Surrogate gas prediction model as a proxy for \( \Delta^{14}C \) based measurements of fossil fuel–\( \text{CO}_2 \)

Kevin J Coakley\(^1\), John B Miller\(^2,3\), Stephen A Montzka\(^2\), Colm Sweeney\(^2,3\),

Ben R Miller\(^2,3\)

---

Corresponding author: Kevin J Coakley, National Institute of Standards and Technology, 325 Broadway, MS 898.03, Boulder, CO, USA. (kevin.coakley@nist.gov)

\(^1\)Statistical Engineering Division,
National Institute of Standards and Technology, Boulder, CO, USA

\(^2\)Global Monitoring Division, National Oceanic and Atmospheric Administration,
Boulder, CO, USA

\(^3\)CIRES, University of Colorado, Boulder, CO
Key Points.

- Proxy–gas model for fossil fuel–CO₂ measurements presented.
- Method may increase number of fossil fuel–CO₂ estimates by up to three times.
- Uncertainties of proxy model predictions quantified.

Abstract. The measured $^{14}$C:$^{12}$C isotopic ratio of atmospheric CO₂ (and its associated derived $\Delta^{14}$C value) is an ideal tracer for determination of the fossil fuel derived CO₂ enhancement contributing to any atmospheric CO₂ measurement ($C_{ff}$). Given enough such measurements, independent top-down estimation of US fossil fuel- CO₂ emissions should be possible. However, the number of $\Delta^{14}$C measurements is presently constrained by cost, available sample volume, and availability of mass spectrometer measurement facilities. $\Delta^{14}$C is therefore measured in just a small fraction of samples obtained by flask air sampling networks around the world. Here, we develop a Projection Pursuit Regression (PPR) model to predict $C_{ff}$ as a function of multiple surrogate gases acquired within the NOAA/ESRL Global Greenhouse Gas Reference Network (GGGRN). The surrogates consist of measured enhancements of various anthropogenic trace gases, including CO, SF₆, and halo- and hydrocarbons acquired in vertical airborne sampling profiles near Cape May, NJ and Portsmouth, NH from 2005 through 2010. Model performance for these sites is quantified based on predicted values corresponding to test data excluded from the model building process. Chi-square hypothesis test analysis indicates that these predictions and corresponding observations are consistent given our uncertainty budget which accounts for random effects and
one particular systematic effect. However, quantification of the combined un-
certainty of the prediction due to all relevant systematic effects is difficult
because of the limited range of the observations and their relatively high frac-
tional uncertainties at the sampling sites considered here. To account for the
possibility of additional systematic effects, we incorporate another compo-
nent of uncertainty into our budget. Expanding the number of $\Delta^{14}C$ mea-
surements in the NOAA GGGRN and building new PPR models at addi-
tional sites would improve our understanding of uncertainties and potentially
increase the number of $C_{ff}$ estimates by approximately a factor of three. Pro-
vided that these estimates are of comparable quality to $\Delta^{14}C$ -based esti-
mates, we expect an improved determination of fossil fuel- $CO_2$ emissions.
1. Introduction

Over the last two hundred years, the amount fraction of CO$_2$ in the atmosphere has increased from approximately 280 micromole of CO$_2$ per mole of dry air to nearly 400 micromole of CO$_2$ per mole of dry air [Ballantyne et al., 2012; Etheridge et al., 1996]. (Henceforth, for the sake of brevity, we denote micromole per mole as ppm.) There is an overwhelming scientific consensus that this increase is due primarily to fossil fuel emissions despite the fact that about half of the CO$_2$ produced by burning fossil fuels is absorbed by oceans and the terrestrial biosphere [Canadell et al., 2007; Knorr et al., 2009; Ballantyne et al., 2012]. Although year-to-year increases in CO$_2$ are well explained by the global average of remote atmosphere measurements, variability at shorter time scales over large land areas is often dominated by terrestrial biosphere exchange. To fully understand the existing array of atmospheric CO$_2$ measurements in terms of potential sources and sinks, determinations of CO$_2$ enhancements due solely to combustion of fossil fuels, $C_{ff}$, are essential. Currently, $C_{ff}$ is estimated from inventories of fossil fuel- CO$_2$ emissions that are based on economic statistics [Gurney et al., 2009; Boden et al., 2014]. Atmosphere-based determinations of a sufficiently large number of $C_{ff}$ values should enable both near real-time monitoring of fossil fuel emissions, and an independent assessment of bottom-up inventories.

Due to nuclear reactions associated with cosmic rays, $^{14}$C is continually produced in the atmosphere and then rapidly oxidized to produce $^{14}$CO$_2$ [Anderson et al., 1947; Lin- genfelter, 1963; Suess, 1965]. Atmospheric testing of nuclear weapons during the 1960s nearly doubled the amount of $^{14}$CO$_2$ in the Northern Hemisphere atmosphere at one point.
in time [Dal and Suess, 1968]. Since then, much of this excess has been absorbed by the oceans and biosphere. Nuclear reactors also produce $^{14}\text{CO}_2$ [Yim and Caron, 2006] at spatially varying rates. Notably, these rates are generally higher in Europe than in the US [Graven and Gruber, 2011; Vogel et al., 2013]. Although the effect of nuclear reactor emissions on $C_{ff}$ has been assessed at CMA and NHA [Lehman et al., 2013], in order to develop proxy prediction models elsewhere, one must assess nuclear reactor effects on a site-to-site basis.

Since underground fossil fuels stored in the rock reservoir were last in contact with the atmosphere hundreds of millions years ago, and $^{14}\text{CO}_2$ decays with a half life of (5700 ± 30) years [National Nuclear Data Center, Brookhaven National Laboratory, www.nndc.bnl.gov], the isotopic ratio $^{14}\text{C}:^{12}\text{C}$ of $\text{CO}_2$ produced by burning fossil fuels is negligible and equated to zero in any analysis. To understand this point more clearly, note that the probability that a $^{14}\text{C}$ nucleus does not decay until after 100 million years is $10^{-5281}$ given that its half life is 5700 years. Since CO$_2$ contributed by nearly all other sources is in near equilibrium with the atmosphere, a measurement of $^{14}\text{C}:^{12}\text{C}$ and its associated derived $\Delta^{14}\text{C}$ value [Stuiver and Pollack, 1977] (note that $\Delta^{14}\text{C}$ corresponds to $\Delta$ in Stuiver and Pollack [1977]) enables one to determine $C_{ff}$ according to relatively simple mass balance considerations [Levin et al., 2003; Levin and Carsten, 2007; Turnbull et al., 2007; Turnbull et al., 2011; Vogel et al., 2010; Miller et al., 2012]. Approximately 5000 high precision measurements of $\Delta^{14}\text{C}$ (with random uncertainty approximately 0.2 %) and associated derived $C_{ff}$ estimates (with random uncertainty approximately 1 ppm), would enable an independent “top-down” (i.e. atmosphere-based) estimation of monthly averages of US national and regional fossil fuel-$\text{CO}_2$ emissions with random uncertainties
in the 5% -10% range [Committee on Methods for Estimating Greenhouse Gas Emissions; National Research Council, 2010; Basu et al., 2016]. However, our ability to make the necessary $\Delta^{14}C$ measurements (e.g. several thousand per year just for the US) is currently constrained by cost, accessibility to accelerator mass spectrometers (AMS) and the volume of air (approximately 2 liters at STP [Lehman et al., 2013]) required to make high precision measurements of $\Delta^{14}C$. Thus, $\Delta^{14}C$ is currently measured in just a small subset of air samples obtained from sampling networks (including the US portion of the NOAA/ESRL Global Greenhouse Gas Reference Network (GGGRN)). A substantial number of new $C_{ff}$ measurements determined from either $\Delta^{14}C$ or surrogate gases measurements would greatly facilitate efforts to verify “bottom-up” inventory approaches for US fossil fuel CO$_2$ emissions.

Various studies have presented univariate proxy models for $C_{ff}$ based on CO [Turnbull et al., 2011; Vogel et al., 2010]. In these studies, $C_{ff}$ is predicted based on interpolated measured ratios of CO:$C_{ff}$. However, spatio-temporal extrapolation of measured ratios of CO:$C_{ff}$ is problematic because the relative intensity of non-fossil fuel CO sources can vary as a complicated function of time and space. Further, CO-based proxies can be problematic because many sources of fossil fuel emissions such as vehicles with diesel engines and power plants have very low CO emission rates. Miller et al. [2012] demonstrated that a wide range of other anthropogenic trace gases such as SF$_6$, and a number of halo- and hydro-carbons that are frequently measured in NOAA network samples are positively correlated with observed $C_{ff}$. Because of these correlations, a multivariate proxy prediction model for $C_{ff}$ based on measured surrogate gases is plausible. If one could accurately predict $C_{ff}$ as a function of multiple anthropogenic trace gases over broad spatio-temporal
regions of the US, such a prediction model could serve as proxy for a $\Delta^{14}$C-based measurement of $C_{ff}$ and expand the temporal density of reliable $C_{ff}$ estimates ($\Delta^{14}$C-based measurements and proxy model predictions) by a factor of approximately three at towers and aircraft vertical profiling sites throughout North America [Sweeney et al., 2015; Andrews et al., 2014]. In this work, we construct such a prediction model based on multiple surrogate gas measurements and quantify its performance. To the best of our knowledge, our model is the first multivariate proxy prediction model for $C_{ff}$.

2. Methods

We develop Projection Pursuit Regression (PPR) prediction models based on $\Delta^{14}$C-derived measurements of $C_{ff}$ and measured enhancements of twelve anthropogenic surrogate gases acquired by NOAA from vertical aircraft profiles at two sites, CMA (Cape May, NJ) and NHA (Portsmouth, NH) (see Figure 1 and Table 1), between 2005 and 2010. A large fraction of these measurements (between 2005 and 2009) were first presented and analyzed in Miller et al. [2012]. The rest of the measurements were acquired between 2009 and 2010.

Given these $C_{ff}$ and surrogate gas data, we determine which trace gases to include in PPR prediction models, as well as their complexity and form, with a statistical learning method [Hastie et al., 2008] called cross-validation [Arlot and Celisse, 2010; Stone, 1974; Stone, 1977]. Following Hastie et al. [2008], we split the observed data into model building data and test data. The model building data subset is then split into training data and validation data subsets. After fitting each candidate model to the training data, we predict $C_{ff}$ values corresponding to observed values in the validation data. We select the model that minimizes the discrepancy between the observed and predicted $C_{ff}$ values in the
validation data. Since both the training and validation data are involved in the model selection process, quantification of prediction model performance based on how well the model fits the validation data is an imperfect and usually overly-optimistic measure of how well the model performs for new data excluded from the model building process. To better quantify model performance for new data, we predict $C_{ff}$ corresponding to test data excluded from the model building process.

We stress that the only inputs to the PPR prediction model are measured surrogate gas enhancements. That is, the time and location of any surrogate gas enhancement measurement are not predictors in the PPR model. Later in this work, we discuss how possible spatial and temporal effects such as systematic temporal variations of surrogate gas emission rates could affect the performance of our prediction model. Next, we present measurement models for $C_{ff}$ and surrogate gas enhancements following Miller et al. [2012].

2.1. Measurement models and data

The estimated value $C_{ff,obs}$ of the theoretical (true) value $C_{ff}$ for any sample is based, in part, on the measured isotopic ratio of $^{14}$C and $^{12}$C, $^{14}R_{obs} = ^{14}C/^ {12}C$, for that sample. Following Stuiver and Pollack [1977], we estimate the theoretical value of $\Delta^{14}$C as $\Delta_{obs}$, where

$$\Delta_{obs} = ^{14}R_{obs}/^{14}R_{standard} \times \Theta_{obs} - 1,$$

and $\Theta_{obs}$ is a correction term (which depends on the measured isotopic ratio of $^{13}$C and $^{12}$C for the sample) that accounts for systematic effects due to mass-dependent fractionation.
Above, isotopic ratio estimates are also corrected for small radioactive decay losses during the interval between acquisition and measurement of the sample.

We model the theoretical values of $C_{ff}$ and surrogate gas enhancements with the simple one-dimensional analytical framework employed by Miller et al. [2012]. In this approach, the theoretical mole fraction of a gas at lower-levels (typically at approximate altitudes of 300 and 2200 m above sea level (asl) ) is the sum of a theoretical background plus a theoretical enhancement due to recent anthropogenic emissions of that gas at the site of interest. Further, we assume that the theoretical background at lower levels is the same as the theoretical background at upper levels in the relatively well-mixed free troposphere (typically 4000 m asl), and that theoretical enhancements are 0 at upper levels. Based on these assumptions, we estimate any particular surrogate gas enhancement (i.e. an additive signal relative to a background) as the difference between measured mole fractions at lower and upper levels of a particular sampling profile.

Following earlier work [Turnbull et al., 2006; Miller et al., 2012], we estimate $C_{ff}$ as $C_{ff,obs}$ where

$$C_{ff,obs} = \frac{C_{obs}(\Delta_{obs} - \Delta_{bg,obs})}{\Delta_{ff} - \Delta_{bg,obs}} - \hat{C}_{cor},$$

(2)

where $\Delta_{ff} = -1$, and the measured values $C_{obs}, \Delta_{obs}, \Delta_{bg,obs}$ are estimates of the unknown theoretical values $C, \Delta$ and $\Delta_{bg}$ (see Appendix A for more details). Above, $\Delta$ is shorthand for measured $\Delta^{14}C$, and $C$ is shorthand for CO$_2$, and $\hat{C}_{cor}$ is an estimate of small contributions to the tropospheric $^{14}C$ budget that may influence the measured enrichment or depletion relative to background. These contributions include the heterotrophic respiration return flux of $^{14}C$ typically photosynthetically assimilated a decade or two ago.
when atmospheric $^{14}$CO$_2$ was higher than it is today. To be consistent with Miller et al. [2012], $\dot{C}_{cor}$ neglects possible point source emissions of $^{14}$CO$_2$ from nuclear power plants.

Any particular measurement of $C_{ff}$ is affected by both random and systematic measurement errors. Following Miller et al. [2012] and Lehman et al. [2013], we estimate the random uncertainty of any $C_{ff}$ measurement as $\hat{\sigma}_{obs,ran} = 1$ ppm. The majority of this uncertainty (0.9 ppm) is due to the random uncertainty in each $\Delta^{14}C$ measurement.

We assume that each $C_{ff}$ measurement has a bias (systematic error) due to: imperfect modeling of backgrounds; imperfect modeling of the transport of emissions associated with heterotrophic respiration; and neglecting the effect of nuclear power plants on $C_{ff}$. These systematic errors vary from observation to observation. Based on previous analysis [Miller et al. 2012; Lehman et al. 2013] we estimate the standard deviation of the systematic error for $C_{ff}$ measurements, $\hat{\sigma}_{obs,sys}$, to be 0.5 ppm. We estimate the combined uncertainty for any $C_{ff}$ measurement as $\hat{\sigma}_{obs,tot} = \sqrt{\hat{\sigma}_{obs,ran}^2 + \hat{\sigma}_{obs,sys}^2} = 1.1$ ppm.

The high variability of the $C_{ff,obs}$ time series (Figure 2) is expected, in part, because of spatial variation of emissions and their ratios [Miller et al., 2012], temporal variation of wind directions and associated gas transport, and the variations in altitude where data are acquired. In particular, measured $C_{ff}$ values at approximately 2200 m asl are typically low and have large fractional uncertainties because the altitude of the boundary between the free troposphere and the PBL fluctuates about 2200 m asl. In Figure 3, we show scatterplots of $C_{ff,obs}$ and those surrogate gas enhancements selected by at least one of the four PPR prediction models that we study. Although only CO, NMHCs and CH$_4$ are potentially co-emitted with CO$_2$ during fossil fuel combustion, enhancements of these and other surrogate gases are positively correlated with $C_{ff,obs}$ since emissions are strongly
correlated over the broad spatio–temporal regions to which our air samples are sensitive [Miller et al., 2012].

2.2. $C_{ff}$ prediction model

The PPR model prediction of $C_{ff}$ is $C_{ff,pred}$ where

$$C_{ff,pred} = \sum_{m=1}^{M} g_m(w_m \cdot X),$$

(3)

where $g_m$ is the $m$th ridge function, $M$ is the total number of ridge functions, $X$ is a multivariate surrogate gas measurement vector with dimension $K$, and $w_m$ is a $K$-dimensional direction vector. Given the mathematical form of the ridge functions, the $M$ direction vectors are estimated from the data. (Logan and Shepp [1975] introduced the term “ridge function” to describe a multivariate function $q$ that maps a $n$-dimensional real vector $X$ into a real scalar $q(e \cdot X)$ where $e$ is a direction vector in $\mathbb{R}^n$ ($n$-dimensional Euclidean space) and $e \cdot X$ is the inner product of $e$ and $X$.) A ridge function varies only along the direction $e$. In our study, the number of surrogate gases $K$ included in the model, the mathematical form of each ridge function, and the number of ridge functions $M$ are determined by cross-validation. In particular, we consider PPR models where the number of ridge functions varies from 1 to 4, and where each ridge function is either a smoothing spline [Craven and Wahba, 1979; Wahba, 1990] with an adjustable effective degrees-of-freedom (edf) or a supersmooth [Friedman, 1984]. PPR is attractive because of its flexibility; any function can in principle be represented with a PPR model [Diaconis and Shahshahani, 1984]. In contrast, simpler multivariate linear models can not in general represent any function. However, for observed data affected by both random
and systematic measurement errors, selecting the form and complexity of the optimal PPR model is non-trivial.

Before constructing a PPR model from any set of model building data, we center and scale measurements for each surrogate gas in the training data by subtracting its sample mean from each measurement and then dividing by the estimated standard deviation of the measurements. The validation data and test data are centered and scaled in exactly the same way.

2.3. Model Selection

Given the number of surrogate gases in the PPR model, we select the optimal combination of the 12 candidate surrogate gases (Table 1) by six-fold cross-validation— an implementation of the K-fold cross-validation method [Hastie et al., 2008]. Every sixth observation of the model building data is assigned to the validation data set. The other observations in the model building data are assigned to the training data set. For each of the six distinct ways to split the model building data into training and validation subsets, we determine a candidate PPR model from the training data and predict $C_{ff}$ values for the validation data set. For each of the six distinct splits of the model building data, we determine the mean-squared difference between measured and predicted $C_{ff}$ for the validation data. The square-root of the mean of these six values is the cross-validation statistic (Table 2). We select the identity of the surrogate gases by minimizing this cross-validation statistic (Figure 4). (Since the model building data is different for each of the four Cases, variations in both the number and identity of selected surrogate gases as well as the selected edf of the ridge functions in the PPR model are expected (Table 2).) We
select the optimal linear multivariate model and the optimal linear single proxy model according to the same cross-validation method described above.

For the data considered here, a PPR model with two ridge functions, where each has a smoothing spline form, works best according to cross-validation. In our study, for each distinct combination of candidate surrogate gases, we vary the effective degrees-of-freedom (edf) of the ridge functions on a grid. For any particular choice of surrogate gases, the optimal edf value yields the minimum cross-validation statistic. We set the maximum value of edf equal to 3 to encourage smooth ridge functions and impose regularity on the predictions. In our analysis, the number of surrogate gases in a candidate subset varies from two to six. For subsets of size seven or larger, the PPR model had convergence problems. In particular, for a subset with seven or more surrogate gases, the resulting cross-validation statistic was sensitive to the order of surrogate gases in the R code [R Development Core team, 2013] implementation of the PPR method. However, for subsets of size six or less, the variation of results due to the order of the gases was very slight.

### 2.4. Analysis

We quantify how well the PPR model determined from model building data predicts test data that are independent of the model building data for four Cases (Figure 2). In Case A, the test data are the first third of the $C_{ff,obs}$ time series. In Case B, the test data are the middle third of the time series. In Case C, the test data are the last third of the time series. In Case D, the test data includes every third observation in the time series. For Cases B and D, we make slight adjustments so that the test data have the same number of observations as Cases A and C.
For each Case, we estimate the uncertainty of the selected model, $\sigma_{\text{pred,post}}$, with a bootstrap method [Efron and Tibshirani, 1993] (see Appendix C for details). Neglecting the uncertainty associated with model selection methods leads to overly optimistic uncertainty estimates [Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Hoeting et al., 1999; Hjort and Claeskens, 2003]. Therefore, we estimate an additional component of uncertainty, $\sigma_{\text{pred,model}}$, that accounts for imperfection in our model selection method. In particular, our estimate $\hat{\sigma}_{\text{pred,model}}$ is the estimated standard deviation of the predictions corresponding to the optimal 2-gas, 3-gas, 4-gas, 5-gas and 6-gas PPR models selected by cross-validation. Our provisional estimate of the combined uncertainty of the PPR model prediction is

$$\hat{\sigma}_{\text{pred,tot}} = \sqrt{\hat{\sigma}_{\text{pred,post}}^2 + \hat{\sigma}_{\text{pred,model}}^2}. \quad (4)$$

Based on Eq. 4, we estimate the combined uncertainty of any deviation as

$$\hat{\sigma}_{\text{dev,tot}} = \sqrt{\hat{\sigma}_{\text{obs,ran}}^2 + \hat{\sigma}_{\text{obs,sys}}^2 + \hat{\sigma}_{\text{pred,tot}}^2}. \quad (5)$$

Given $\hat{\sigma}_{\text{dev,tot}}$, we test the hypothesis that the expected difference (deviation) between $C_{ff,\text{obs}}$ and $C_{ff,\text{pred}}$ is 0 ppm based on a chi-squared test statistic. To complement the hypothesis test analysis, for each test data set, we estimate a trend (and an associated confidence band) in the deviation ($C_{ff,\text{pred}} - C_{ff,\text{obs}}$) as a function of $C_{ff,\text{pred}}$ with a nonparametric smoothing method.

To illustrate the PPR approach, we show the two ridge functions for Case C (Figure 5) corresponding to the selected PPR model determined from the model building data. Given the first estimated direction vector $w_1$ and associated ridge function, we form residuals...
(observation - ridge function) and show how well the second ridge function predicts these residuals. The PPR model prediction is the sum of the two ridge functions.

3. Results and Discussion

3.1. Prediction model uncertainty

In Figures 6a, 6b, 6c we show how $\hat{\sigma}_{\text{pred,post}}$, $\hat{\sigma}_{\text{pred,model}}$, and $\hat{\sigma}_{\text{pred,tot}}$ vary with predicted $C_{ff}$ for Case D. As discussed later in Section 3.3, we expect $\hat{\sigma}_{\text{pred,post}}$ to generally increase as $C_{ff,pred}$ increases to large values because random prediction uncertainties for extreme values of $C_{ff}$ are expected to be larger than random prediction uncertainties corresponding to midrange values. The approximate j-shape variation of $\hat{\sigma}_{\text{pred,post}}$ with predicted $C_{ff}$ is consistent with this expectation (Figure 6a). More specifically, for $C_{ff,pred}$ greater than 0.5 ppm, $\hat{\sigma}_{\text{pred,post}}$ generally increases as $C_{ff,pred}$ increases. For $C_{ff,pred}$ less than 0.5 ppm, $\hat{\sigma}_{\text{pred,post}}$ generally increases as $C_{ff,pred}$ decreases. Similar remarks apply to $\hat{\sigma}_{\text{pred,tot}}$. For Cases A, B, C and D, the 0.16, 0.5 and 0.84 the quantiles of the empirical distribution of $\hat{\sigma}_{\text{pred,tot}}$ are: (0.27,0.33,0.49) ppm; (0.14,0.21,0.39) ppm; (0.17, 0.22, 0.37) ppm; and (0.15, 0.22, 0.49) ppm. Because of the j-shape dependence of $\hat{\sigma}_{\text{pred,tot}}$ on $C_{ff,pred}$, one should interpret the empirical quantiles of $\hat{\sigma}_{\text{pred,tot}}$ determined from all the test data with caution.

3.2. Model Performance

A standard way to quantify prediction model performance is to estimate the square root of the mean squared (RMS) deviation between observed and predicted values (see Appendix B for mathematical details). The estimated RMS deviation between predicted and observed $C_{ff}$ for all observations in the test data ranges from 1.06 ppm to 1.37 ppm.
for the four distinct definitions of the test data considered (Table 3). As expected, estimated RMS deviations determined from all observations in the test data are larger than corresponding RMS deviations determined for the model building data. For all four ways of defining the test data, the estimated RMS deviation for the subset $C_{ff,pred} > 2$ ppm is larger than the estimated RMS deviation for the subset $C_{ff,pred} \leq 2$ ppm (Table 3). Uncertainties for individual deviations are also larger, in general, for the $C_{ff,pred} > 2$ ppm subset (Figure 7). As discussed later in Section 3.3, one expects the random uncertainty of any prediction model to be higher for extreme $C_{ff}$ values relative to random uncertainties for mid-range $C_{ff}$ values. This may explain why estimated RMS deviations are larger for the $C_{ff,pred} > 2$ ppm subset compared to the $C_{ff,pred} \leq 2$ ppm subset. Additionally, physical effects could contribute to inflation of RMS deviations. For instance, RMS deviations may be larger for $C_{ff,pred} > 2$ ppm due to incomplete mixing of local emissions at measurement sites which would create systematic errors in $\Delta_{gas}: C_{ff}$ ($\Delta_{gas}$ is shorthand for surrogate gas enhancement) ratios for larger $C_{ff}$ values.

Although a useful performance metric, the RMS deviation statistic does not quantify the uncertainty of the PPR model prediction for any particular observation. Further, the RMS deviation statistic does not inform us if the predicted values are consistent with observations given measurement and prediction uncertainties. To determine if predictions are consistent with observations, we test the null hypothesis that the expected value of the difference between observed and predicted $C_{ff}$ is 0 ppm given our estimates of measurement and prediction uncertainties. For each of the four Cases, we determine a chi-squared goodness-of-fit statistic

$$\chi_{obs}^2 = \sum_i \frac{(C_{ff,pred}(i) - C_{ff,obs}(i))^2}{\hat{\sigma}_{dev,tot}^2(i)}, \quad (6)$$
and an associated p-value (Table 4). When the chi-squared test statistic is determined from all 94 observations in each test data set, the p-values for Cases A, B, C and D are 0.12, 0.95, 0.90 and 0.94 respectively. Since these p-values are high, the evidence for rejecting the null hypothesis is weak.

As a further check of the consistency of the prediction model with observations, we plot trend estimates determined with the LOCFIT method [Loader, 1999; Loader, 2010] for the deviation (predicted minus observed) as a function of the predicted value and associated approximate 95% confidence bands (Figure 7). In our implementation of LOCFIT, we fit a local polynomial of degree 0 and set the span parameter to 0.75. The confidence bands are determined by a parametric bootstrap method [Efron and Tibshirani, 2003]. For Cases A, B, C and D, the fraction of observations for which the approximate 95% confidence bands for the trend include 0 ppm are 0.71, 1, 1, 1.

To study possible variation of model performance with season, we compare predicted and observed $C_{ff}$ values for test data corresponding to (November-February) and (May-September) (Figures 8 and 9). For each Case, even though the proxy prediction model is determined from the complete model building data from all seasons, the visual agreement between predicted and observed values for seasonal subsets is good. For the test data, the estimated RMS deviations for the summer subset were, on average, 14 percent higher than the estimated RMS deviations for the winter subset (Table 3).

3.3. Other sources of systematic uncertainty

In general, if the fractional uncertainties of observations are relatively large and/or the total number of observations is not sufficiently large, hypothesis testing may not reveal real systematic differences between observations and model predictions. Since the un-
certainties for $\Delta^{14}$C-based $C_{ff}$ measurements analyzed in this work are typically large relative to associated individual measurement values and the range of measured values, more data (particularly measurements corresponding to larger $C_{ff}$ values) may be necessary to produce evidence for possible additional systematic uncertainties not accounted for in our current model. Below, we discuss plausible sources of such additional uncertainty.

As stated earlier, we expect each $C_{ff}$ observation ($C_{ff,obs}$) to have a non-zero bias (systematic error). Ideally, the unobserved biases of $C_{ff}$ measurements should have a mean value close to 0 ppm and vary independently from observation to observation. Since we have no empirical method to verify these modeling assumptions, it is possible that the mean systematic error of any set of $C_{ff}$ measurements may vary from 0 ppm in a scientifically significant way. To illustrate this effect, suppose that actual systematic errors for any set of measurements are independent realizations of exponentially distributed random variables with theoretical mean and standard deviation of $\hat{\sigma}_{obs,sys} = 0.5$ ppm. For this case, the expected bias for each measurement is 0.5 ppm. This would introduce a systematic error of approximately 0.5 ppm into all proxy model predictions. Worse yet, for such an exponential bias model, hypothesis testing could confirm that (biased) observations and (biased) predictions are consistent. Since our goal is to construct a confidence interval for (true) $C_{ff}$ value rather than the associated expected value of a biased measurement of $C_{ff}$, we must consider the above effect as an additional source of systematic uncertainty. Further, our estimate $\hat{\sigma}_{pred,tot}$ (Eq. 4) does not account for systematic uncertainties in measured surrogate gas in the test data due to imperfect background correction, spatial effects due to determining one PPR model from two sites rather one PPR model for each
site, systematic temporal variations of surrogate gas emission rates, and systematic effects due to incomplete mixing of emissions at measurement sites.

To account for the systematic effects above, we incorporate an additional component of uncertainty, $\hat{\sigma}_{\text{pred,extra}} = 1$ ppm, into our prediction uncertainty budget. Our decision to set $\hat{\sigma}_{\text{pred,extra}} = 1$ ppm is based on scientific judgment informed by an additional study (not presented here) of how variability in surrogate gas background estimates affects predicted $C_{ff}$ values. Hence, our final estimate of the combined uncertainty of $C_{ff}$ is $
abla\sqrt{\hat{\sigma}_{\text{pred,post}}^2 + \hat{\sigma}_{\text{pred,model}}^2 + \hat{\sigma}_{\text{pred,extra}}^2}$ (see Figure 6d). We note that an uncertainty budget with $\hat{\sigma}_{\text{pred,extra}} = 0$ ppm is sufficient to explain deviations between observed and predicted $C_{ff}$ values according to hypothesis test results (Table 4). Further, any additional systematic uncertainty might vary as a function of $C_{ff,pred}$ (or as a function of other quantities) contrary to our approach.

We note that the combined uncertainty for the prediction of $C_{ff}$ due to systematic effects is $\hat{\sigma}_{\text{pred,sys}} = \sqrt{\hat{\sigma}_{\text{pred,model}}^2 + \hat{\sigma}_{\text{pred,extra}}^2}$ which is greater than $\hat{\sigma}_{\text{obs,sys}}$ for all observations. Also, the combined uncertainty for the prediction of $C_{ff}$ due to random effects is $\hat{\sigma}_{\text{pred,post}}$ which is generally less than $\hat{\sigma}_{\text{obs,ran}}$ (Figure 6a). For Case D, for 77 of 94 observations, the combined uncertainties (due to random and systematic effects) for $C_{ff}$ predictions are slightly less than the uncertainty of any $\Delta^{14}$C-based measurement of $C_{ff}$ (1.1 ppm) (Figure 6d). This phenomenon is possible because the random uncertainty of any model prediction decreases as the number of observations $n_{\text{obs}}$ that the model is determined from increases. In our case the PPR model is determined from 174 observations where each observation is a vector where one component is observed $C_{ff}$ and the other components are observed surrogate gas enhancements. Hence, for our problem, if the component of
prediction uncertainty due to systematic effects is sufficiently small relative to 1.1 ppm and \( n_{\text{obs}} \) is sufficiently large, and the random uncertainty of surrogate gas enhancements that are input into the PPR prediction model are sufficiently small, the phenomenon is plausible. Moreover, theory predicts this phenomenon for idealized univariate regression models where there are no systematic errors [Mendenhall and Cincich, 1992]. For this idealized case, the width of a confidence interval for the true value (predicted by the fitted regression line) shrinks towards 0 as the number of data points that the regression model is determined from increases without limit. Further, theory predicts wider confidence intervals for extreme observations relative to mid-range observations. In general, one expects the the random uncertainty of any prediction model (including our PPR proxy model prediction) to be larger for extreme observations relative to mid-range observations.

In Figure 6a, the approximate j-shape relationship between the random uncertainty of predicted \( C_{ff} \) and predicted \( C_{ff} \) may result because the distribution of unobserved \( C_{ff} \) is skewed toward lower values.

As discussed earlier, for each of the four Cases of interest, we select PPR predictions models based on model building data acquired at both sites. We expect prediction model performance to be best when model building data is acquired at the same site where we predict \( C_{ff} \). Moreover, prediction model performance could vary with the time of surrogate gas measurements and depend on the time period corresponding to when model building data is acquired. Such temporal variations could be diurnal, weekly, seasonal or long term. Long term variations are highly likely because of emissions of discontinued refrigerants that change in time, or relative emission rates that that vary due to changes in industrial, technological or energy production processes or regulations. Hence, over
sufficiently long time periods, one should recalibrate any surrogate gas prediction model to avoid degradation of model performance. A simple form of recalibration would be to split up the model building data into (nonoverlapping) blocks according to time. One would then determine a PPR model for each block. We note that systematic temporal variability effects could introduce temporal trends in deviations between predicted and observed values of $C_{ff}$. However, for all four cases, when we fit linear trend models to deviations from both sites, the evidence for trends is weak. For Cases A, B, C and D the estimated slopes and associated uncertainties (in parentheses) are -0.01(0.21) ppm/year, 0.14(0.18) ppm/year, 0.08(0.22) ppm/year and 0.02(0.07) ppm/year respectively. Further, model performance for Case C (where the PPR model is extrapolated into the future, and model performance for Case D (where the model building data in interwoven with the test data) are similar according to RMS deviation statistics (Table 3) and p-values from hypothesis testing (Table 4).

3.4. Alternative proxy models

We also predict $C_{ff}$ with a standard linear multivariate prediction model. For all cases considered, this linear model yields uniformly higher cross-validation statistics compared to the PPR model (see Figure 4 for an example). For Case B, where cross-validation selects CO as the best surrogate gas for a single proxy linear proxy model, the cross-validation statistics for the optimal single proxy linear model, the optimal multivariate linear model, and the optimal PPR model are 1.33 ppm, 1.19 ppm and 1.08 ppm respectively. For this same case, a single proxy model based on CO yields RMS deviation estimates (and approximate 95 percent confidence interval) of 1.23(1.00,1.53) ppm and 1.39(1.00,2.09) ppm for the subsets corresponding to May to September and November to February.
In contrast, for the same subsets, the PPR model yields 1.11 (0.89,1.44) ppm and 1.06 (0.76,1.45) ppm.

Although cross-validation statistics serve a critical role for selecting the optimal prediction model, they do not provide a clear picture of how the prediction models vary with respect to one another. To get some insight into this variation, we quantify the difference between the single proxy and multivariate models with respect to the PPR model for the Case B test data (Figure 10). The deviation between the single proxy and PPR prediction models is most dramatic for large PPR prediction values. In general, the multivariate model prediction for $C_{ff}$ is less than the PPR model prediction for cases where the PPR prediction is greater than approximately 1 ppm.

In addition to the multivariate linear and PPR prediction models, we also considered a multivariate adaptive regression spline (MARS) model [Friedman, 1991]. According to our cross-validation criterion, the PPR model outperformed the MARS for all cases considered. In some applications, researchers fit PPR models to transformed rather than raw observations. We explored this approach by fitting a PPR model to $\log(\alpha + C_{ff,obs})$ values where $\alpha = 4$ ppm. (We require a positive $\alpha$ because some $C_{ff,obs}$ values are negative.) According to our cross-validation criterion, this approach under-performed the approach presented in this work.

4. Summary

For each of four ways of defining test data, we selected a PPR proxy prediction model by cross-validation based on joint analysis of data from NOAA aircraft profiling sites CMA and NHA. The estimated RMS difference between predicted and observed $C_{ff}$ for the test data excluded from the model building process ranged from 1.06 ppm to 1.37
ppm depending on how the test data was defined (Table 3). We also quantified prediction uncertainty for each observation in each test data set (Figure 6). We tested the hypothesis that predicted and observed values are consistent given their associated uncertainties (Table 4). Since evidence of systematic deviations between observed and predicted $C_{ff}$ based on p-values from hypothesis tests is weak, we conclude that development of a scientifically useful multivariate proxy prediction model for $C_{ff}$ is a realistic goal.

In future studies, we plan to expand $C_{ff}$ estimates beyond those based on $\Delta^{14}$C measurements, by applying our methods to NOAA GGGRN aircraft and tower sites in the US. To account for spatial and temporal variations of emission ratios, prediction models would be determined at each site. At each new site, to suppress effects due to systematic temporal variations of surrogate gas emission rates, we currently plan to determine PPR prediction models for contiguous two-year long blocks of model building data. For sites that lack $\Delta^{14}$C measurements, development of proxy prediction models based on local model building data is not possible. For such cases, one might predict $C_{ff}$ based on proxy models developed at other sites with similar $\Delta_{gas} : C_{ff}$ ratios. How well such an approach would work is a research question. At many sites, we expect the range of $C_{ff}$ to be larger than for the measurements analyzed in this work. For such sites, on average, we expect lower fractional uncertainties of $\Delta^{14}$C-based measurements of $C_{ff}$ and predictions of $C_{ff}$ with lower fractional uncertainties.

Additional $C_{ff}$ values determined from surrogate gases ($C_{ff, pred}$) may facilitate a more accurate estimate of US national fossil fuel emissions determined by an atmospheric inverse approach (e.g. [Basu et al., 2016]). The utility of proxy model predictions that we plan to acquire at other sites will depend on their yet to be determined uncertainties.
which we expect to differ from what we determined for the NHA-CMA sites. In future studies, we expect that high resolution atmospheric transport modeling along with more observations from aircraft vertical profiles and upwind sampling locations at points of continental inflow, such as the Pacific and Gulf coasts, may yield improved determinations of background values for both $\Delta^{14}C$ and surrogate gases. Further, improved representation of heterotrophic respiration and nuclear reactor emissions will also help us reduce and better characterize systematic uncertainties. Because systematic errors can never be eliminated from either $C_{ff}$ or $\Delta_{gas}$ measurements, high precision, low bias measurements of $\Delta^{14}C$ remain critical for improving determination of US fossil fuel emissions based on additional proxy model predictions of $C_{ff}$.

**Acknowledgements.** We thank A. Possolo, A. Pintar, D. Samarov and J.C. Turnbull for helpful comments, the NOAA Climate Program Office AC4 for funding the research that produced the measurements analyzed here, and C. Siso, A. Karion and others from NOAA and CIRES for technical and logistical contributions related to sampling and analysis of flasks. Contributions to this work by staff of NIST (an agency of the US government) are not subject to copyright in the US and represent an execution of the NIST Greenhouse Gas and Climate Science Measurements Program (Special Programs Office, Associate Director for Laboratory Programs, NIST, U.S. Dept. of Commerce). Data and software scripts are available in supplementary information.

5. Appendix A

Below, we summarize the theoretical model for $C_{ff}$ and its associated empirical estimate (Eq. 2) presented in [Miller et al., 2012]. Following Turnbull et al. [2006], the theoretical CO$_2$ mole fraction is modeled as due to three sources: background, local fossil fuel emis-
sions, and biospheric sources. This model neglects possible effects due to emissions from nuclear power plants. Given the unobserved true values of $C$, $\Delta$, $C_{bg}$, $C_{bio}$, $\Delta_{bg}$, and $\Delta_{bio}$, one gets the following model equations

$$C = C_{bg} + C_{ff} + C_{bio},$$

(7)

and

$$\Delta C = \Delta_{bg}C_{bg} + \Delta_{ff}C_{ff} + \Delta_{bio}C_{bio}.$$  

(8)

We note that Eq. 8 is valid for cases where $\Delta$ is small [Vogel et al., 2013]. Since the CMA and NHA sites are far from major sources of fossil-fuel CO$_2$ emissions, we expect negligible systematic error associated with the Eq. 8 modeling assumption. We decompose $C_{bio}$ into the sum of a photosynthetic term, $C_{photo}$ and a respiratory term $C_{resp}$. Further, we assume that $\Delta_{photo} = \Delta_{bg}$. Based on Eqns. 7 and 8, we estimate $C_{ff}$ as $C_{ff,obs}$ where

$$C_{ff,obs} = \frac{C_{obs}(\Delta_{obs} - \Delta_{bg,obs})}{\Delta_{ff} - \Delta_{bg,obs}} - \hat{C}_{resp}(\hat{\Delta}_{resp} - \Delta_{bg,obs})/\Delta_{ff} - \Delta_{bg,obs},$$

(9)

and the estimates $C_{obs}, \Delta_{obs}, \Delta_{bg,obs}$ are determined by measurement, and the estimates $\hat{C}_{resp}$ and $\hat{\Delta}_{resp}$ are determined by a synthesis of experimental measurement and computational modeling. The second term on the right hand side of Eq. 9 is the correction term, $\hat{C}_{cor}$ that appears in Eq. 2. This term typically takes values in the range 0.4 ppm to 0.8 ppm during the summer and 0.2 ppm to 0.3 ppm during the winter [Lehman et al., 2013].
6. Appendix B

In this work, we report estimates of theoretical (true) root-mean-square (RMS) deviations between measured and predicted $C_{ff}$ and associated confidence intervals for the unknown theoretical RMS deviation. Since measured $C_{ff}$ is a function of isotopic ratio measurements, the “true” RMS deviation is undefined if one assumes the minor and major isotope measurements are realizations of Poisson random variables. This is so because the expected value of the ratio of two Poisson random variables is infinite \cite{Coakley2005, Coath2013}. To ensure that “true” RMS deviation is well defined and reported confidence intervals for $C_{ff}$ are sensible, we restrict analysis to the subsample of $C_{ff}$ measurements where denominator terms in isotopic ratios are positive, and $\Delta_{bg,obs} \neq \Delta_{ff}$. Our subsampling restriction also ensures that the expected value and theoretical standard deviation of $C_{ff,obs}$ are well-defined. For $C_{ff}$ studies, this subsampling restriction has no practical effect on data acquisition since the probability that a $C_{ff}$ measurement falls outside the subsample where the “true” RMS deviation is defined is negligible.

7. Appendix C

The PPR model prediction for $C_{ff}$ values in the test data depends on input surrogate gas measurements in the test data, $S_T$, and the model building data $D_M$ that determines the PPR model. Each observation in the model building data is a vector of dimension K+1 where one component is a $C_{ff,obs}$ value and the other K components are surrogate gases. Suppose we write the PPR prediction for the $i$th observation in the test data as

$$C_{ff,pred}(i) = f(D_M, S_T(i)),$$

(10)
where \( f \) denotes the PPR model. To get \( \hat{\sigma}_{\text{pred,post}} \) our bootstrap method simultaneously resamples \( D_M \) and \( S_T(i) \) by nonparametric and parametric bootstrap methods respectively. The nonparametric bootstrap resampling scheme for \( D_M \) accounts for observation-to-observation variation of systematic measurement errors and random uncertainties in both \( C_{ff} \) and surrogate gas measurements. The parametric bootstrap resampling scheme for \( S_T(i) \) accounts for random uncertainties but not observation-to-observation variation of systematic measurement errors. Our method does not account for the possibility that the mean systematic measurement error over all observations may not be 0 ppm. This additional uncertainty is accounted for with a method described in Section 3.3.

References


Graven, H.D. and Gruber N (2011), Continental-scale enrichment of atmospheric 14CO2 from the nuclear power industry: potential impact on the estimation of fossil fuel-derived


Table 1. Surrogate gases.

<table>
<thead>
<tr>
<th>Trace Gas index</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CO</td>
</tr>
<tr>
<td>2</td>
<td>SF₆</td>
</tr>
<tr>
<td>3</td>
<td>HFC-134a</td>
</tr>
<tr>
<td>4</td>
<td>HCFC-22</td>
</tr>
<tr>
<td>5</td>
<td>HCFC-142b</td>
</tr>
<tr>
<td>6</td>
<td>C₂Cl₄</td>
</tr>
<tr>
<td>7</td>
<td>CH₂Cl₂</td>
</tr>
<tr>
<td>8</td>
<td>HFC-152a</td>
</tr>
<tr>
<td>9</td>
<td>C₆H₆</td>
</tr>
<tr>
<td>10</td>
<td>CH₄</td>
</tr>
<tr>
<td>11</td>
<td>N₂O</td>
</tr>
<tr>
<td>12</td>
<td>CFC-11</td>
</tr>
</tbody>
</table>

Table 2. Selected surrogate gases for PPR prediction model and observed RMS deviation (observed - predicted) for training and validation data. For each of the four Cases, we select a corresponding optimal PPR model by minimizing our cross-validation statistic which is the estimated root-mean-square deviation between predicted and observed values in the validation data (third column in this table).

<table>
<thead>
<tr>
<th>CASE</th>
<th>RMS deviation training data (ppm)</th>
<th>RMS deviation validation data (ppm)</th>
<th>selected gases</th>
<th>selected edf of ridge functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.927</td>
<td>1.012</td>
<td>(2, 3, 4, 5, 9, 11)</td>
<td>2.82</td>
</tr>
<tr>
<td>B</td>
<td>1.027</td>
<td>1.081</td>
<td>(1, 3, 4, 6, 9)</td>
<td>2.95</td>
</tr>
<tr>
<td>C</td>
<td>1.079</td>
<td>1.106</td>
<td>(4, 6, 7, 9, 11)</td>
<td>2.66</td>
</tr>
<tr>
<td>D</td>
<td>1.051</td>
<td>1.096</td>
<td>(4, 6, 9)</td>
<td>2.90</td>
</tr>
</tbody>
</table>
Table 3. For the PPR model, estimates and approximate 95% bootstrap confidence intervals for the root-mean-square value of $C_{ff,pred} - C_{ff,obs}$ for selected subsets of both model building and test data for four cases. In Case A, test data are first third of time series. In Case B, test data are middle third of time series. In Case C, test data are last third of time series. In Case D, every third observation is in the test data. Results are in units of μmole/mole (ppm). In this study, $\hat{\sigma}_{pred,extra} = 0$.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Case A</th>
<th>Case B</th>
<th>Case C</th>
<th>Case D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Building Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{C}_{ff,PPR} &lt; 2$ ppm</td>
<td>0.87 (0.78,0.97)</td>
<td>0.99 (0.89,1.12)</td>
<td>1.04 (0.92,1.18)</td>
<td>0.98 (0.86,1.13)</td>
</tr>
<tr>
<td>$\hat{C}_{ff,PPR} \geq 2$ ppm</td>
<td>1.11 (0.89,1.42)</td>
<td>1.18 (0.88,1.59)</td>
<td>1.28 (0.98,1.66)</td>
<td>1.31 (1.04,1.66)</td>
</tr>
<tr>
<td>all data</td>
<td>0.93 (0.84,1.05)</td>
<td>1.03 (0.93,1.18)</td>
<td>1.09 (0.98,1.22)</td>
<td>1.06 (0.95,1.20)</td>
</tr>
<tr>
<td>November to February</td>
<td>0.97 (0.78,1.23)</td>
<td>0.85 (0.68,1.17)</td>
<td>0.94 (0.77,1.15)</td>
<td>0.89 (0.70,1.14)</td>
</tr>
<tr>
<td>May to September</td>
<td>0.95 (0.80,1.14)</td>
<td>1.14 (0.99,1.35)</td>
<td>1.22 (1.01,1.45)</td>
<td>1.19 (1.02,1.41)</td>
</tr>
<tr>
<td>Test Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{C}_{ff,PPR} &lt; 2$ ppm</td>
<td>1.25 (1.06,1.49)</td>
<td>0.94 (0.82,1.10)</td>
<td>0.98 (0.86,1.12)</td>
<td>0.95 (0.83,1.08)</td>
</tr>
<tr>
<td>$\hat{C}_{ff,PPR} \geq 2$ ppm</td>
<td>1.73 (1.31,2.68)</td>
<td>1.39 (1.03,1.87)</td>
<td>1.41 (0.94,2.29)</td>
<td>1.51 (1.11,2.09)</td>
</tr>
<tr>
<td>all data</td>
<td>1.37 (1.19,1.66)</td>
<td>1.06 (0.93,1.25)</td>
<td>1.07 (0.93,1.31)</td>
<td>1.12 (0.97,1.35)</td>
</tr>
<tr>
<td>November to February</td>
<td>1.09 (0.90,1.33)</td>
<td>1.06 (0.76,1.45)</td>
<td>0.95 (0.63,1.39)</td>
<td>1.23 (0.75,2.18)</td>
</tr>
<tr>
<td>May to September</td>
<td>1.51 (1.24,2.03)</td>
<td>1.11 (0.89,1.44)</td>
<td>1.15 (0.96,1.48)</td>
<td>1.13 (0.92,1.36)</td>
</tr>
</tbody>
</table>
Table 4. Chi-square goodness-of-fit statistic $\chi^2_{obs}$ (Eq. 6), degrees-of-freedom (df), and associated p-values for testing the hypothesis that the observed and predicted test data are consistent.

<table>
<thead>
<tr>
<th>Case</th>
<th>$\chi^2_{obs}$</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cape May, NJ data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case A</td>
<td>71.5</td>
<td>56</td>
<td>0.08</td>
</tr>
<tr>
<td>Case B</td>
<td>43.5</td>
<td>58</td>
<td>0.92</td>
</tr>
<tr>
<td>Case C</td>
<td>42.3</td>
<td>47</td>
<td>0.67</td>
</tr>
<tr>
<td>Case D</td>
<td>46.3</td>
<td>51</td>
<td>0.66</td>
</tr>
<tr>
<td>Portsmouth, NH data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case A</td>
<td>38.6</td>
<td>38</td>
<td>0.44</td>
</tr>
<tr>
<td>Case B</td>
<td>29.0</td>
<td>36</td>
<td>0.79</td>
</tr>
<tr>
<td>Case C</td>
<td>34.7</td>
<td>47</td>
<td>0.91</td>
</tr>
<tr>
<td>Case D</td>
<td>27.5</td>
<td>43</td>
<td>0.97</td>
</tr>
<tr>
<td>all data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case A</td>
<td>110.1</td>
<td>94</td>
<td>0.12</td>
</tr>
<tr>
<td>Case B</td>
<td>72.5</td>
<td>94</td>
<td>0.95</td>
</tr>
<tr>
<td>Case C</td>
<td>77.1</td>
<td>94</td>
<td>0.90</td>
</tr>
<tr>
<td>Case D</td>
<td>73.7</td>
<td>94</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Figure 1. Example vertical profiles (a) above site CMA (offshore from Cape May, NJ) from Feb. 21, 2007 and (b) above site NHA (offshore from Portsmouth, NH) from July 10, 2008. Black, red and blue pluses connected with a line represent nine air samples collected between the surface and 8 km asl, with measured mole fractions shown for CO₂, CO, and HFC-134a, respectively. Brown circles represent values from the three typical altitudes where Δ¹⁴C is analyzed. Grey bars highlight the nominal altitudes at which all gases were measured. (reprinted figure from [Miller et al. 2012])
Figure 2. Observed $C_{ff}$ time series. We split the data into a model building part and a test part. The model building part is further split into training and validation data for model selection. For Case A, the test data corresponds to data acquired before the time associated with the first dashed line. For Case B, the test data corresponds to data acquired between the two solid lines. For Case C, the test data corresponds to data acquired after the second dashed line. For Case D, every third observation is in the test data.
Figure 3. Scatterplots of $\Delta^{14}$C-based measurements of $C_{ff}$ and measured surrogate gas enhancements selected for inclusion in a PPR prediction model for at least one of the four Cases. Above, ppb and ppt denote mole fractions equal to $10^{-9}$ and $10^{-12}$ respectively.
Figure 4. For Case B, we show cross-validation statistics corresponding to optimal multivariate linear model and the optimal nonlinear Projection Pursuit Regression models as a function of the number of surrogate gases in the prediction model.
For Case C the selected gases for the Projection Pursuit Regression model are HFC-22, C\textsubscript{2}Cl\textsubscript{4}, CH\textsubscript{2}Cl\textsubscript{2}, C\textsubscript{6}H\textsubscript{6} and N\textsubscript{2}O. The direction vectors for the two ridge functions are $w_1 = (0.937, 0.118, -0.152, 0.277, -0.094)$ and $w_2 = (-0.903, 0.311, -0.028, 0.181, 0.232)$.

(a) first ridge function (solid line) and observed $C_{ff}$. (b) second ridge function (solid line) and residual about first ridge function (observed $C_{ff}$ minus first ridge function). (c) scatterplot of projections of surrogate gas measurements. (d) observed versus predicted $C_{ff}$ and line of equality. Prediction is sum of the two ridge functions.
Figure 6. Components and combined components of uncertainty of predicted $C_{ff}$ as a function of predicted $C_{ff}$ for Case D. (a): $\hat{\sigma}_{\text{pred,post}}$. (b): $\hat{\sigma}_{\text{pred, model}}$. (c): $\sqrt{\hat{\sigma}_{\text{pred,post}}^2 + \hat{\sigma}_{\text{pred, model}}^2}$. (d): $\sqrt{\hat{\sigma}_{\text{pred,post}}^2 + \hat{\sigma}_{\text{pred, model}}^2 + \hat{\sigma}_{\text{pred, extra}}^2}$ where $\hat{\sigma}_{\text{pred, extra}} = 1$ ppm. The dashed horizontal reference line corresponds to 1.1 ppm. Results for the Cases A, B and C are similar.
Figure 7. Deviations (predicted-observed) versus predicted $C_{ff}$ for test data. (a) Case A. (b) Case B. (c) Case C. (d) Case D. For each deviation, $C_{ff,pred} - C_{ff,obs}$, we show an associated approximate 68% confidence interval ($C_{ff,pred} - C_{ff,obs} - \hat{\sigma}_{dev,tot}, C_{ff,pred} - C_{ff,obs} + \hat{\sigma}_{dev,tot}$). Solid and dashed lines correspond to trend estimates and associated approximate 95% confidence bands. For Cases A,B,C,D, the confidence band for the trend estimates includes 0 ppm for the following fraction of observations: 0.71, 1, 1 and 1. Above, $\hat{\sigma}_{pred,extra} = 0$ ppm.
Figure 8. Observed $C_{ff}$ versus proxy model prediction of $C_{ff}$ for subset of test data corresponding to November through February. (a): Case A definition of test data. (b): Case B definition of test data. (c): Case C definition of test data. (d): Case D definition of test data. Line of equality shown in black. RMS deviations between observed $C_{ff}$ and predicted $C_{ff}$ and associated approximate 95 percent confidence intervals for Cases A,B,C and D are (in ppm units); 1.09 (0.90,1.33) 1.06 (0.76,1.45) 0.95 (0.63,1.39) 1.23 (0.75,2.18) Above, $\hat{\sigma}_{pred,extra} = 0$ ppm.
Figure 9. Observed $C_{ff}$ versus proxy model prediction of $C_{ff}$ for subset of test data corresponding to May through September. (a): Case A definition of test data. (b): Case B definition of test data. (c): Case C definition of test data. (d): Case D definition of test data. Line of equality shown in black. RMS deviations between observed $C_{ff}$ and predicted $C_{ff}$ and associated approximate 95 percent confidence intervals for Cases A,B,C and D are (in ppm units): 1.51 (1.24,2.03), 1.11 (0.89,1.44), 1.15 (0.96,1.48), and 1.13 (0.92,1.39). Above, $\hat{\sigma}_{pred,extra} = 0$ ppm.
Figure 10. Test data Case B. (a): selected surrogate gases in single proxy prediction model is CO. (b): multivariate proxy model includes 3 surrogate gases: HFC-22, C₂Cl₄, and C₆H₆. (c): difference between single proxy and PPR prediction models. (d): difference between multivariate and PPR prediction models. Uncertainty intervals not shown.