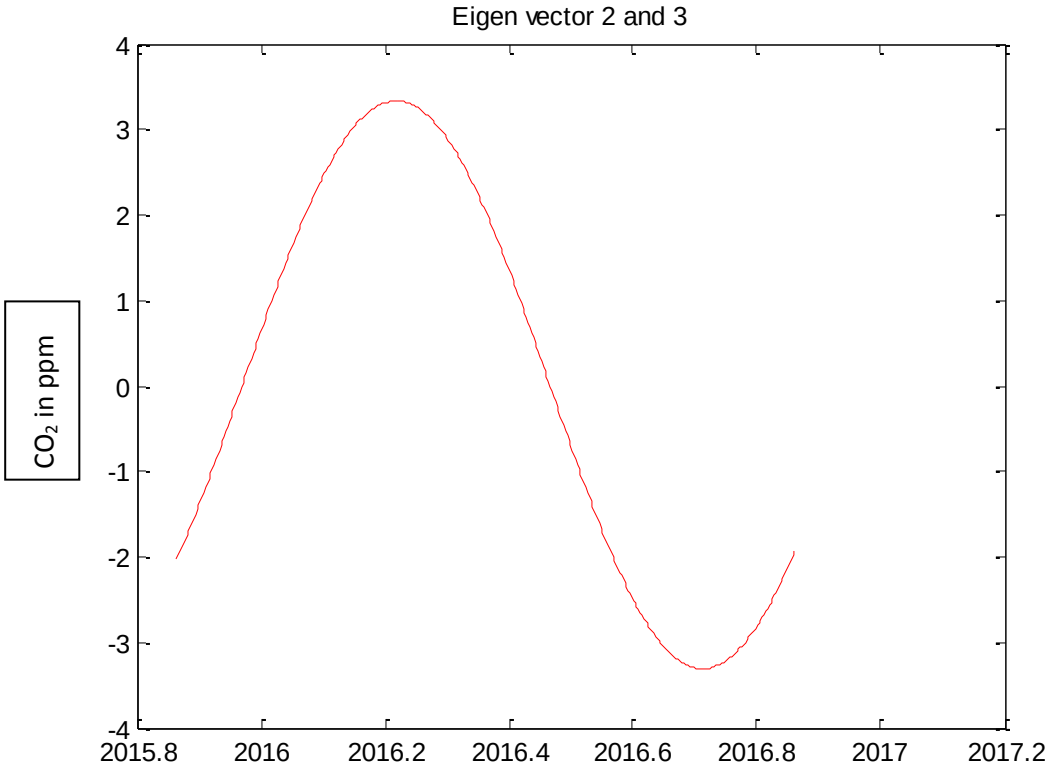


Figure A6: One year SSA forecast wavelet component is reproduced using eigenvectors 2 and 3.

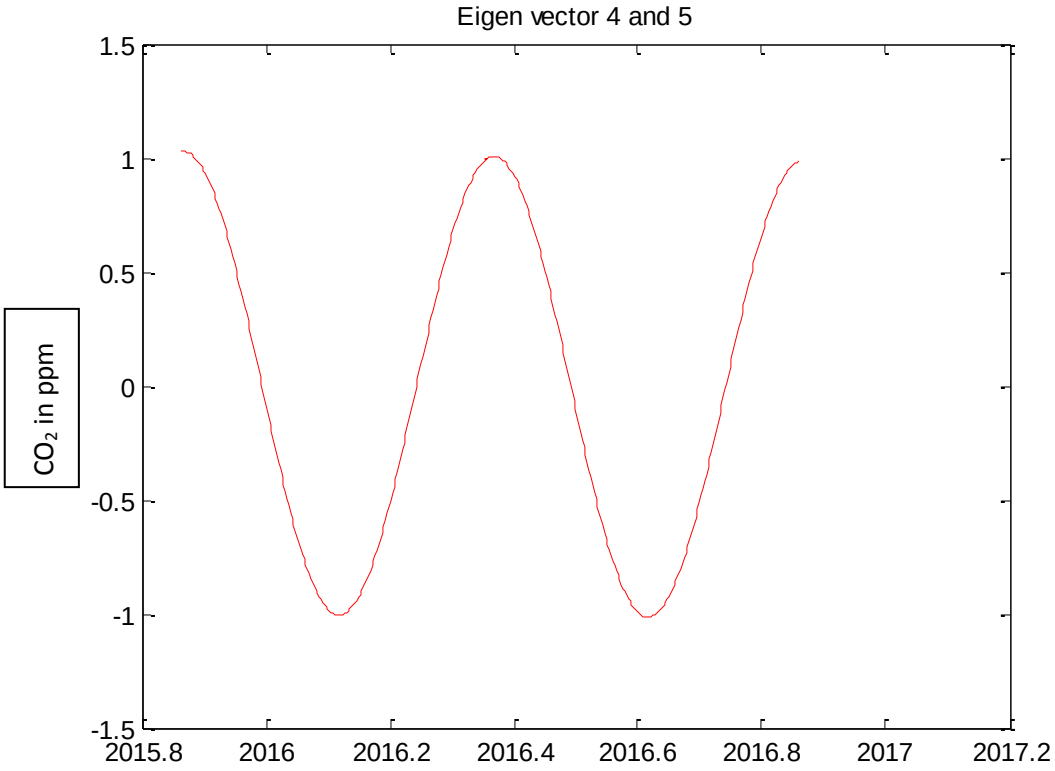


Figure A7: The wavelet leaf refinement is reproduced using SSA eigenvectors 4 and 5.

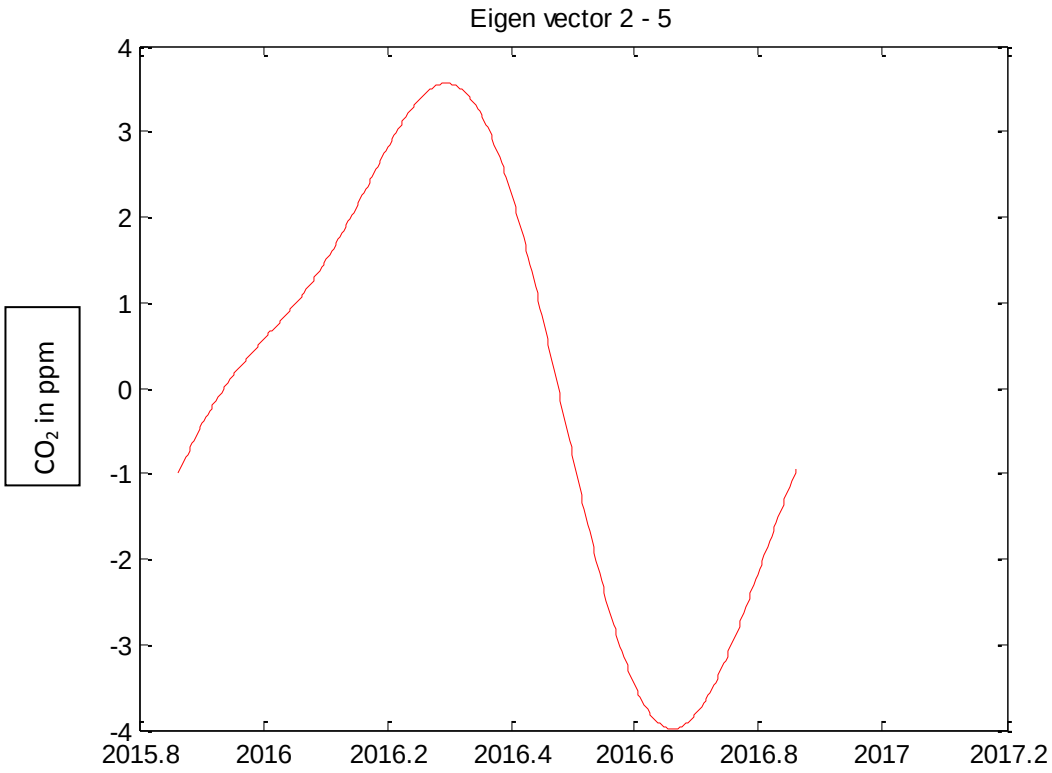


Figure A8: One year SSA forecast using eigenvectors 2, 3, 4 and 5.

Comparing the Models: Back-testing Error Analysis

Figure A9 shows the standard deviation between the forecast (made without the data in the forecast period) and the data later obtained during the forecast period, in a backtesting simulation. The errors are reduced with the leaf term in the wavelet. For comparison we also show the results without using any wavelet at all – just SSA. The errors are comparable with those of the wavelet including the leaf term. This shows the robustness of the model.

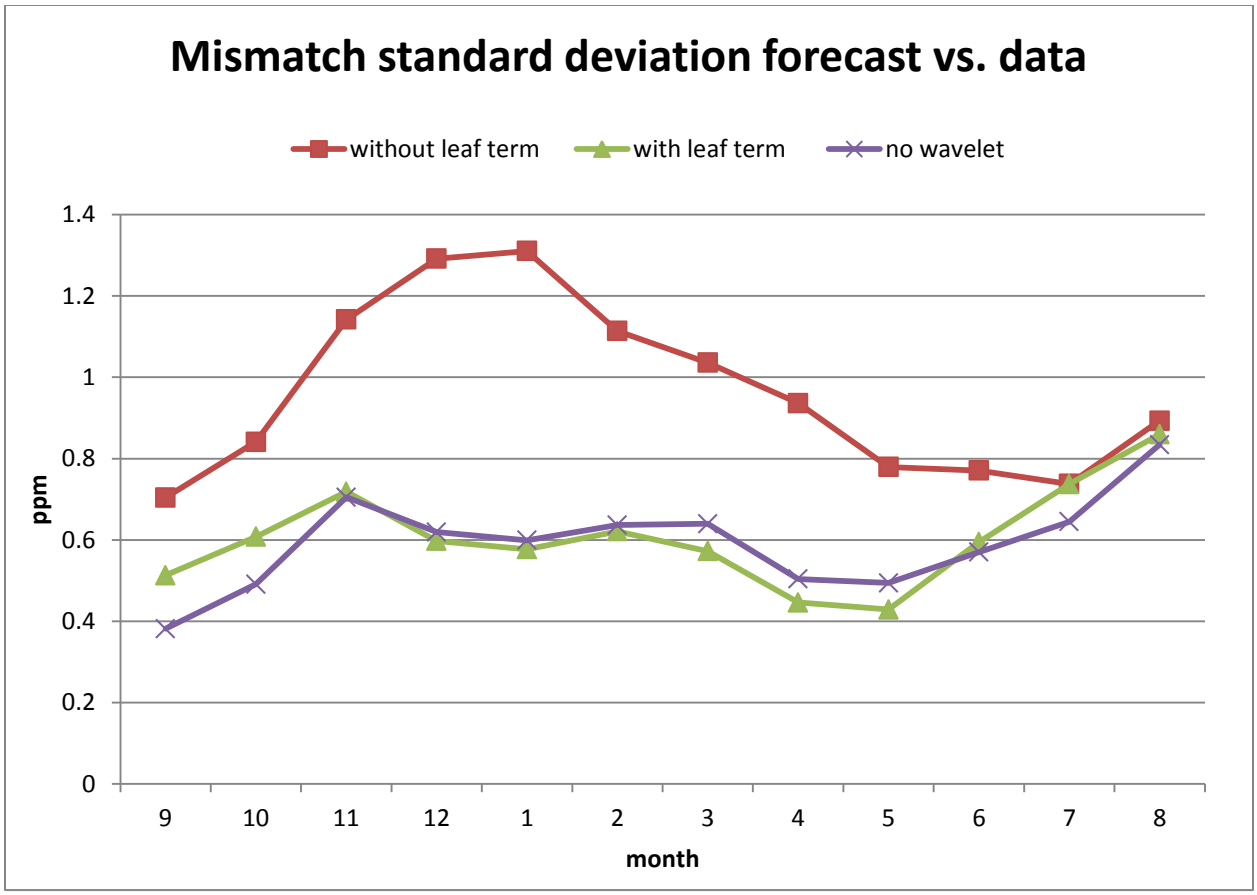


Figure A9: Forecast error comparison, wavelet without leaf term, wavelet with leaf term and SSA fit without the wavelet. The forecast error's standard deviations for the wavelet without the leaf term are bigger during fall and winter. The forecast error's standard deviations for wavelet with the leaf term and SSA without the wavelet are flat across seasons, consistent with each other, and smaller than using the wavelet without the leaf term. "Month" is a label (1 is January).

We prefer using the wavelet since the wavelet terms capture information about the physical world. While we have just shown that the same terms are contained in the SSA eigenvectors without the wavelet, these terms are extracted after the SSA fit.

Appendix III. Long-Term Forecast Simulation

For completeness, we have conducted long-term five-year forecasts using an historical simulation[17, v]. The simulation is first "trained" by applying SSA to five years of CO2 data. We then take the five-year SSA fit and break it up into five one-year segments[18]. We append each such one-year segment to the last point in the five-year training period. That gives five one-year forecast paths, each equal to a one-year SSA segment, fanning out from the last point of the training period.

We then use the same five one-year SSA segments from the original five-year SSA fit as subsequent one-year forecasts for the second forecast year. Each of these one-year segments is appended to the ends of all five first year forecasts. That gives 25 forecast paths in the second year of the simulation. We continue for three more years obtaining 125 forecast paths for the third year, 625 forecast paths for the fourth year, and 3,125 forecast paths for the fifth year.

Finally we plot the monthly data for the five years in the forecast simulation period (which were not used in the procedure) and compare the data with the forecast paths to see how the simulation stacks up with the actual data.

The dates for the two five-year simulations are in Table 1:

Time period	Simulation No. 1	Simulation No. 2
Training period	May 15, 1999 to Dec. 11, 2004	May 1, 2004 to Dec. 12, 2009
Forecast simulation period	Dec. 12, 2004 to Dec. 10, 2009	Dec. 13, 2009 to Dec. 11, 2014

Table A1: Dates defining the two five-year simulations

[17] The method used here has some common elements with a partly historical simulation using SSA for the “Macro” component of a “real-world” model for Potential Future Exposure in counterparty Risk. See Dash and Bondioli, ref.

[18] The training and forecast periods are not exactly 5 years long.

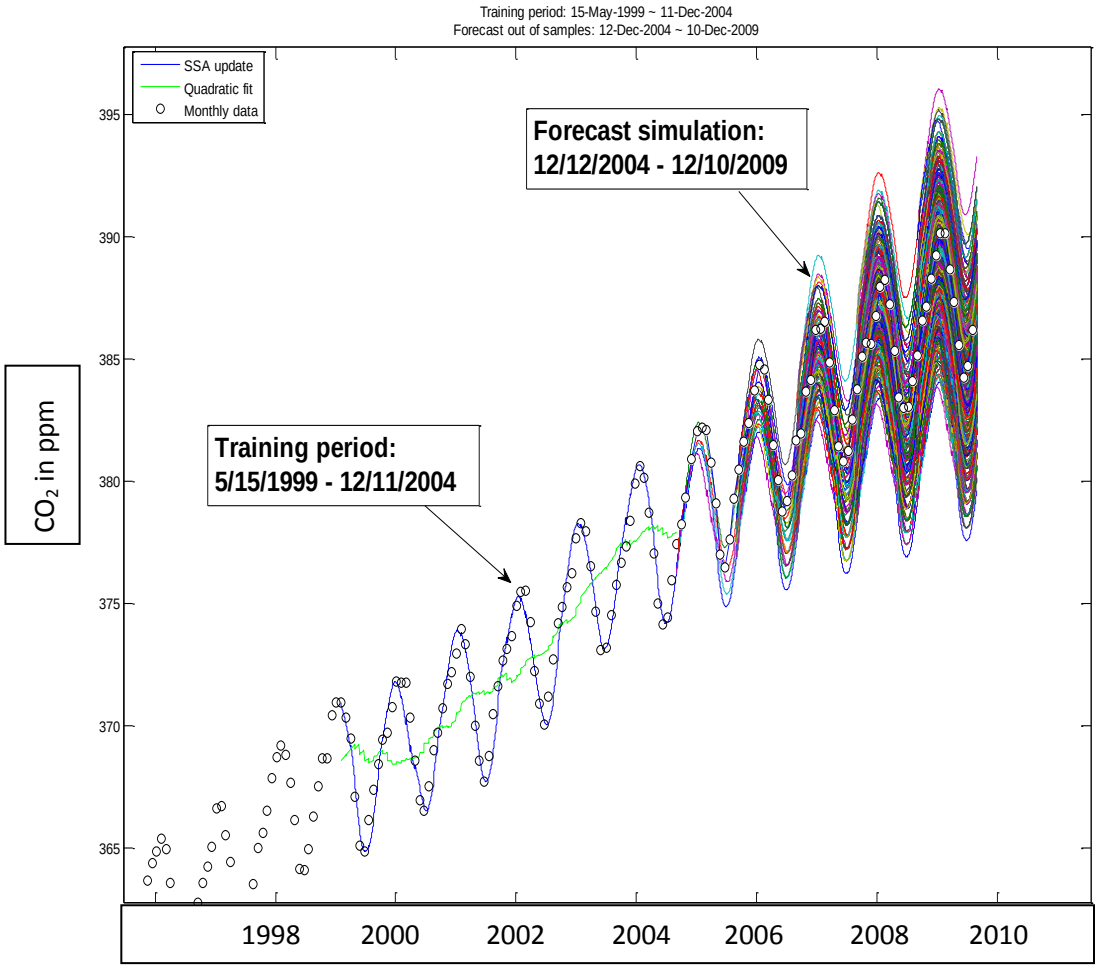


Figure A10: Five year forecast simulation No. 1 for Dec. 12, 2004 to Dec. 10, 2009

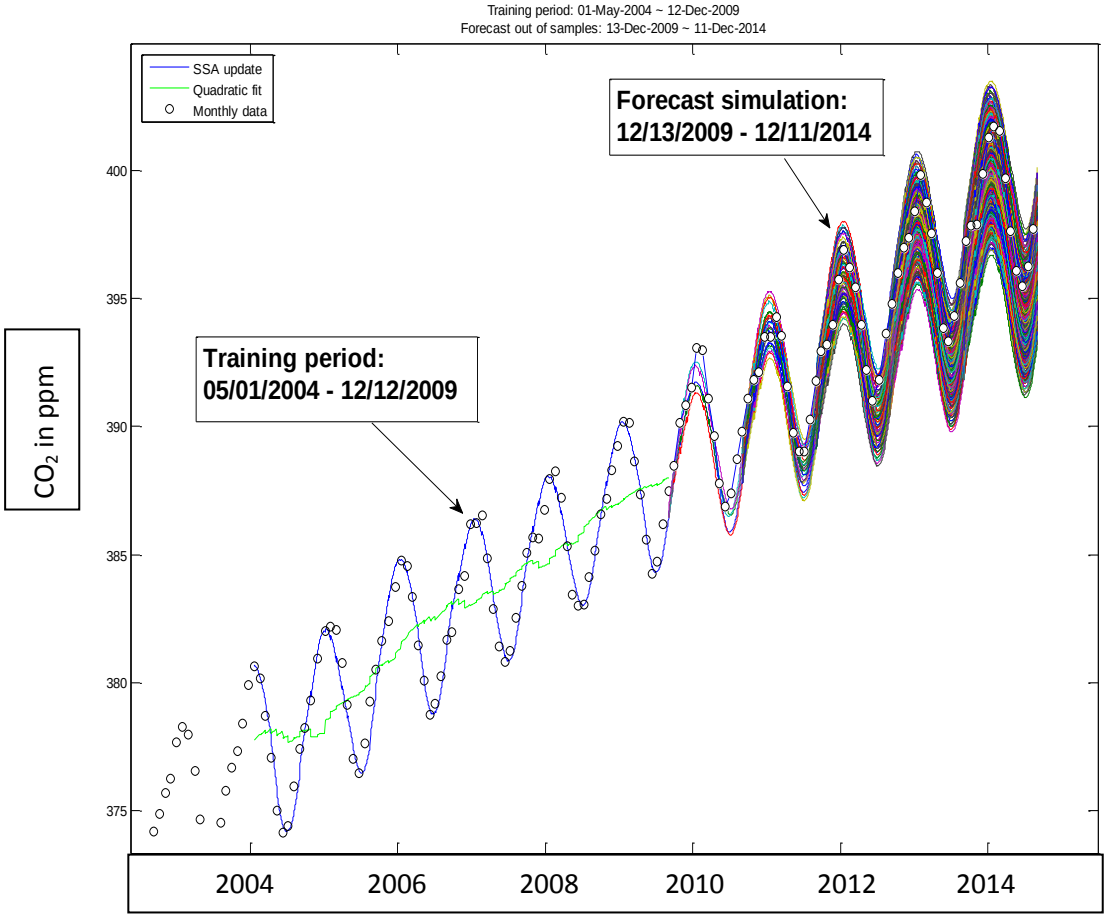


Figure A11: Long term forecast simulation No. 2 for Dec. 13, 2009 to Dec. 11, 2014

Analysis of Five-Year Simulation Results

- Simulation No. 1 reasonably reproduces the data for the forecast period 2004-2009 using the training period during 1999–2004. The 2004-2009 data fall within the ensemble of the forecast simulation paths, though are biased slightly higher than the path average.
- Simulation No. 2 also reasonably reproduces the data for the forecast period 2009-2014 using the training period during 2004–2009. The 2009-2014 data are systematically above the average of the forecast simulation paths. This implies the data of the 5-year period 2009-2014 have increased faster than the data of the 5-year period 2004-2009.

Figure A12 shows 5-year CO_2 trends defined as the quadratic terms from the wavelets, as described above. The trends have increased with time. This verifies the results of Simulation No. 1 and Simulation No. 2.

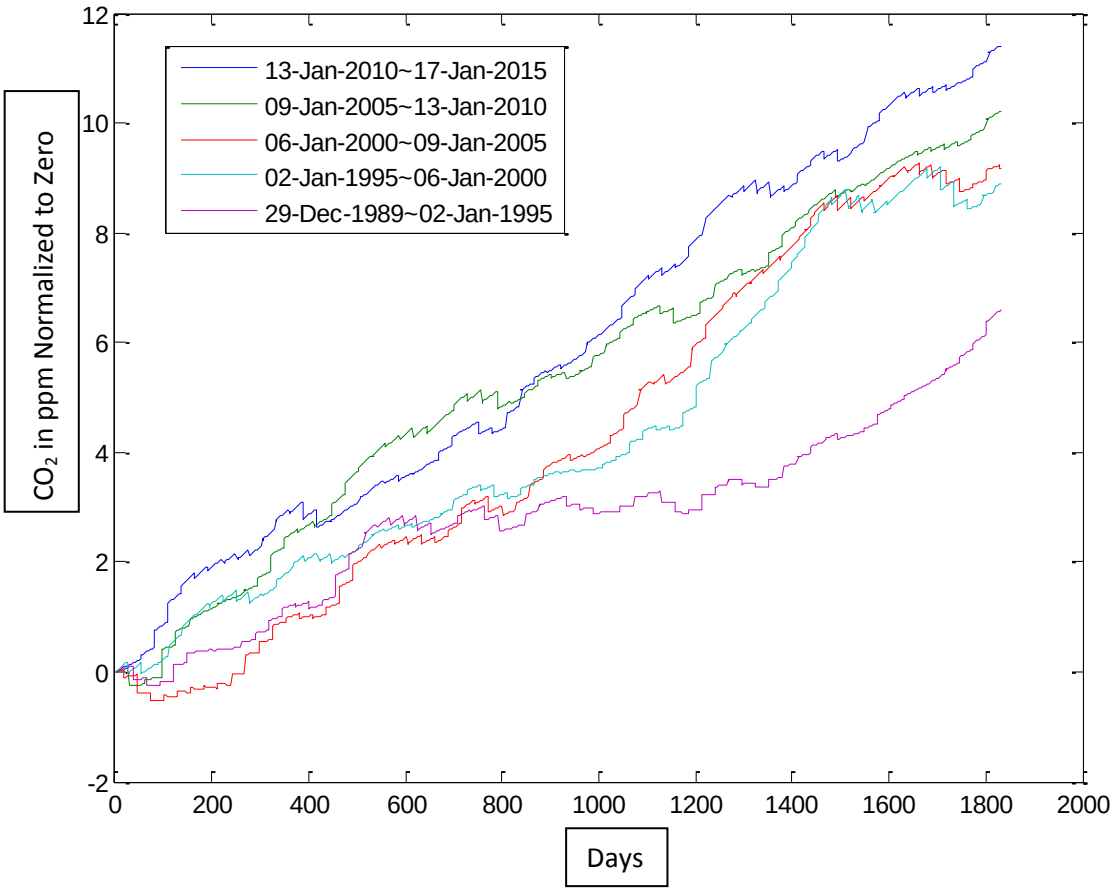


Figure A12: The CO_2 trends in ppm for 5-year periods, normalized to zero at the start.

We also verify the simulations by tracing the historical scenarios in the simulated paths:

1. Rank the five special paths generated each year in the historical (training) period.
2. Rank the quadratic fit slopes for each year in the historical period
3. Rank the shift (to the forecast time) plus the scenario minimum value for each year in the historical period
4. Rank the shift (to the forecast time) plus the scenario maximum value for each year in the historical period
5. Rank the shift (to the forecast time) plus the scenario average value for each year in the historical period
6. Check all the ranks in 2 through 5 to see if they match the simulation path ranking in 1.

This scenario tracing verification procedure is performed for both simulation No. 1 and simulation No. 2.

The table and figure below show the verification details for simulation No. 2. We find that the shift plus scenario maximum value rank, the shift plus scenario minimum value rank and the shift plus scenario's average value rank exactly match simulation path rankings. This result holds for both simulations. The quadratic fit slope rankings are very close to simulation path rankings.

Hist Scenario	Simulation Path Rank	Quadratic Slope Rank	Shift plus Min Rank	Shift plus Max Rank	Shift plus Average Rank
Year 1	2	2	2	2	2
Year 2	1	4	1	1	1
Year 3	4	1	4	4	4
Year 4	3	3	3	3	3
Year 5	5	5	5	5	5

Table A2: Scenario tracing for simulation No. 2

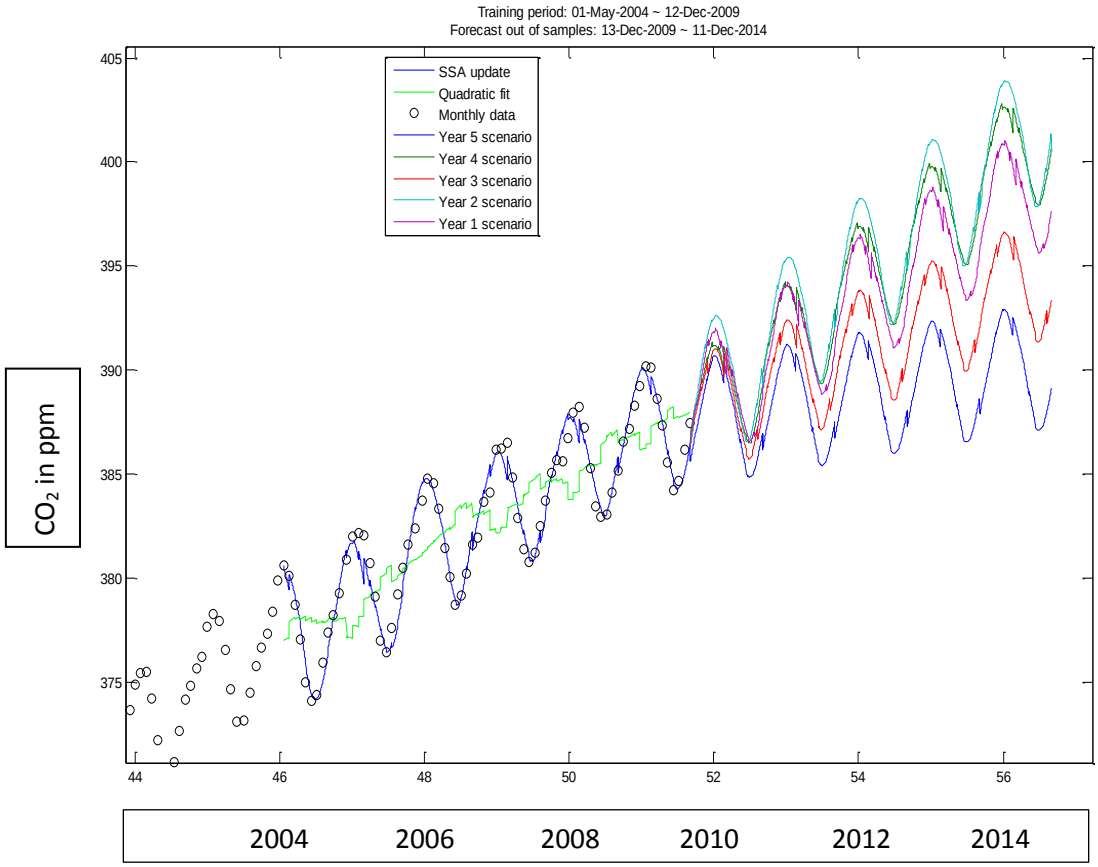


Figure A13: Scenario tracing for simulation No. 2

The future scenarios of long-term simulation are roughly normal distributed (See figures below) and their standard deviations grow at the rate of square root of time. For example in the simulation No. 2, the ratio of the future scenarios' standard deviation at fifth year and the future scenarios' standard deviation at fourth year is 1.13 while $\sqrt{\frac{5}{4}} = 1.12$. These graphs show that the widths increase as a random walk would behave, the square root of time, showing that there are no systemic errors.

Figure A14 is the histogram of widths of the scenario at the end of the fourth year, along with the Gaussian normal distribution fit (in red).

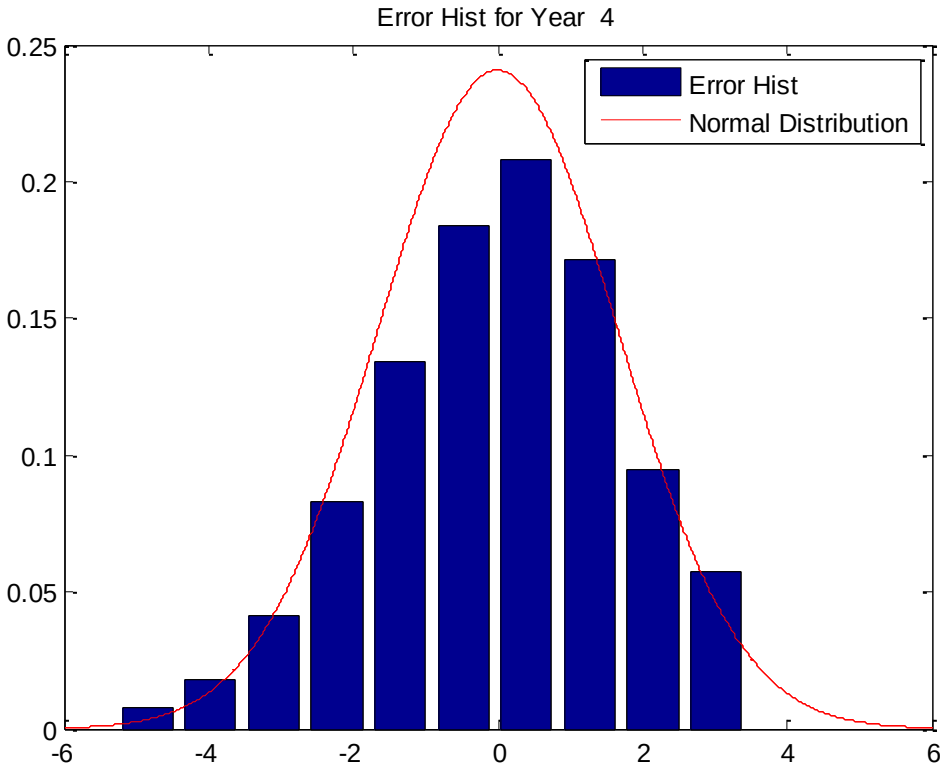


Figure A14: Future scenario error distribution at fourth year for simulation No. 2

Figure A15 is the histogram of widths of the scenario at the end of the fifth year, along with the fit of a Gaussian normal distribution (in red).

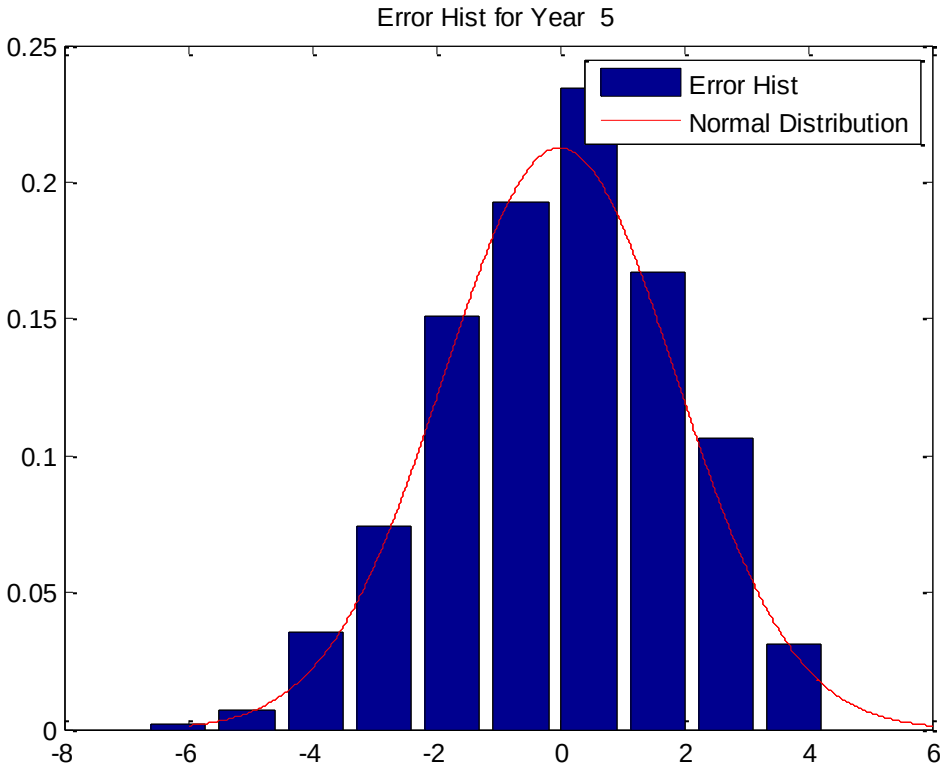


Figure A15: Future scenario error distribution at the fifth year for simulation No. 2

Appendix IV. SSA – Theory and Algorithm

Singular Spectrum Analysis (SSA) Review

The basic idea behind SSA[19] is an expansion of a given time series $\mathrm{X}(\mathrm{t})$. The expansion has information at different points in time, and hence is able to extract time-series trends. The SSA expansion is in a sense "bootstrapped" from information about $\mathrm{X}(\mathrm{t})$ itself through its autocorrelations. SSA produces a "moving average" with cleverly chosen weights. No distribution assumption for the time series $\mathrm{X}(\mathrm{t})$ is made[20]. The details are rather complicated. Here we merely give the explicit reconstruction formula, the main tool we need.

Given a time series $\{\mathrm{X}(\mathrm{t})\}$ for different times, SSA is an expansion with coefficients that are products of components $\{E_k(j)\}_{j=1...M}$ of eigenvectors $\{E_k\}_{k=1...\mathrm{M}}$ of the "lag covariance matrix" C_X. The eigenvectors are also called EOFs (Empirical Orthogonal Functions). The elements of C_X are autocorrelations of the time series, with autocorrelation time lags indexed by $j=1...M$. The reconstruction component $R_k(t)$ used as part of the approximation $\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t})$ to $\mathrm{X}(\mathrm{t})$ at a particular time t is obtained from the k^{th} SSA eigenvector E_k (t is an integer here indicating the time)[21], and is

$$R_k(t)=\frac{1}{M}\sum_{j,j'=1}^{M}\mathrm{X}(\mathrm{t}+\mathrm{j}'-\mathrm{j})E_k(j)E_k(j') \qquad (2.1)$$

Note that the data point $\mathrm{X}(\mathrm{t})$ is present on the right hand side of Eq. (2.1) from the term $j=j'$. The approximation $\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t})$ to $\mathrm{X}(\mathrm{t})$ including κ reconstruction components from SSA is

---

[19] Multivariable SSA (MSSA) formalism is a straightforward extension to several time series. The correlations and lagged correlations between the various series enters. This is used when multiple time series are related (e.g. for a yield curve with different tenors). When there is only one time series, the algorithm for MSSA reduces to the algorithm for SSA.

[20] Different notations: The top and bottom of the Wikipedia page (ref) contain different notations. The connection is: $(X,L,K,N)_{\text{Wiki Page End}}=(\mathrm{D}^T,\mathrm{M},\mathrm{N}',N)_{\text{Wiki Page Top}}$.

[21] A "padded" version of SSA involves lower and upper limits on the sum that are t-dependent. We show the simplest "unpadded" version here. Actually this formula holds sufficiently far away from end points. See the discussion below.

$$\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t}) = \sum_{k=1}^{\kappa} R_k(t) \tag{2.2}$$

Using SSA first consists of deciding which time series[22] to associate with $\mathrm{X}(\mathrm{t})$. Other parameters are the maximum lag M, the number of components κ and the data time window over which SSA is applied. These choices are made from physical considerations, and are not given by the theory.

In practice, the number of components κ retained depends on the application. For price-based SSA, the number of components can be determined such that the eigenvalues are above a noise threshold, for example determined by random matrix theory (ref [vi]). If all components are kept ($\kappa = M$), the series $\mathrm{X}(\mathrm{t})$ is exactly reproduced, i.e. $\mathrm{X}^{(\mathrm{approx})}(\mathrm{t}) = \mathrm{X}(\mathrm{t})$ at each time. This is because the eigenvectors $\{E_k\}$ form a "complete set" in mathematical terminology.

Weighted Moving Average Expression for SSA Approximation

We can rewrite Eq. (2.2) for $\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t})$ as weighted moving average, which we believe is useful for intuition.

We set $i = j' - j$ and $i' = j' + j$. The index i is used for the time shift away from time t for the contribution of $\mathrm{X}(\mathrm{t}+i)$ in the moving average to $\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t})$. We get[23]

$$\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t}) = \sum_{i=1-M}^{M-1} C_i^{(\kappa)}\, \mathrm{X}(\mathrm{t}+i) \tag{2.3}$$

After some algebra, the weights $C_i^{(\kappa)}$ turn out to be

$$C_i^{(\kappa)} = \frac{1}{M} \sum_{k=1}^{\kappa} \sum_{\mathrm{i'}=2+|i|}^{2M-|i|} E_k\left(\frac{i'+i}{2}\right) E_k\left(\frac{i'-i}{2}\right)_{\substack{i' \text{ is even (odd)} \\ \text{if } i \text{ is even (odd)}}} \tag{2.4}$$

[22] SSA in principle can also be performed on a time-differenced time series, though we do not do this here.

[23] Again this holds away from boundaries. Near boundaries the sum cuts off.

Note that $C_{-i}^{(\kappa)} = C_i^{(\kappa)}$.

As an example $\left\{C_i^{(\kappa=6)}\right\}$ is plotted in Figure A18 for the three different initial specifications described in the text.

SSA Forecasting and Updating Algorithm

There are two ways SSA is used in this paper.

1. Forecasting one month ahead of the date that the last data point was received

2. Updating the same monthly forecast at a later date when the next data point is received

For either of the above cases, the same SSA algorithm is used (ref [vii]) with the inputs as described in the text. The SSA algorithm iteratively produces estimates of time series points (either a month ahead for the forecast or the previous month for the update), which are then successively modified in the iteration until a convergence criterion is satisfied. An initial assumption must be used to start the procedure. These initial values do not affect the final results significantly, but do affect the convergence speed.

The iterative procedure in the SSA algorithm has a double loop. In the inner loop, we use SSA reconstruction with a certain number of eigenvectors used, and where the maximum number of iterations is set as a parameter. The outer loop progressively increases the number of leading eigenvectors for the inner loop. This is done until convergence is obtained. The reconstructed time series in each inner loop generates estimates for the missing points in the original time series.

Our algorithm is actually different from the original algorithm in one point. The original algorithm[viii] uses a maximum number of leading eigenvectors in the outer loop, set as a parameter at setup. Our algorithm uses a convergence test in the outer loop to decide how many eigenvectors are used. This improves convergence speed without sacrificing accuracy. We have shown that our algorithm's convergence speed is almost two times faster than the original in our tests.

Use of SSA in practice and end-point considerations

In practice, the algorithm is updated weekly with monthly forecasts. This has a technical benefit. The algorithm has end-point effects for $\mathrm{X}(\mathrm{t})$, because near the endpoint the reconstruction formula can only have previous observations. Away from the endpoint, the reconstruction formula can have both future and past observations. We avoid these end-

point problems by including forecasting a month in advance, and then using weekly updating in practice. Thus, a given forecast is only used in time regions reasonably far away from the problematic end points.

In any case, the numerical results do not indicate significant end-point effects.

The Parameters for SSA Forecasting / Updating Algorithm

The algorithm is highly configurable and has several parameters. The following outlines the parameters definitions; the numerical values are in table A3:

1. M : The lag window size for SSA. It determines the longest periodicity of time series captured by SSA.
2. η : The initial number of eigenvalues to start the algorithm's outer loop. The value of this parameter affects the algorithm's convergence search path and speed.
3. s : The incremental step size for the algorithm's outer loop. It also affects the algorithm's convergence search path and speed.
4. λ : The maximum number of algorithm's inner loop times for each outer loop. The algorithm uses this parameter to control inner loop times for each outer loop. If the inner loop times reaches the limit set up by λ , the algorithm immediately stops the inner loop and jumps to the next outer loop (even if the convergence test is negative).
5. ε : The convergence test threshold. Choosing the value for this parameter is the matter of balancing the convergence accuracy and the convergence speed. A large threshold makes the convergence test easily passed, and thus makes the whole algorithm converge fast.

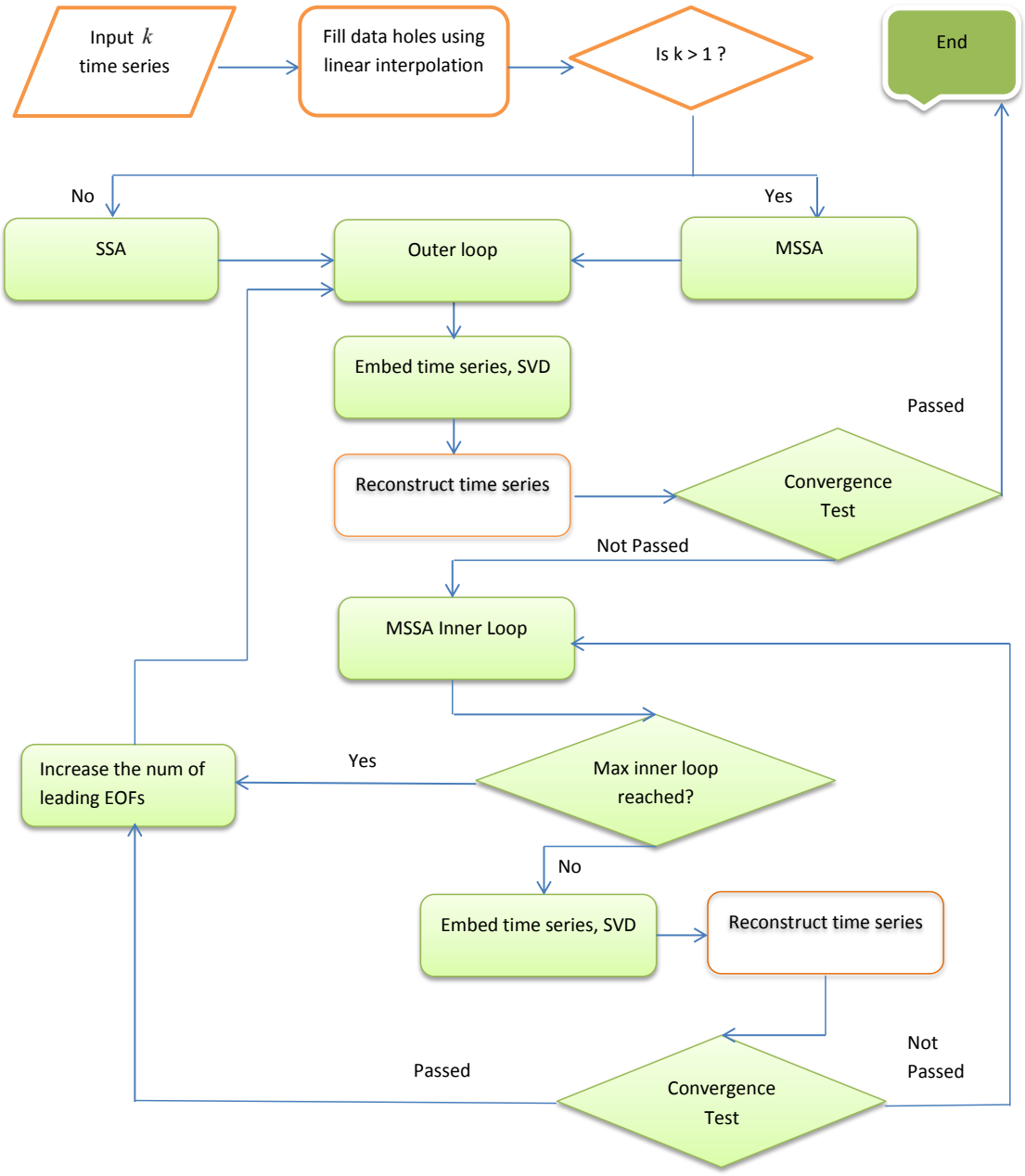
Parameter	Value
M	366
η	1
s	1
λ	200
ε	0.01

Table A3: **Parameter setup for CO_2 forecast**[24]

[24] The Modulating Gaussian Function's parameters (used by the algorithm) are automatically determined by monthly data fitting and are not listed here.

SSA Algorithm Flow Chart

Figure A16 below illustrates the SSA algorithm as described above [25].



[25] All light green boxes are configurable by the algorithm's parameters. The same diagram holds for the multivariable SSA (MSSA), as indicated.

Appendix V: Dependence of SSA Forecasts on initial future CO_2 specifications

The SSA forecast is an iterative procedure that adjusts future values via SSA until convergence occurs[26]. The initial specification for the future values is needed to start the iterative procedure. The choice of initial specification for the future values affects the SSA forecast accuracy and the algorithm's convergence speed.

We compare three choices for SSA initial future CO_2 specifications. The first choice is to use the most recent 3-year historical CO_2 monthly data to fit the wavelet function, and then extend the fitted wavelet function to the future time horizon as the initial future specification for the SSA forecast. The second choice is flat extrapolation of the last CO_2 data point as the initial future specification. The third choice is zero future CO_2 as the initial future specification; while this choice is unphysical, it serves as a no-prior-information benchmark.

The results are that the "wavelet" initial future specification is somewhat more accurate and has faster convergence than the other two choices.[27]

Figure A17 shows the 10-year backtest for the three methods for the CO_2 initial future specification compared to the benchmark SSA update described in the text. The "wavelet" initial future specification result is very close to the "zero" initial future specification result, and both are somewhat better than the "flat extrapolation" initial future specification.

[26] Convergence is defined by successive iteration results differing less than a preassigned small number (the convergence test threshold). See appendix IV.

[27] The "wavelet" initial method also improves SSA performance around/beyond the "current" time boundary for the forecast.

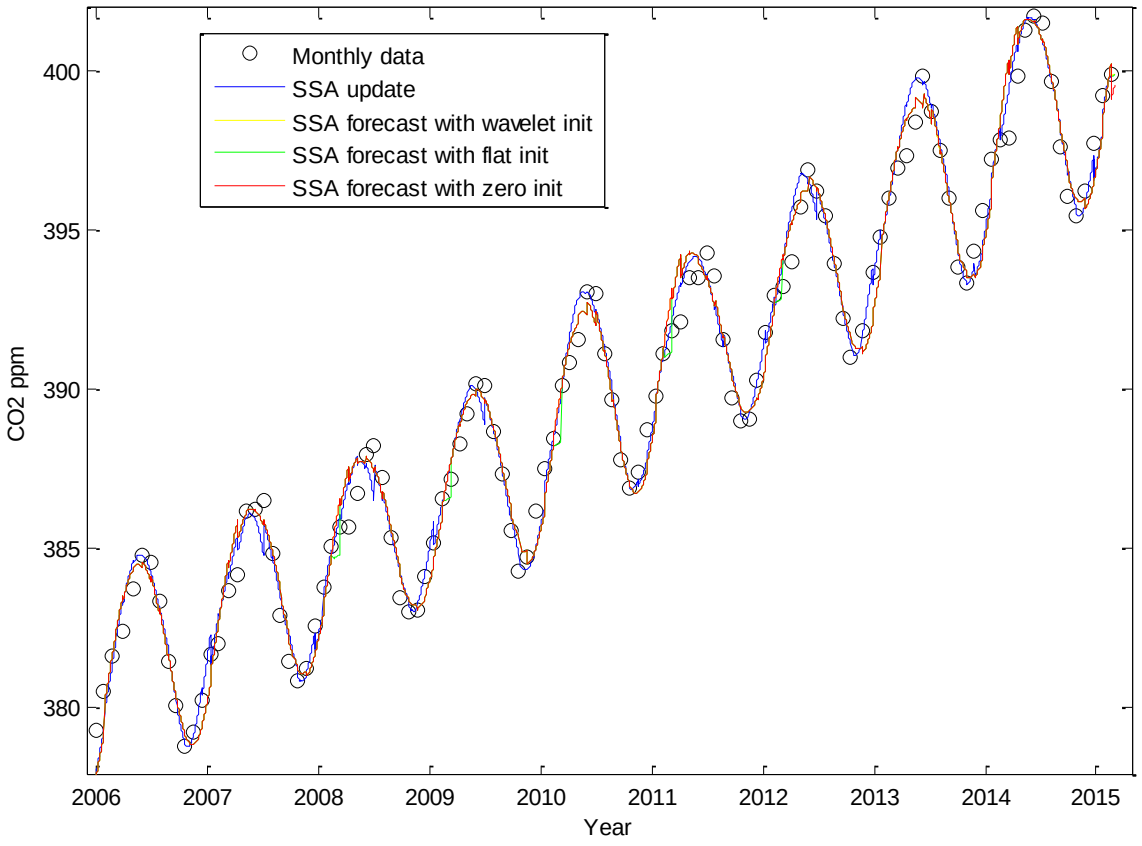


Figure A17: Back test for SSA forecast with three initial specification methods, compared with the SSA update (**blue**). SSA forecast with wavelet CO_2 initial specification (**yellow**, visually indistinguishable) overlaps the SSA forecast with zero CO_2 initial specification (**red**). Both are somewhat better than flat CO_2 initial specification (**green**), which shows some deviation half way going up.

Initial specifications and SSA algorithm convergence

The "wavelet" initial specification convergence speed is faster than "zero" initial future-specification convergence speed. This is because the zero initial value is very far from the correct value, and consequently the SSA forecast takes more time to converge. This section tests the approach to convergence and looks at the very first step using the initial eigenvectors from the different initial specifications and not iterating.

Recall (see Appendix IV) that the SSA approximation away from boundaries is is $\mathrm{X}^{(\mathrm{approx};\kappa)}(\mathrm{t}) = \sum_{i=1-M}^{M-1} C_i^{(\kappa)}\, \mathrm{X}(\mathrm{t}+i)$. Near boundaries the sum cuts off.

Figure A18 shows the SSA reconstruction component weighted-moving average coefficients $\{C_i^{(\kappa=6)}\}$ for the three initial future specifications, using the initial eigenvectors from the different initial specifications and not iterating. The corresponding CO_2 values resulting from these initial eigenvectors: = 400.8 ppm (wavelet CO_2 initial specification **blue**), = 400.1 ppm (flat CO_2 initial specification **green**), and = 63.5 ppm (zero CO_2 initial specification **red**), far from the physical CO_2 value of around 400 ppm. In this last case the non-zero values of $\{C_i^{(\kappa=6)}\}$ are concentrated at one year. Physically, if there is no information initially, the best first-stage approximation for the roughly periodic CO_2 value is the value one year away. Note $C_{-i}^{(\kappa)} = C_i^{(\kappa)}$. For the forecast, the values around - one year are actually relevant. Note also that $M = 366$ days.

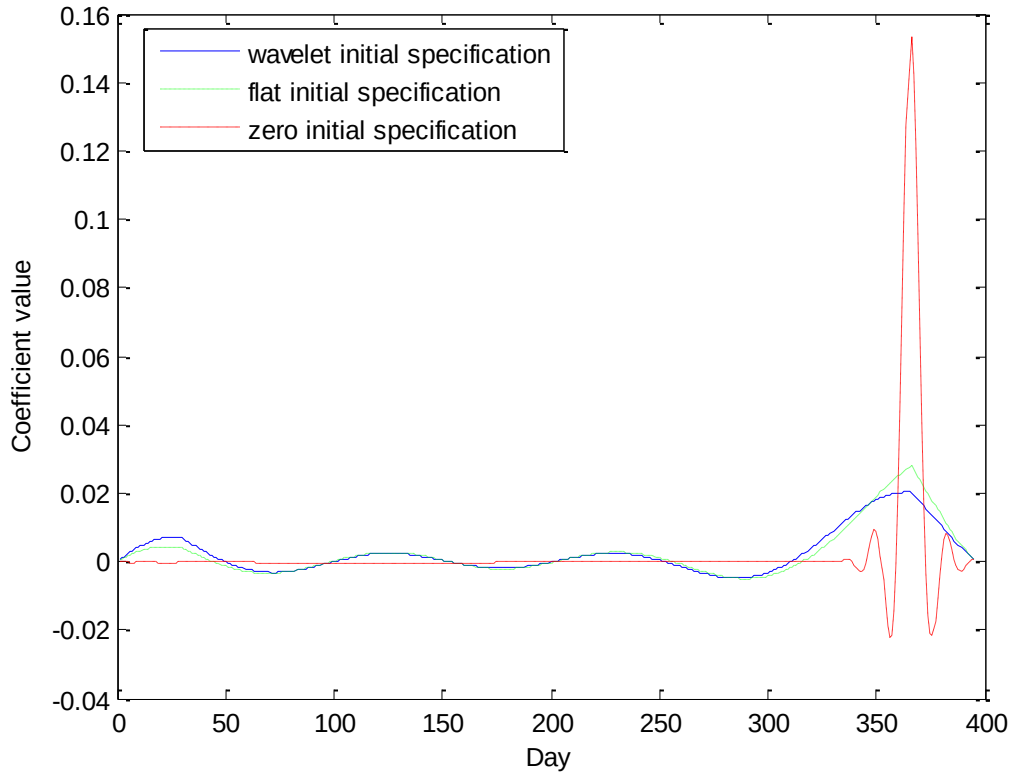


Figure A18: Plot of the total reconstruction coefficients $\{C_i^{(\kappa=6)}\}$ in the moving average reconstruction approximation with $\kappa = 6$ and $i \geq 0$, using the initial eigenvectors and not iterating. The "Day" axis (calendar days) is the index i. $\{C_i^{(\kappa=6)}\}$ are similar for the wavelet (**blue**) and flat (**green**) initial CO_2 specifications. $\{C_i^{(\kappa=6)}\}$ for the zero (**red**) initial CO_2 specification are concentrated at $\pm$one year.

References

[i] IPCC, 2014: *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp. on the IPCC AR5 Synthesis Report website

[ii] Steffen et al. 2015. Planetary Boundaries: Guiding human development on a changing planet. *Science* Vol. 347 no. 6223

[iii] Ghil, M. and Allen, M. R. "Advanced Spectral Methods for Climatic Time Series," *Reviews of Geophysics*, 40, 1 (2001)

[iv] Dettinger, M. and Ghil, M., "Seasonal and interannual variations of atmospheric CO_2 and climate," *Tellus* 50B, 1 (1998).

[v] Dash, J. W., Bondioli, M., and Yang, X., *THE MACRO-MICRO MODEL, TRENDS VS. NOISE, AND SSA - I*, Bloomberg LP working paper (2016), SSRN http://papers.ssrn.com/abstract=2808176

[vi] Akemann, G. et al (editors), *The Oxford Handbook of Random Matrix Theory*, Oxford U. Press, (2011)

[vii] Dash, J. W. and Zhang, Y., *CLEANING FINANCIAL DATA USING SSA AND MSSA*, Bloomberg LP Working paper (2016), SSRN http://papers.ssrn.com/abstract=2808156

[viii] (A). Kondrashov, D. and Ghil, M. *Spatio-temporal filling of missing points in geophysical data sets*, Nonlinear Processes in Geophysics, 13, 151-159 (2006);

(B). Musial, J.P., Verstraete, M.M. and Gobron, N. *Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series*, Atmospheric Chemistry and Physics, 11, 14259-14308 (2011)

File: BLOOMBERG CARBON CLOCK TECHNICAL WORKING PAPER RevSep2016 Posted.docx
Date: 9/19/2016 9:32:00 AM