
To: Patrick Barickman
From: Jon Wilkey
Date: February 29, 2016
RE: Monthly report on integrating equipment-based emission factors into R model

During the month of February, work on the R model has focused on (a) updating the UDAQ copy of the R model (to include the changes made to the R model between delivery in Oct. 2015 and the start of this project), (b) importing data from the OGEI database, and (c) preliminary work on the data analysis and simulation of equipment-based emission factors. Each of these items are discussed briefly below.

Updating the UDAQ copy of the R Model

At the conclusion of the previous R modeling project, a final copy of the model was delivered to UDAQ through Bitbucket (a Git version control repository service), and ownership of the repository was transferred to UDAQ. Once work started on incorporating equipment-based emission factors, a decision had to be made to either (a) update the UDAQ repository (so that it included important changes that had been made to the model during Q4 2015) or (b) create a new repository. The group decided that creating a new repository was the best option. The new repository (https://bitbucket.org/wilkey/ub_o-g_emissions) was created, based on the most recent version of the R model used in my thesis¹.

Unrelated components in the model (jobs, taxes, water usage, etc.) were stripped out, download and installation instructions were updated, and UDAQ staff were granted permissions to read and write to the new repository.

Importing data from the OGEI database

I received sample copies of the OGEI database and inventory workbooks at the beginning of February. Initially, I attempted to use a package in R called “RSQLite” to work with each of the tables in the sample OGEI.db file directly. However due to an unknown issue with either RSQLite or the structure of the tables in OGEI.db, I was unable to read from the tables directly without creating serious data errors. For example, when reading from the “apis” table in OGEI.db, R incorrectly interpreted the API numbers in the “api” column so that instead of reading “43-047-52863” it would read “9767844”. Some of the errors may stem from mixed types of variables being located in the same column (e.g. the “api” column is listed as being an integer in the database but includes character hyphens in some of the API number entries). With additional development time, I’m certain that those bugs could be ironed out and connecting directly to the database file would work. However I found that it was faster (for now) to simply switch to importing CSV copies of each table (exported from the OGEI.db using an SQLite browser program). This approach has the downside of requiring an additional user step (i.e. generating the *.csv files), but the result is error-free.

Preliminary data analysis and simulation work

With the data from the OGEI.db file in hand, the majority of my time this month was spent on determining the best method for including equipment-based emission factors in the model’s Monte-Carlo simulation. Originally, I had intended on using the emissions calculated by the formulas in the

¹ Wilkey, J. 2016. “Economic and Environmental Impacts of Oil and Gas Development on the Uinta Basin.” University of Utah.

workbook and database to create relationships between emission sources their total emissions. For example, emissions from well completions (E_{WC} in ton/yr), are calculated in the workbook based on how much diesel fuel is consumed during completion:

$$E_{WC} = \frac{D \cdot EF \cdot (1 - R)}{2000}$$

where D is annual diesel usage (gal/yr), EF is an emissions factor constant (lb/gal), and R is the control percentage (from flaring). The user inputs in the workbook calculation are D and R . Following my original plan, well completion emissions would have been analyzed and simulated by:

1. Creating a cumulative distribution function (CDF) for E_{WC} (based on the results calculated in the workbook/database table).
2. Randomly drawing from that CDF to determine well completion emissions for each well that was completed during the Monte-Carlo simulation.

This effectively folds the two user inputs (D and R) into a single CDF, leveraging the work that has already been put into the workbook and creating a simpler calculation procedure in the R model. However it also eliminates the detailed information that has been collected in the database about both D and R , which in turn reduces the number of ways in which the model's inputs and options can be manipulated and/or investigated by the user. Additionally, there are some equipment types (such as RICE and turbines) for which it would be inappropriate to assume a continuous range of emissions.

Given those drawbacks, I decided to use the same emissions calculation procedure laid out in the workbook in the R model. Under this approach, the algorithm for analyzing the data and simulating emissions for each equipment type is to:

1. Create a CDF for each user input in the workbook for that piece of equipment
2. Randomly draw from each of those CDFs for each well in the Monte-Carlo simulation
3. Calculate the resulting emissions using the equations provided by the workbook

In addition to the advantage of retaining all of the available input information, this approach also allows for verification of the code written in R (by directly comparing emissions results in R with those from the workbook).

At present, functions have been written in R for the calculation of emissions from each of the equipment types listed in the workbook. In order to use these functions, the data analysis steps for each of the user inputs in the workbook must be written. Work in the month of March is expected to focus primarily on that task.

To: Patrick Barickman, Whitney Oswald, Brock LeBaron
From: Jon Wilkey
Date: March 31, 2016
RE: March 2016 monthly report on integrating equipment-based emission factors into the R model

During the month of March, work on the R model has focused on analyzing the information contained in each of the OGEI database tables. The goal of this data analysis work was to produce all of the information necessary to enable a Monte-Carlo simulation of the inputs required for calculating emissions from the equipment types in UDAQ's emissions inventory workbooks. That task was completed, and a full report on the results of that analysis is attached in the file "OGEI Data Analysis" (as both an *.html and *.pdf file).

To summarize the contents of that report, a method has been developed to analyze all of the unique combinations of the input variables contained in the OGEI database by:

1. Finding all unique combinations of the necessary input variables in each table
2. Counting the frequency with which each combination is repeated
3. Determining the cumulative probability of each combination occurring

The resulting cumulative probability table (CPT) contains one row for each combination of inputs along with the cumulative probability of that combination occurring. The CPT can then be combined with the existing Monte-Carlo structure of the R model (randomly picking a number between 0 and 1) to select a set of inputs for calculating equipment-based emissions. Like other parts of the R model, the script for generating the CPTs (and the OGEI Data Analysis report itself) are fully automated, and can be rerun as the OGEI database is updated.

With the data analysis work complete, I believe I am on track for delivery of the finished R model to UDAQ by the end of the month of April. Over the coming month I'm planning on building the necessary Monte-Carlo simulation functions for randomly drawing from the CPT tables. After connecting those functions to the previously written functions for calculating equipment-based emissions, this project will be completed. The User Manual documentation will be updated to reflect the changes made over the October 2015 version of the model. Please let me know if UDAQ staff would like to have an in-person training session to cover these changes.

In addition to the OGEI database analysis work, I also devoted some time to engaging Doug Henderer (Newfield Corporate Air Quality Manager) regarding model predictions for drilling and production over the next five years. I updated and ran the model to produce predictions for the 2016 – 2020 time period. Doug and I disagreed over the model's projections about future drilling activity over that time period. Ultimately, I believe we concluded that we're each tracking different things. Doug was tracking how many new wells had been started (or spudded), while in the R model I track when wells first start producing. We agreed that there is a time delay between these two events, and I performed a regression analysis to determine what the best time delay was between energy prices and when a well first produces. At present, the model estimates the number of wells drilled using the equation:

$$W_t = a \cdot OP_{t-1} + b$$

where W is the number of wells that first produce in month t , OP is oil price, and a and b are fitted coefficients. However stated more generally, the model is:

$$W_t = a \cdot OP_{t-x} + b$$

where x is the time delay between past oil price and the number of wells that produce for the first time. In the regression analysis, I searched for the optimal value of x over the interval of 1 to 24 months, finding that the best fit (as measured by R^2 values) was achieved with $x = 3$ months. This raised the R^2 value of the oil price drilling model from ~ 0.70 to 0.74 . However the revised fit does not substantially change the predicted drilling activity over the 2016 – 2020 period (which is ultimately driven by EPA expectations of returning to $\sim \$60/\text{bbl}$ oil within the next few years). If requested I'd be happy to investigate this particular issue further.

To: Patrick Barickman, Whitney Oswald, Brock LeBaron
From: Jon Wilkey
Date: April 28, 2016
RE: April 2016 monthly report on integrating equipment-based emission factors into the R model

During the month of April, work on the R model has focused on the Monte-Carlo (MC) simulation of emissions using the equipment-based emissions calculation functions and OGEI database cumulative probability tables (CPT), developed in February and March, respectively. See attached for a summary of the MC simulation algorithm. The output of the simulation gives total emissions for each of the following (all as a function of time):

- Each species (PM10, PM25, SO_x, NO_x, VOCs, CO, and CH₂O)
- Each equipment category (well completions, RICE & turbines, separators and heaters, dehydrators, tanks, truck loading, pneumatic controllers, pneumatic pumps, and fugitive emissions)
- Each iteration of the Monte-Carlo simulation

At present, the model calculates (for each pollutant species) total emissions, the quantiles of the MC simulation results, and the % contribution of each type of equipment to the total emissions. A sample output (with comparison to the previous activity-based emissions results) is shown in Figure (1).

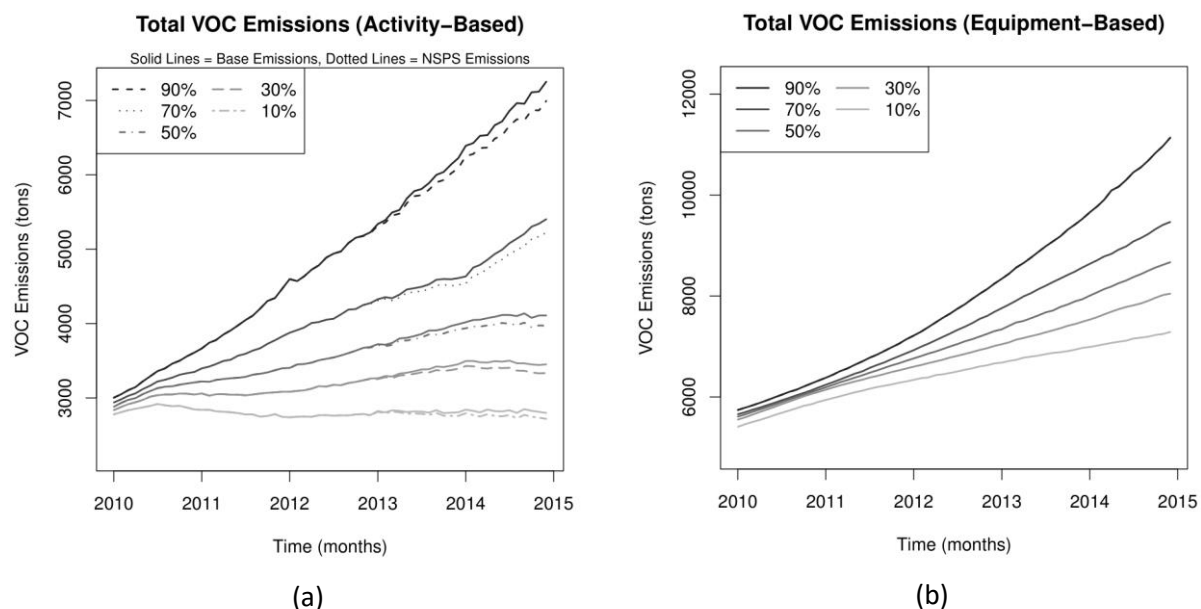


Figure 1: Total VOC emissions over the 2010 – 2014 period using 10³ MC iterations according to (a) activity-based emissions factors and (b) equipment-based emissions calculations. VOC emissions in (a) are primarily from well completions (which leads to the somewhat rough line pattern), gas transmission, and gas production. VOC emissions in (b) are primarily due to pneumatic controllers (30%), fugitive emissions (27%), pneumatic pumps (20%), and dehydrators (17%).

Given the number of factors available (7 species, 9 types of equipment, new and existing wells, each specified quantile of MC simulation output, monthly/annual/etc. time series aggregation) there are many possible ways that the data could be visualized or compared. I'd be happy to code any type of plot or comparison that UDAQ would like to get out of the model, or to prepare the results data for export to a different data visualization tool. Please let me know what type of post-processing options you'd like to use.

Aside from implementing post-processing options, I'm happy to report that at this point equipment-based emission factors have been fully integrated into the R model. Tasks (1) – (3) in the work plan have been completed, and the only remaining deliverables are to (a) update the documentation in the User Manual and (b) provide training to UDAQ staff. Model documentation should be updated by May 6, 2016. Please let me know how you'd like to cover training on the model changes.

MC Simulation Algorithm

The algorithm applied in the MC simulation is as follows:

1. Randomly draw a row from each equipment type's CPT table (e.g. well completion, RICE & turbines, tanks, etc.) for each well.
2. For each well and each type of equipment, calculate the monthly emissions for each pollutant species applicable to that type of equipment following the same procedure applied in the OGEI spreadsheet.
3. Create a production "identity" matrix with i rows (one for each well) and j columns (one for each time step in the simulation). If well i is in operation (i.e. producing either oil or gas) at time j , then the value at $[i, j] = 1$, otherwise it equals 0.
4. For each type of equipment and each pollutant species, multiply the vector of calculated emissions (containing one element for each well) by the identity matrix element-wise. This results in one matrix for the emission of each pollutant species for each type of equipment, with one row for each well and one column for each time step.
5. Calculate total emissions by species
 - a. Add together the emissions matrices from all types of equipment for the same species (e.g. PM10 emissions = (Well completion PM10) + (RICE & Turbine PM10) + (Separator & Heater PM10)).
 - b. Calculate the column sums of the combined matrix to find total emissions across all wells as a function of time.
6. Calculate the % contribution of each type of equipment to each species
 - a. Divide the column sums of the emissions matrix for each type of equipment by the result from Step (5.b) above.

Notes:

- Well completions emissions are applied once per well (whenever a well is originally drilled or reworked). Well completion emissions for new wells and reworked wells are assumed to be the same.
- All emissions are constant in time, with the following exception:
 - Truck loading emissions are a function of oil production volume. Instead of calculating an emissions vector for truck loading in Step (2) for truck loading, an emissions factor for each well is calculated, and that vector is multiplied element-wise by the oil production matrix.
 - I tried to also make tank emissions a function of oil production volume by randomly drawing from the ratio of (VOC emissions from tanks) / (oil throughput) in the OGEI database and multiplying that ratio by the oil production matrix. However there appears to be an inverse relationship in the ratio, see Figure (2). Low throughput tanks have high emissions, while high throughput tanks have low emissions (perhaps the low volume tanks are older and leak more?). Therefore random selection in the MC simulation can combine high emissions ratios with high production wells and return unrealistically large VOC emissions results. As a workaround, the code randomly draws emissions results from tanks directly from the CPT. Tank emissions are therefore both constant and are

not a function of oil throughput, but are representative of VOC emission rates for tanks in the Uinta Basin.

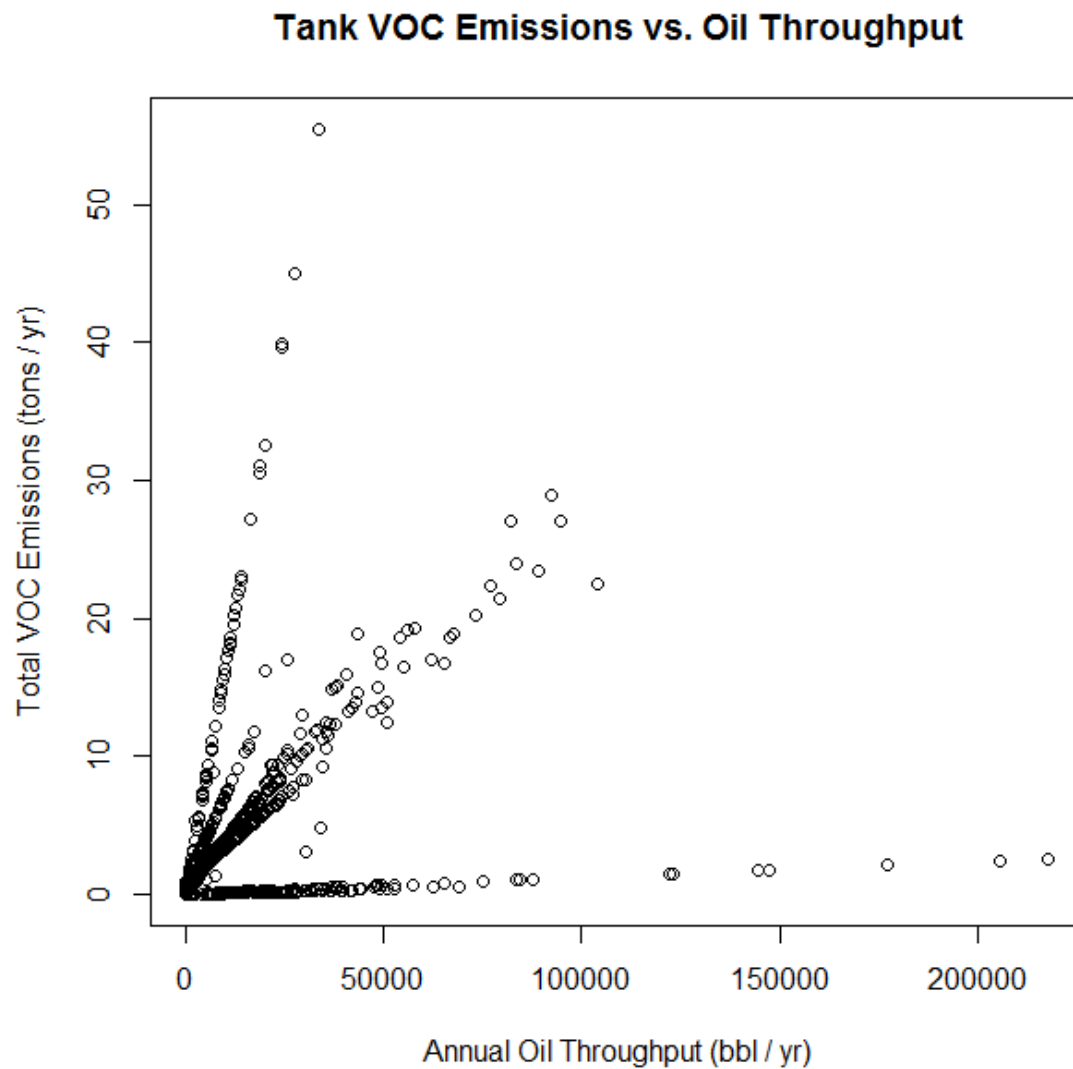


Figure 2: Tank VOC emissions vs. oil throughput data from OGEI database (March 30, 2016 data).

To: Patrick Barickman, Whitney Oswald, Brock LeBaron
From: Jon Wilkey
Date: June 30, 2016
RE: June 2016 monthly report on integrating equipment-based emission factors into the R model

During the month of June, work on the R model has focused finishing the model features discussed during the June 6th project meeting. Detailed change notes are given below. At this point, I believe that all coding work on the model is finished. The User Manual will be updated to reflect the final revisions and will be delivered by July 15, 2016, along with ownership of the R model's version control repository.

Change Log

1. User input options:

- a. Revised save/load options. If specified in the options file, the entire R workspace (comprised of results from the data analysis, Monte-Carlo simulation, and post-processing) is saved at the conclusion of the model run, and can be reloaded at a later time.
- b. Implementation dates of emissions reductions can now be specified using calendar dates (in "YYYY-MM-DD" format) instead of using model time steps.
- c. Emission reductions
 - i. Added the following selection criteria as built-in options (for each piece of equipment):
 1. Date that a well is drilled
 2. Implementation date of the emissions reduction
 3. Well type – oil, gas, or both
 4. Jurisdiction (i.e. surface lease owner) – can be Federal, State, Fee, Indian, or all of the above
 5. County – Uintah, Duchesne, or both
 6. Annual emissions thresholds (for each type of equipment and pollutant)
 7. Maximum monthly production rate (for both oil and gas). Goal here is that you can select only those wells capable of producing $\geq n$ barrels of oil or MCF of gas.
 - ii. Finally, reductions now only apply to wells with uncontrolled sources (accidental double-reductions are no longer possible)
 - iii. Only the records that meet all specified selection criteria have reductions applied to them.

2. Post-processing

- a. Excel export includes worksheets showing total emissions by species.

Several other features were implemented with the following changes:

3. Calculate tanks emissions by operator.

- a. Issue: I've found that picking the operator is redundant. Ultimately, the model calculates the ratio $r = (\text{mass of VOCs emitted}) / (\text{volume of fluid produced})$. The fluid produced is oil or condensate (there was no relationship between water throughput and tank VOC emissions). In the data analysis performed last month there is clear evidence that the value of r varies by operator. However if we randomly pick the operator for new wells, and then use the r -value fitted to that operator (which so far has always been a straight line with an intercept of zero and a slope of r), then really we're still randomly picking r . Since the frequency with which a given operator/ r -value occurs is identical, we're just adding an additional data analysis and simulation step.

Another potential reason to go with individual operators would be to differentiate between r -values for oil and condensate throughput (as most entries in the OGEI database are either all oil or all condensate throughput). However, DOGM doesn't track this difference, and simply reports oil and condensate production together as "oil" production. Consequently, the rest of the R model uses the same approach, and doesn't differentiate between the two fluid types either.

- b. Revision: The Monte-Carlo simulation picks the r -value directly. I sum together oil and condensate throughput volumes in the OGEI database when determining r -values. Any results that predict emission rates higher than the maximum emissions reported in the OGEI tanks database are capped at that maximum value.

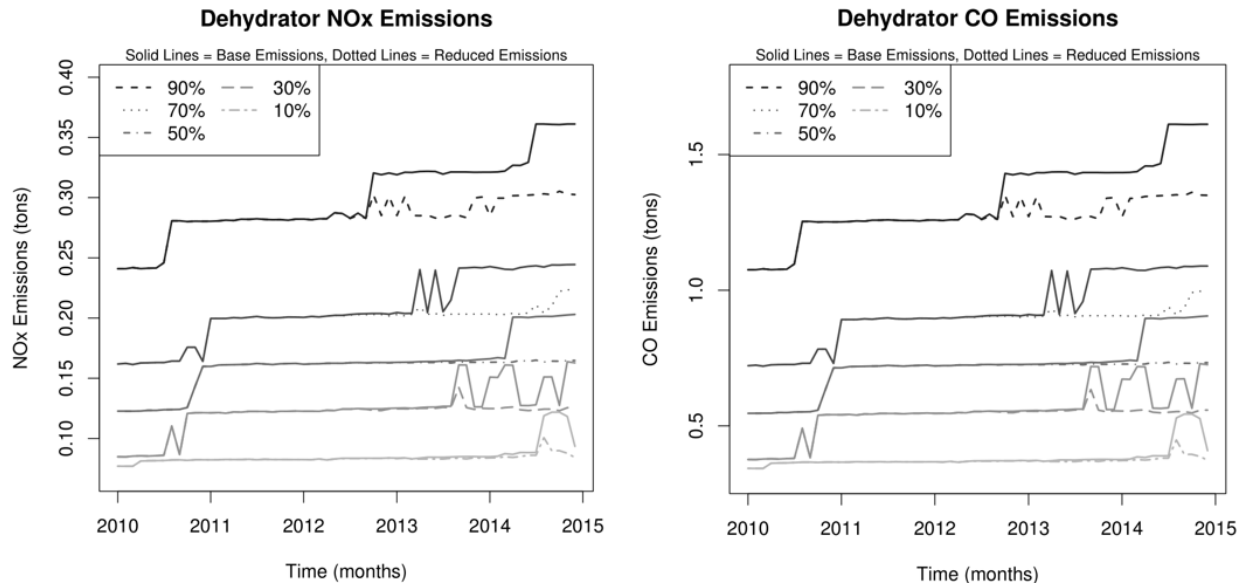
4. Use actual entries in OGEI database for prior wells instead of randomly drawing from cumulative probability tables (CPT).

- a. Issue: There are many complications to matching prior wells with rows in the OGEI database. Many wells do not have matching rows in the OGEI database for each type of equipment. Additionally, some wells are associated with multiple facilities in the OGEI database (and therefore have multiple matching rows in the CPTs). For example, the well API # 4301330403 (operator ID # 22, facility ID # 5277) is associated with two different entries in the RICE & turbines table (RICE ID # 960 and 961), which have substantially different emissions. In order to implement this feature, we'd need to:
 - i. Match wells that have a single matching entry in the OGEI database (before starting the Monte-Carlo simulation)
 - ii. Randomly draw from the CPTs for prior wells without matches in the database
 - iii. Randomly select between entries in the database for prior wells with multiple matches

- b. Revision: There isn't enough time left to implement this feature. Also, given how much random selection already occurs for prior wells with no/multiple matches, I think that there's limited value in finishing this feature. I've commented out the code used to find matching rows in the function that performs the data analysis of the OGEI database (but it's still there in case anyone wants to try tackling this issue in the future).

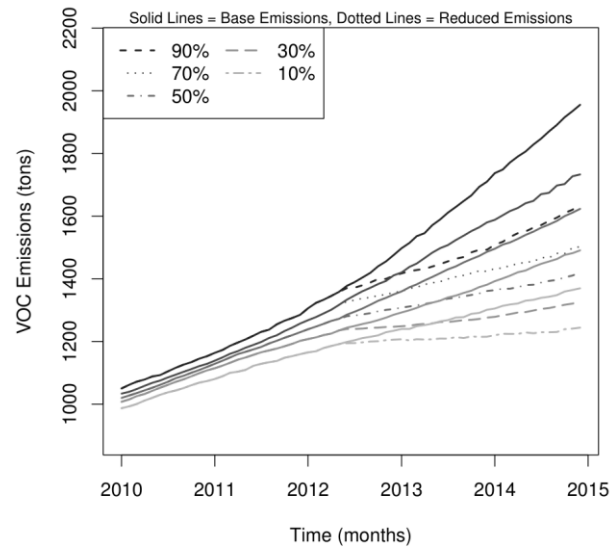
Finally, the following bugs were fixed/investigated:

5. Confirmed that 6 ton/yr VOC emissions threshold rate was working as intended for tanks. Only 5.92% of tanks in the OGEI database have VOC emissions ≥ 6 ton/yr, which is why there appears to be almost no reduction in VOC emissions from tanks when that emissions threshold is applied to the model's emissions reductions criteria.
6. Why are the plots of for emissions from dehydrators so odd?
 - a. Plots of CO and NOx emissions from dehydrators look very strange:



However after digging through the code I found that 99.89% of results for dehydrator NOx and CO emissions are == 0 (dehydrator NOx and VOC emissions are solely from combustor and combustor pilot light activity). So what we're really seeing is a bunch of noise around the very few results that are non-zero. By comparison, the VOC emissions from dehydrators are normal:

Dehydrator VOC Emissions



To: Patrick Barickman
From: Jon Wilkey
Date: September 30, 2016
RE: September 2016 monthly report on price forecasting for the R model

During the month of September work on this project focused on the following tasks:

- Collect long-term oil and natural gas price forecasts from alternative sources
- Review literature for energy price forecasting models

A summary of the work completed on each task is presented below. Over the coming month the information gathered will be compared against historical prices to determine the accuracy and performance of various forecasting sources and methods.

Alternative Sources of Energy Price Forecasts

The first task of this project was to search for and collect price forecasts from non- U.S. Energy Information Administration (EIA) sources. At present, the R model uses EIA's Annual Energy Outlook (AEO) energy price forecasts to estimate (a) the number of new wells drilled in the Uinta Basin and (b) the economic value of each existing well (for the purpose of determining when a well is abandoned). While EIA's AEO is the standard reference for oil and natural gas price forecasts, a number of other governments, agencies and businesses publish their own forecasts. These forecasts could be used as an alternative to the AEO forecast and/or combined to create a meta-forecast.

A summary of the price forecasts found so far is shown in Figures (1)-(3) (attached at the end of this report). Of the non-EIA sources, the most potentially useful long-term forecast is from Canada's National Energy Board (CNEB)¹. Like EIA's forecast, the CNEB forecast covers both oil and natural gas prices and is publicly accessible. Also noteworthy is the U.K.'s Dept. of Energy & Climate Change (DECC) forecast², although only their oil forecast is directly comparable (their natural gas price forecasts are for the European market). Unlike EIA, the CNEB and DECC forecasts have limited historical records. CNEB published forecasts in 1994, 1999, 2003 (all inaccessible online), 2007, 2009, 2011, 2013, and 2016 (all available online). The DECC was created by the U.K. in 2008, and it appears that they began releasing annual forecasts in 2012. The last source of long-term publicly available price "forecasts" is the Organization of Petroleum Exporting Countries (OPEC) World Oil Outlook³. Forecasts is used in quotes here however because OPEC defines the prices they report as their price assumption, and it appears that they use this assumption as an input in their work rather than produce it as a result (the International Energy Agency, IEA, does this as well). OPEC's forecasts are available (for oil only) on an annual basis starting in 2007.

There are many sources of short- to medium-term forecasts that could be useful for making (or informing) short-term projections in the R model. Notable agencies that produce short- and medium-term forecasts include the International Monetary Fund (IMF, 1-year, with uncertainty estimates)⁴, World Bank (10 year)⁵, IEA (publicly available 5-year price assumptions, long-term forecast must be purchased)⁶, and the Organization for Economic Co-operation and Development (OECD, 5-year forecast). With the exception of the World Bank, these agencies only forecast oil prices. I was able to find several

forecasts from the private sector (Deloitte⁷, Moody's⁸, and the Economist's Intelligence Unit⁹) that were publicly accessible, but historical data from these sources (if it exists) is behind a pay wall.

Energy Price Forecasting Models

The second task of this project was to review the literature on the subject energy price forecasting techniques. As an alternative to (or in combination with) using a price forecast from an outside source, we could use any of a number of forecasting models described in peer-reviewed literature to project oil and gas prices in the Uinta Basin. Of the set of articles I've found on the subject, Behmiri and Manso (2013)¹⁰ has the best review of price forecasting techniques, which I've summarized and outlined below.

1. Econometric Models

a. Time Series Models

- i. Predict future energy price based on historical energy price data
- ii. Commonly used models:
 1. Auto-Regressive Integrating Moving Average (ARIMA)
 - a. Models that include/exclude any combination of terms for:
 - i. Random fluctuation of prices around a mean
 - ii. The mean changes as a function of time
 - b. Example: Geometric Brownian Motion (the only econometric method currently in the R model) is an ARIMA type model with one random fluctuation term (of constant variance) and one term that acts as the mean and changes linearly over time
 2. (Generalized) Auto-Regressive Conditional Heteroscedastic (ARCH/GARCH)
 - a. Extension of ARIMA models which allows the variance (i.e. volatility) in the random price steps to change over time
 - b. Example: variance in next time step = variance in current time step + randomly selected value
 3. Markov Switching Model of Condition Mean
 - a. Describes the price trend with several different equations (e.g. bear market, bull market, etc.)
 - b. Randomly select which equation to use to determine prices in each time step of the forecast

b. Financial Models

- i. Predict future energy price based on futures trading in the commodity market

c. Structural Models

- i. Predict future energy price based on economic fundamentals
 1. For oil, commonly used fundamentals are oil inventory levels, oil consumption and production, GDP, interest rates, and OPEC behavior

2. Machine Learning Methods

- a. Give a defined algorithm a set of inputs and outputs and use the algorithm to create a relationship between them. Types of machine learning algorithms that have been applied in the literature (as of Behmiri and Manso's review paper):
 - i. Artificial neural networks

- ii. Support vector machines
- iii. Bayesian networks
- iv. Genetic algorithms

I'm currently in the process of testing the feasibility of implementing each of the forecasting methods described by Behmiri and Manso in the current R model. There are a number of pre-built packages in R that directly support modeling and fitting time series, and the only input required is the historical price data, which we already have on hand. The financial and structural models would be more difficult to implement because they require additional input data that we haven't collected (futures prices, oil inventories, etc.) and I'd recommend against trying to implement either model type (given the time frame of this project). Machine learning is a popular topic in data science and there are hundreds¹ of R packages that could be applied to this problem. The two concerns that I have about implementing a machine learning approach are (a) picking which algorithm(s) to try, and (b) the complexity of implementing or interpreting their results (and communicating how the algorithm works).

¹ For example, the "caret" package in R supports 230 different algorithm types
<http://topepo.github.io/caret/available-models.html>

Figures

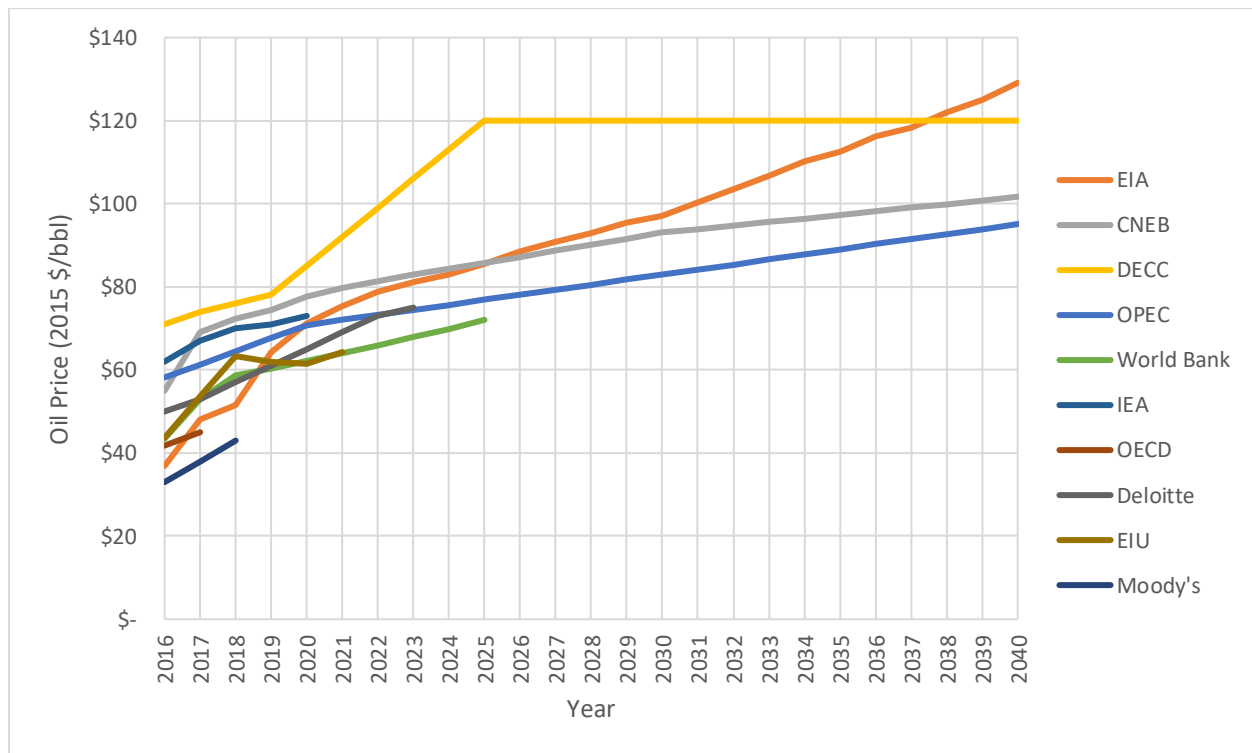


Figure 1: Oil price forecasts from various sources^{1,2,11,3,5,6,12,7,9,8}. Notes:

1. All prices are for West Texas Intermediate (WTI) oil in 2015 U.S. dollars (USD) per barrel (bbl), except:
 - a. OPEC prices are representative of the “OPEC Resource Basket” (ORB) in 2014 USD/bbl.
 - b. World Bank prices were originally in 2010 USD/bbl.
 - c. Both OPEC and World bank prices were adjusted to 2015 USD using the Consumer Price Index.
2. OPEC and IEA prices are assumptions, not forecasts.

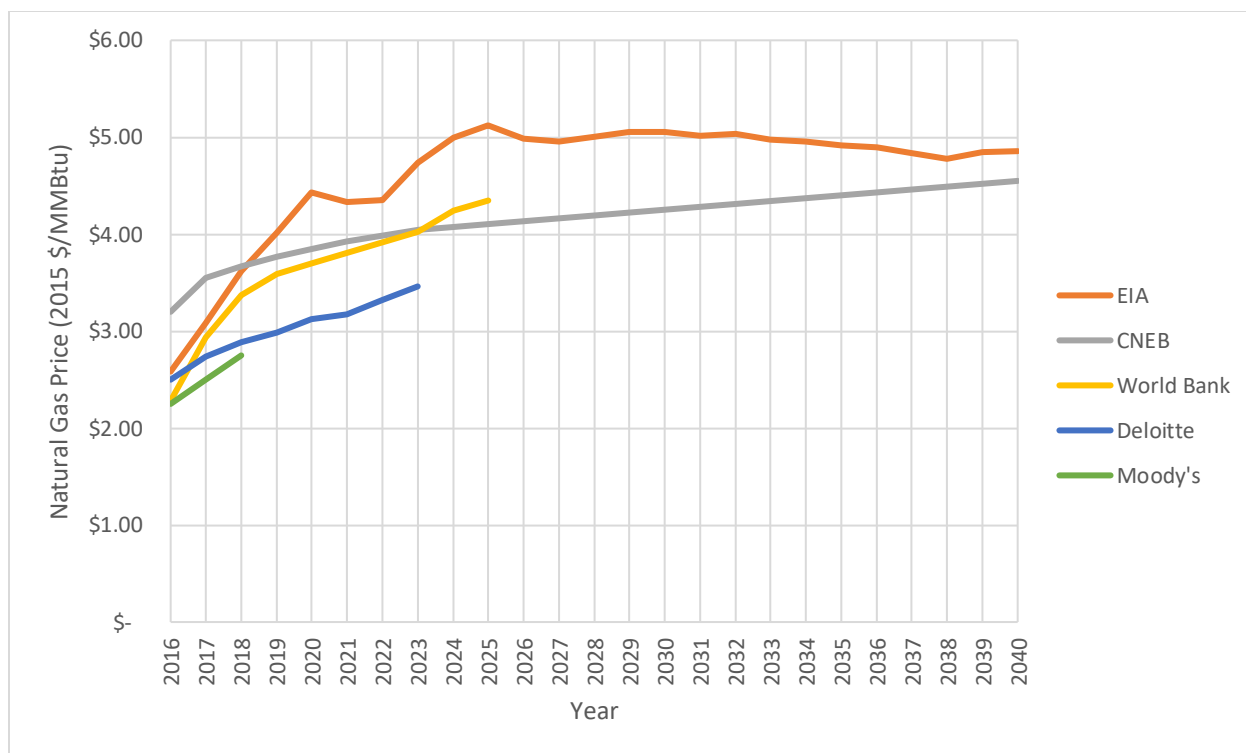


Figure 2: Natural gas price forecasts from various sources^{1,11,5,7,8}. Notes:

1. All prices are for Henry Hub natural gas in 2015 USD per million Btu (MMBtu). Conversions:
 - a. World Bank prices originally reported in 2010 USD/MMBtu, adjusted using CPI.
 - b. Deloitte forecast was in 2015 USD per MCF, adjusted assuming heating value of 1040 Btu per cubic foot.

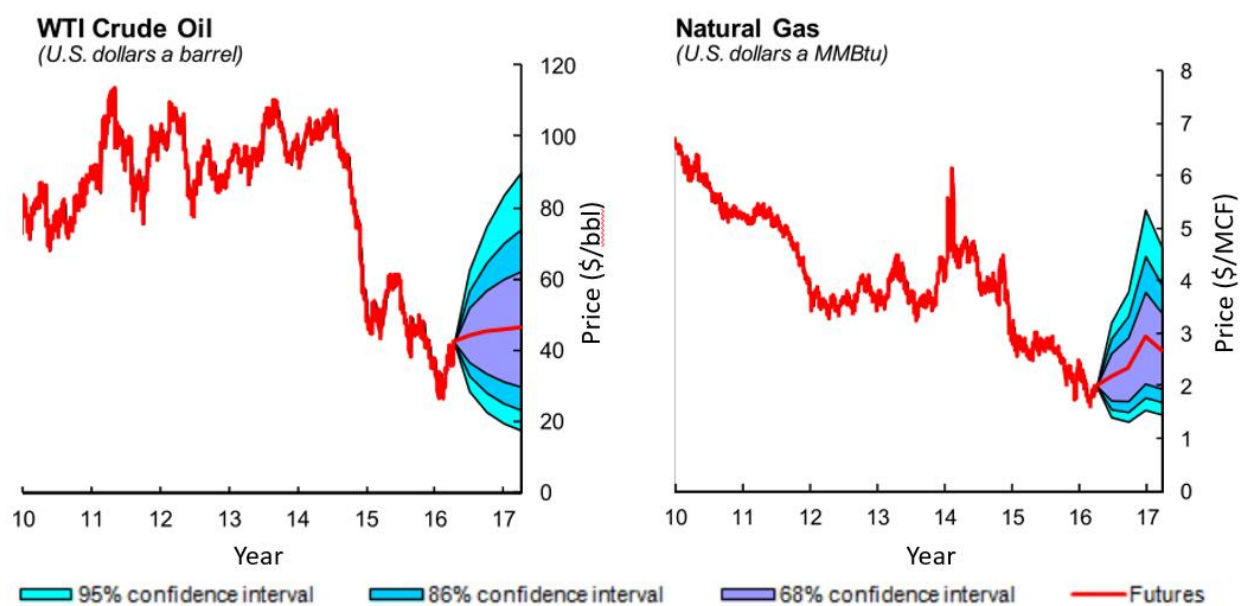


Figure 3: IMF price forecasts for oil and natural gas (2015 dollars)⁴.

References

1. Canada National Energy Board. Benchmark Prices. Canada's Energy Future 2016: Energy Supply and Demand Projections to 2040. <https://apps.neb-one.gc.ca/fttrppndc/dflt.aspx>
2. Department of Energy & Climate Change. *DECC 2015 Fossil Fuel Price Projections*. London, United Kingdom; 2015. <https://www.gov.uk/government/publications/fossil-fuel-price-projections-2015>
3. OPEC. *World Oil Outlook*. Vienna, Austria: Organization of the Petroleum Exporting Countries; 2015. http://www.opec.org/opec_web/en/publications/340.htm
4. International Monetary Fund. *Commodity Price Outlook and Risks - April 2016*. Washington, DC; 2016. <http://www.imf.org/external/ns/cs.aspx?id=329>
5. World Bank. *Commodity Price Forecasts - July 2016*. Washington, DC; 2016. <http://pubdocs.worldbank.org/en/764161469470731154/CMO-2016-July-forecasts.pdf>
6. International Energy Agency. *Medium-Term Market Report 2015*. Paris, France; 2015. <http://www.iea.org/publications/freepublications/publication/medium-term-oil-market-report-2015.html>
7. Deloitte. *Price Forecast.*; 2016. <http://www2.deloitte.com/ca/en/pages/resource-evaluation-and-advisory/articles/deloitte-canadian-price-forecast.html>
8. Marshall T. Moody's cuts oil price estimates as supply glut continues. *Moody's Investors Service*. https://www.moody's.com/research/Moodys-cuts-oil-price-estimates-as-supply-glut-continues--PR_342747. Published January 21, 2016
9. The Economist Intelligence Unit. Global Forecasting Service - Commodity Price Forecasts. The Economist. <http://gfs.eiu.com/Article.aspx?articleType=cf&articleId=1174608101&secId=0>. Published 2016
10. Bashiri Behmiri N, Pires Manso JR. Crude Oil Price Forecasting Techniques: A Comprehensive Review of Literature. *SSRN Electron J*. 2013;30-48. doi:10.2139/ssrn.2275428.
11. U.S. Energy Information Administration. *Annual Energy Outlook - Table A1*. Washington, DC; 2016. <http://www.eia.gov/forecasts/aeo/>
12. OECD. *Economic Outlook Annex Table 16: Oil and Other Primary Commodity Markets*. Paris, France; 2016. <http://www.oecd.org/eco/outlook/wages-costs-unemployment-and-inflation.xls>

To: Patrick Barickman
From: Jon Wilkey
Date: November 2, 2016
RE: October 2016 monthly report on price forecasting for the R model

During the month of October work on this project focused on analyzing the various methods of creating energy price forecasts (EPFs) identified in literature review conducted during the month of September. A summary of this analysis is presented below, categorized by EPF type (structural, time-series, and user-input) and concluding with a comparison of the relative merits of each approach. Based on these results, the recommended methods for generating EPFs are given in the table below:

EPF Method	Number of Years into the Future that EPF is Being Made			
	0 – 1	1 – 5	5 – 10	10 – 20+
Structural		X	X	
Time-Series			X	X
User-Input	X			

Work for the month of November is expected to focus on adding options for each type of EPF method to the R model, documenting model changes in the User Manual, and final delivery of the R model repository.

Structural EPFs

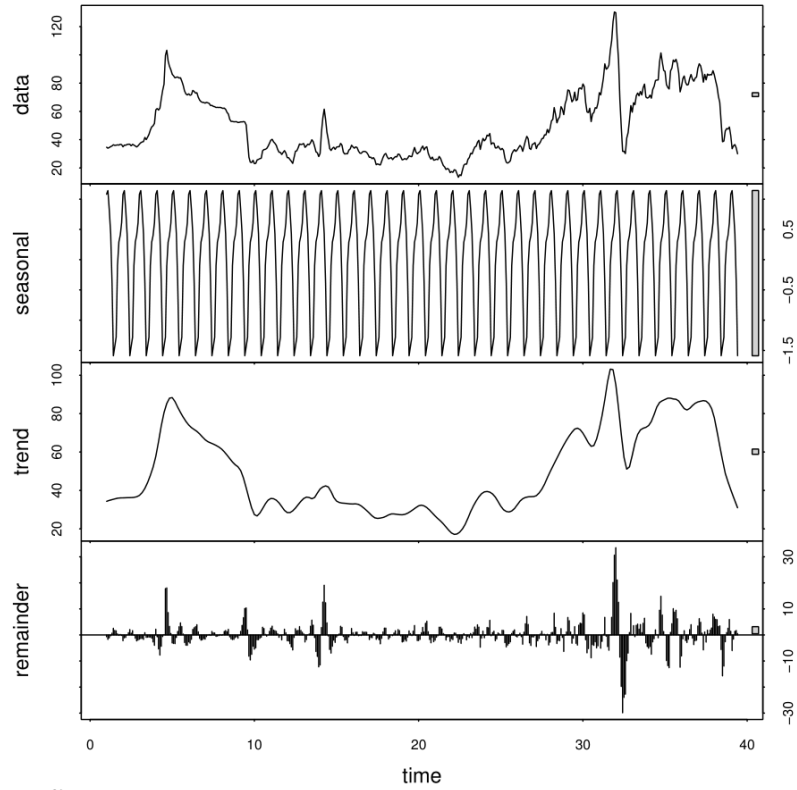
EPFs in this category attempt to model the dynamics of supply and demand that ultimately lead to observed energy prices. Of the long-term, publicly accessible EPFs reviewed in last month's report, only U.S. EIA and U.K. Dept. of Energy and Climate Change (DECC) attempt to apply independent structural EPFs that could be used in the R model. However, given that the DECC has only published four EPFs (2012 – 2015), there is little basis for assessing the accuracy of the DECC's EPFs and the R model's current EIA-based methodology remains the best EPF in this category.

Time-Series EPFs

In a time-series forecast, the future values of some quantity is predicted solely based on the past values of that same quantity. In general, the quantity to be forecasted is assumed to be comprised of (a) seasonal variation, (b) an underlying trend, and (c) noise. There are a number of methods for fitting the underlying trend, but in general the most widely used is the ARIMA method (autoregressive integrated moving average), which utilizes the past values (p), differences in past values (d), and the moving average of the fitting error (q) as predictor variables. ARIMA models are classified based on how many of each type of these predictor variables (p, d, q) they use, and certain combinations are known to have characteristic behaviors (e.g. they decay to zero, result in a random walk around a constant mean, etc.).

To applying the time-series approach to our specific oil and gas price history, we begin by taking the decomposition of past oil and gas prices into the seasonal/trend/noise components as shown in Figure 1 on the following page.

(a)



(b)

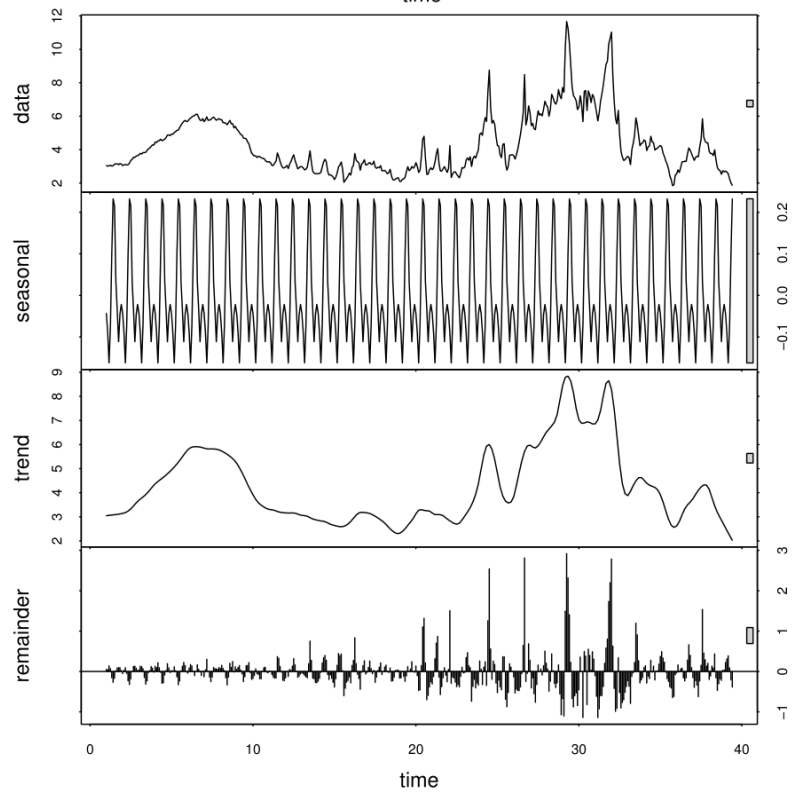


Figure 1: Time-series decomposition of (a) oil (\$ / bbl) and (b) natural gas (\$ / MCF) prices in the Uinta Basin. Note that the “time” values on the x-axis are years since 1977.

From these decompositions, we can draw the following conclusions:

1. Seasonal variation in oil prices ($\pm \$1.5/\text{bbl}$) and natural gas prices ($\pm \$0.2/\text{MCF}$) is negligible.
2. Over their entire price history, the trend for both oil and gas is flat. There are clearly periods where prices went up during the early 1980s and mid 2000s, but long-term there is no clear upward/downward trajectory.
3. There is substantial noise in the remainder term ($\pm \$30/\text{bbl}$ for oil and $+\$3 -\$1 / \text{MCF}$ for gas).

Consequently, we can expect that an ARIMA time-series fit to the price data will project a constant future price (as the median forecast) with substantial variation around that constant value.

The best method I've found for fitting ARIMA(p,d,q) models in R is using the forecast¹ package, which allows the user to either fit a specified ARIMA model or to automatically select the best ARIMA model based on AIC (Akaike information criteria) values (a measure of the goodness of fit for a statistical model). An example twenty year forecast for both oil and gas prices is shown in Figure 2, which was generated by testing various ARIMA models on the oil and gas price histories from July 1977 (the start of the oil and gas price history data set) to December 1995.

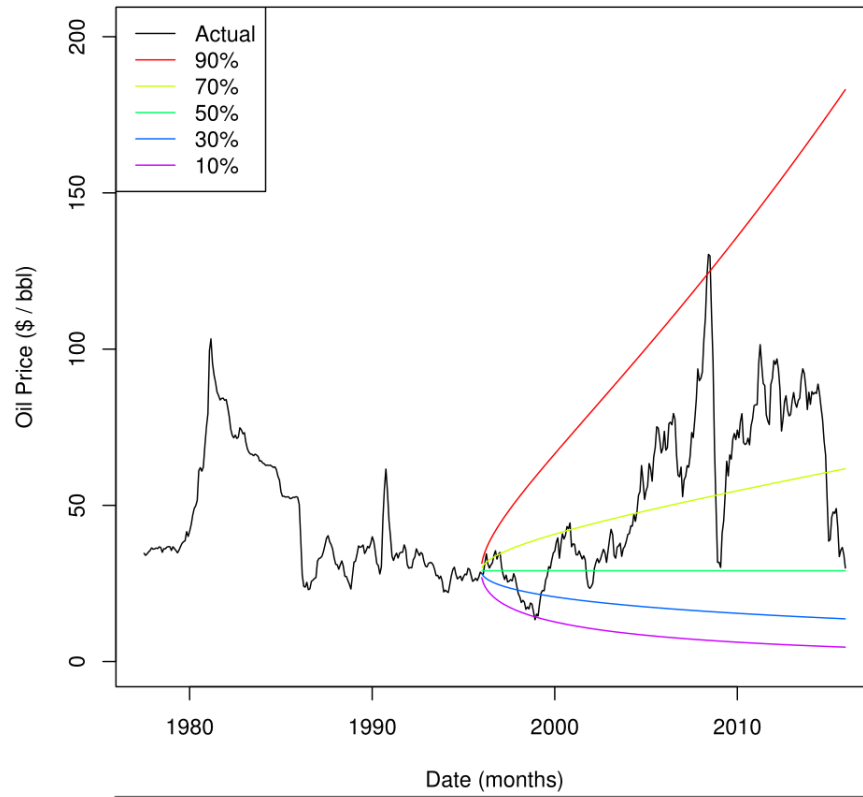
It should be noted that the auto-ARIMA fit function can produce substantially different fits based on the selection of training data, and that the best fit for a given set of training data (as measured by AIC values) may not actually be the best match to the test data used for cross-validation. For example, the auto-ARIMA function found that the best fit for oil prices given the 1977-1995 data was a (0,1,1) model, which produces the results shown in Figure 2(a) that fully encompass the observed test prices within the 10th-90th percentiles. For natural gas, the best auto-ARIMA fit for over that same period is a (1,1,4) model (see Figure 3), which produces a narrower price forecast than most of the observed natural gas prices during the test period. For that reason, the use of a (0,1,1) model as in Figure 2(b) is preferable. When the auto-ARIMA is fit to the entire price history data set (1977-2015), we again get a different set of best-fit models for oil (1,1,0) and gas (0,1,2) prices. I'll be performing additional work over the coming month to clarify whether a specific ARIMA fit should be used or if the model should be allowed to select the ARIMA model automatically.

User-Input EPFs

The final option for handling EPFs is to allow the user to specify the price directly. The R model can already support this approach as-of the version deliver to UDAQ in July, however it could be further expanded to support additional options such as directly specifying the number of wells drilled each month.

¹ Hyndman RJ and Khandakar Y (2008). "Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software*, **26** (3), pp. 1-22. <http://www.jstatsoft.org/article/view/v027i03>

(a)



(b)

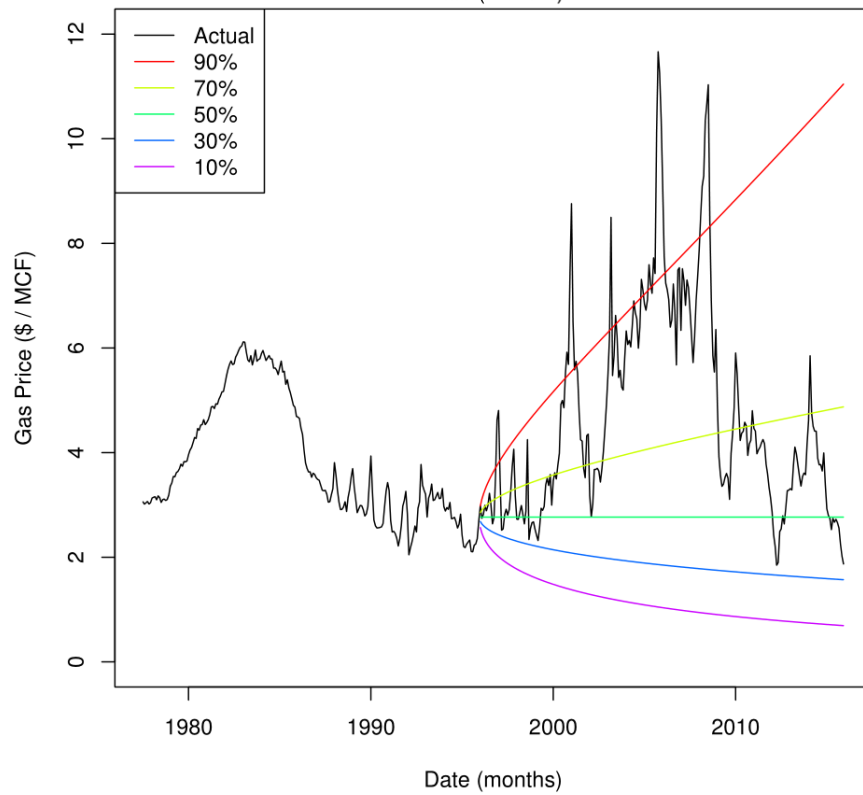


Figure 2: Twenty year ARIMA(0,1,1) forecasts of (a) oil and (b) natural gas prices.

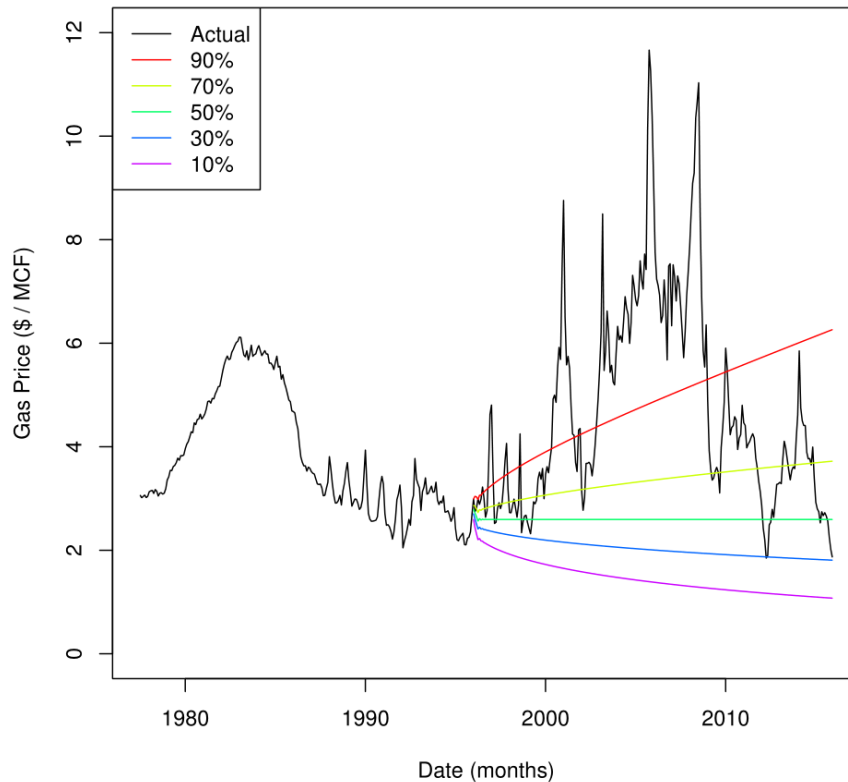


Figure 3: ARIMA(1,1,4) forecast for natural gas prices using the same training set as Figure 2(b). Note that the result produces a narrower forecast than actually observed during the majority of the 20-year test period. The (1,1,4) model is the best fit to the training data, but performs worse than the (0,1,1) model in Figure 2(b) during the test period.

Comparison of EPF Methods

The best choice of EPF model depends on the time-frame over which model predictions are being made:

1. Annual forecast (≤ 1 year)
 - a. User-input is highly preferable. Within a 1 year period it's reasonable to make any of the following simplifying assumptions to eliminate the uncertainty and complexity of EPFs from the R model:
 - i. Hold current oil and gas prices constant
 - ii. Assume prices are equal to EIA's forecast (on average EIA's relative error 1 year out is only 15%)
 - iii. Directly specify the number of new wells you expect to see in the Uinta Basin over the coming year.
2. Short-term forecast (1 – 5 years)
 - a. Up to five years, EIA's structural forecast is preferred.
 - i. There's substantial uncertainty in EPFs over a 5-year period, so user-input is likely to be wrong unless the user has access to more detailed and direct knowledge about future development than is publicly available.

- ii. There are enough observations about EIA's forecasting error to reasonably incorporate that error into their structural price forecast.
 - iii. EIA's structural forecast can predict trends up/down which the time-series approach tends to ignore (given the past price history).
- 3. Medium-term forecast (5 – 10 years)
 - a. Time-series is slightly preferred.
 - i. There are enough observations about EIA's error rate to use the structural forecast, but the error rates are skewed because the 1999-2005 forecasts from EIA are all essentially the same forecast.
 - ii. However, EIA has made some forecasts (such as their 2008 forecast) that have performed reasonably well over the medium-term.
 - iii. There are no downsides to the time-series approach other than (again) the tendency of the ARIMA fits to predict a flat constant price.
- 4. Long-term forecast (10 – 20 years)
 - a. Time-series approach is highly preferable.
 - i. At twenty years, all models will be wrong, but as the cross-validation tests with the ARIMA(0,1,1) fit show in Figure 2, at least the time-series approach can fully encompass the range of observed prices.
 - ii. There are too few observations to use the current EIA-based EPF method.
 - 1. Beyond 16 years it's mathematically impossible to use (at least two observations of the observed error rate are required in order to estimate a distribution)
 - 2. Even where mathematically possible, the error rates are not representative beyond 10 years (at which point every forecast from EIA is the identical 1999-2005 forecast)
 - 3. At some point in the (distant) future there will be enough information to use this approach.

To: Patrick Barickman
From: Jon Wilkey
Date: December 2, 2016
RE: November 2016 monthly report on price forecasting for the R model

During the month of November work on this project focused on implementing price forecasting options for short, medium, and long term time horizons, as well as updating the user manual to reflect the changes. A summary of the change-log is given below. At this point, the model is ready for final delivery to UDAQ pending approval and payment of the contract with the Univ. of Utah.

Change-Log

1. Major changes
 - a. Added ARIMA forecasting for medium/long term time horizon
 - i. Fits two ARIMA models:
 1. A user-specified ARIMA model
 2. A automatically selected ARIMA model
 3. Plots long-term forecasts of both ARIMA fits so that user can judge which type (user-specified or automatically selected) of model to use for simulation.
 - b. Overhauled energy price forecasting options
 - i. New possible forecasting models:
 1. User-specified ARIMA model
 2. Automatically selected ARIMA model
 3. EIA forecast with error propagation
 4. Actual energy prices (for cross-validation use only)
 5. Constant prices equal to average of last N months prior to start of simulation period
 6. User-specified price paths
 - ii. Oil and gas price forecasts can now be specified independent of each other
 - c. Drilling forecast changes
 - i. Added optimization step that searches for the best time delay between energy prices and well completions using the function:

$$W_t = a \cdot OP_{t-N} + b \cdot GP_{t-N} + c$$

where W is the number of wells completed during month t , OP is oil price, GP is gas price, (a,b,c) are coefficients fitted by linear regression, and N is a vector of 1 to an upper limit specified by the user. The model now fits the above equation (and its simpler variants which include only oil or gas prices separately) for each possible value of N and automatically selects the value that returns the lowest residual sum of the squares. The selected value of N is then automatically applied during the Monte-Carlo simulation.

- ii. Added option to allow the user to directly specify the drilling schedule (in the same way that energy prices can be specified).
 - d. Removed and Geometric Brownian Motion (GMB) price forecasting model (which is redundant now that ARIMA forecasting is present - a (0,1,0) ARIMA model is identical to GBM model)
 - e. Revised user-manual to cover the modeling changes.
2. Minor changes
- a. Updated all model input files and IO_options.R script settings
 - b. Revised Excel export feature
 - i. Previous version of this feature required the installation of Java
 - ii. New version uses a different package that instead requires and uses Rtools
 - 1. Rtools is a standalone program that is part of the R-project and is used for building R and R packages.
 - 2. Made the switch to using Rtools because:
 - a. It poses a lower security risk than a Java installation
 - b. It does not require updating by the user
 - c. Suppressed spurious warning messages in function that fits log-normal distributions to oil decline curve coefficients
 - d. Added option to export plots from postProcess script as either:
 - i. A single PDF file with many single page plots
 - ii. Individual PDF files for each plot
 - e. Automated some labor-intensive input calculations, specifically:
 - i. First purchase price specifications (user had to look at price data and enter the values for the time step just prior to the start of the simulation period)
 - ii. The number of wells drilled just prior to the start of the simulation period.
 - f. Added automatic deletion of large intermediate *.csv files created as part of the dogmDataUpdate function.