**Uncertainty of oil field GHG emissions resulting from information gaps: a Monte Carlo approach**

**Supplemental Information**

Kourosh Vafi

Adam R. Brandt

Department of Energy Resources Engineering, Stanford University

Stanford, California, 94305, United States

## S1. Methodology

We selected 8 parameters based on a previous study which described key variables controlling GHG emissions estimates produced by OPGEE.[1] Data for these 8 parameters were taken from public datasets (see Table S1 below). We used *EasyFit* to find the best-fitting distribution to fit to each dataset.[2] *EasyFit* ranked the goodness of fit of all available distribution functions. We determined the best fitting distribution using the Kolmogorov-Smirnov and Chi Square metrics. In most cases we found the first-ranked distribution based on Kolmogorov–Smirnov test were better visual fits to the data.

Nevertheless, the distribution ranking provided by *Easyfit* does not guarantee that the best fitting distribution function is an acceptable fit (i.e., all fits could be poor, including the best fit). In some cases none of available distribution functions in the library were a fit which met a visual inspection test. In such cases, we split the domains of the data, creating more specific datasets with smaller bounds. If we could not find a satisfactory fit after dividing the data using logical divisors (e.g., heavy crude oil and light crude oil) then we used probability tables instead. Table S1 summarizes the chosen probability distributions and the source of each dataset. Figures S1 to S8 show the underlying data and resulting distribution functions.

After selecting a distribution, set of distributions, or probability table that best fits the underlying data, we used *EasyFit* then to generate 10000 random numbers for each dataset. The 10,000 entries for each of 8 input parameters were recorded in a database.

**Table S1. Sources of data and selected best-fitting probability distributions**

| Parameter | Min | Max | Mean | N data | Reference | Distribution |
|---|---|---|---|---|---|---|
| **API** | 6 | 64 | 27 | 1303 | 3 | Exponential Power [a] |
| **Depth** | 10 | 17610 | 4614 | 1381 | 3 | Johnson SB |
| **GOR ( scf/bbl)** | 0 | 21380 | 1288 | 179 | 4,5 | Multiple |
| $API < 20$ | | | | 45 | | Generalized Pareto |
| $20 \leq API \leq 30$ | | | | 69 | | Generalized Extreme Value |
| $API > 30$ ( $GOR \leq 3538$) | | | | 58 | | Generalized Extreme Value |
| $API > 30$ ($7765 \leq GOR \leq 9927$) | | | | 2 | | Non-continuous [b] |
| $API > 30$ ($12182 \leq GOR \leq 21380$) | | | | 5 | | Non-continuous [b] |
| **Oil production rate (bbl/d)** | 28 | 83365 | 3495 | 152 | 6 | Multiple |
| $28 \leq Prod.\ rate \leq 4996$ | | | | 137 | | Fatigue Life (3P) [c] |
| $7408 \leq Prod.rate \leq 18863$ | | | | 8 | | Non-continuous [b] |
| $30628 \leq Prod.rate \leq 83343$ | | | | 7 | | Non-continuous [b] |
| **WOR** | 0 | 146 | 15 | 152 | 6 | Weibull |
| **SOR** | 0 | 9 | 3 | 23 | 6 | Johnson SB |
| **RSPC** [d] | 0 | 1 | 0 | 23 | 6 | Multiple |
| Frac. cogen = 0 | | | | 9 | 6 | Probability table |
| $0.07 \leq Frac.Cogen \leq 1.0$ | | | | 14 | 6 | Uniform |
| **Method of recovery** | - | - | - | 147 | 6 | Non-continuous [e] |

a- *EasyFit* refers to this distribution also as Error distribution which should not be confused with Error Function distribution. We quoted the exact *EasyFit*'s definition of this probability distribution in section S3 to avoid any ambiguity.

b- First a probability value is randomly generated based on uniform distribution then the input parameter is assigned based on the range of the probability which is indicated in the table.

c- Chi-Square test rank 1 distribution used

d- Ratio of Steam Produced by Cogeneration

e- See Table S2

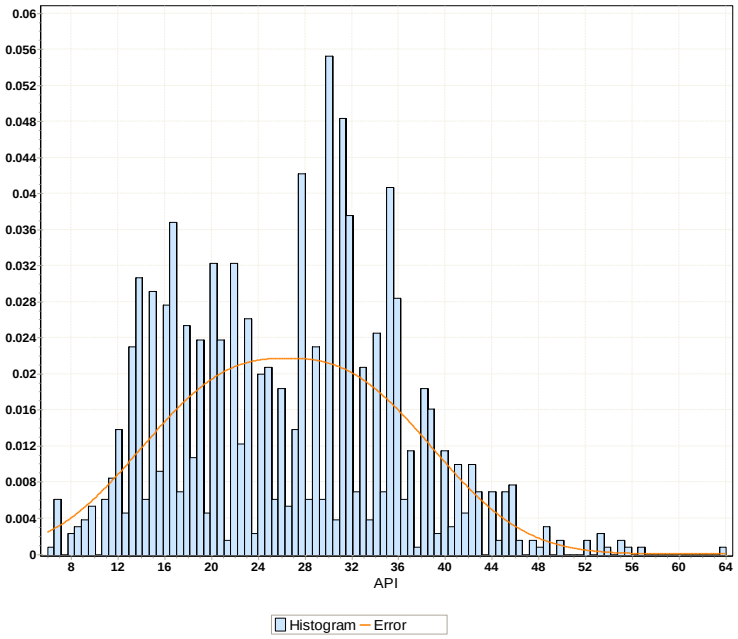


**Figure S1. API gravity distribution in California. Best-fitting: Error distribution or Exponential Power distribution (see section S3).**

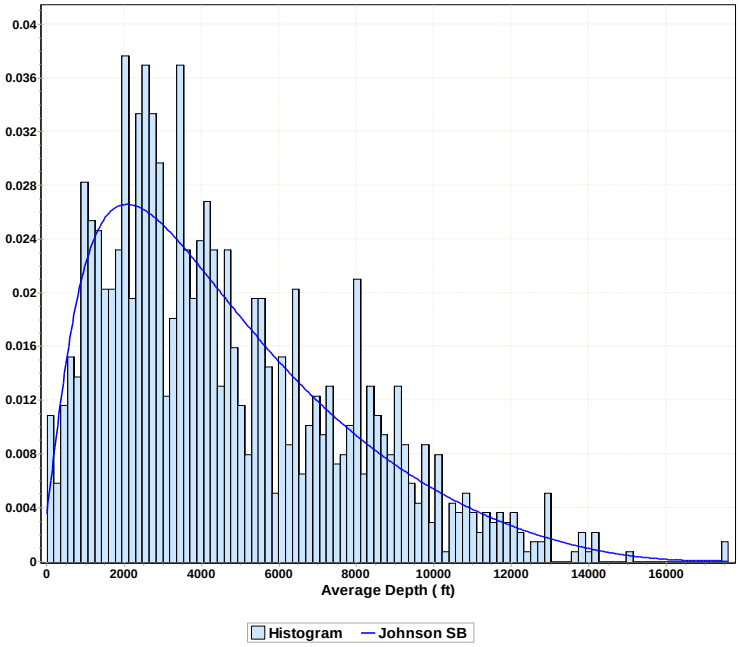


**Figure S2. Average reservoir depth distribution in California. Best-fitting: Johnson SB distribution.**

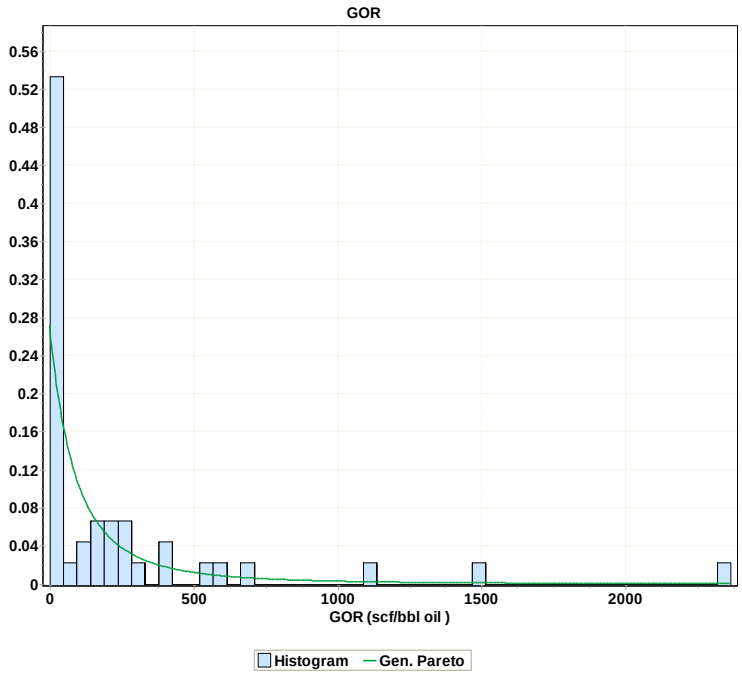


**Figure S3. GOR probability distribution in California for API< 20. Best-fitting: Generalized Pareto.**

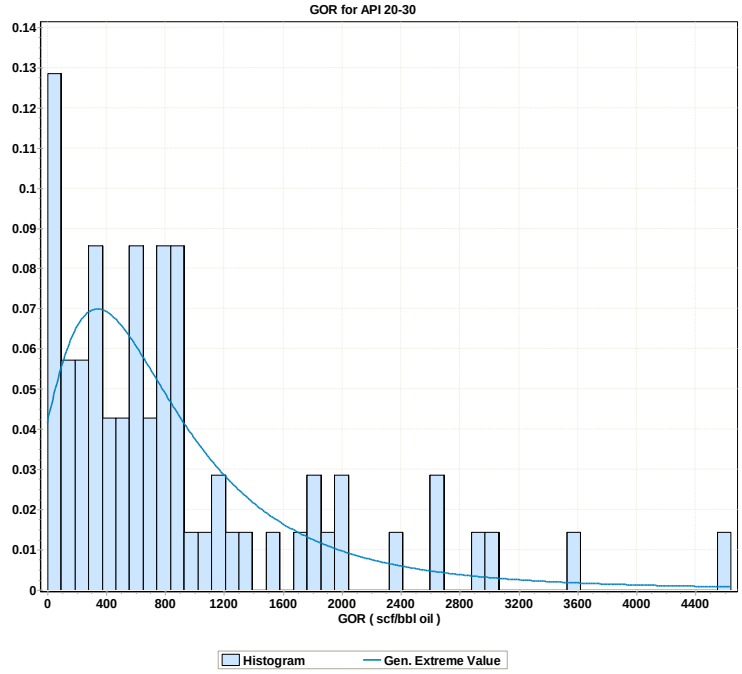


**Figure S4. GOR probability distribution in California for API gravity between 20 and 30 –Best-fitting: Generalized Extreme Value.**

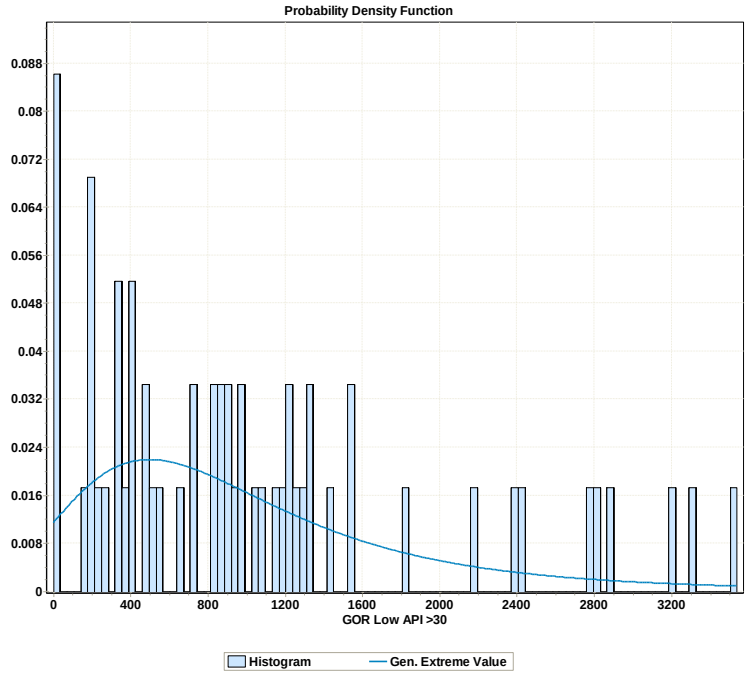


**Figure 5. GOR distribution in California for API gravity > 30 and GOR below 3540 scf/bbl. Best-fitting: Generalized Extreme Value.**

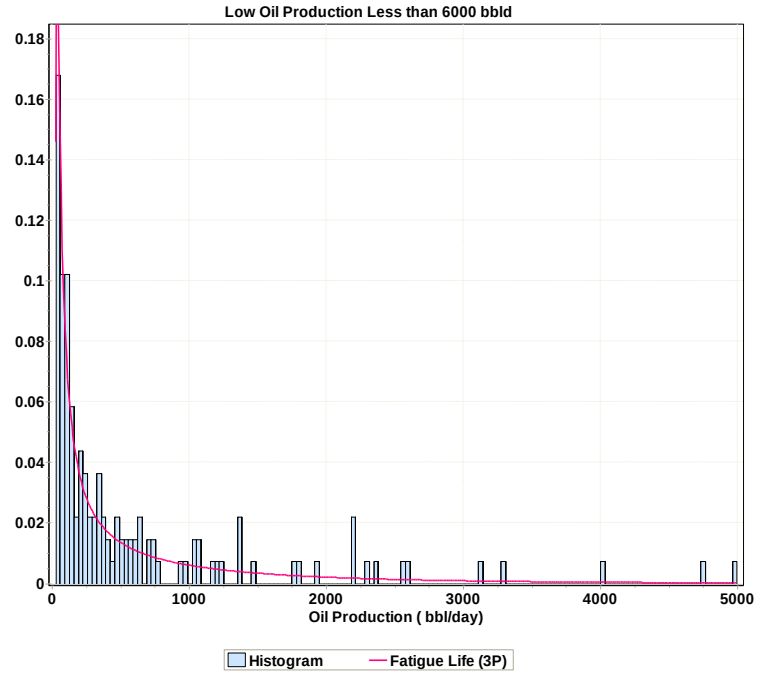


**Figure S6. Probability distribution of crude oil in California for production rate below 5000 (bbl/d). Best-fitting: Fatigue Life (3P).**

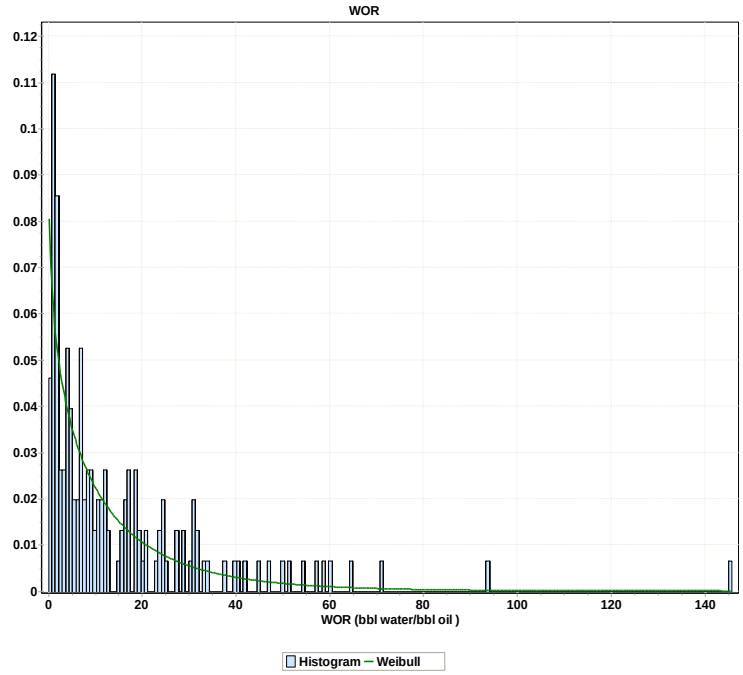


**Figure S7. WOR distribution in California. Best-fitting: Weibull distribution.**

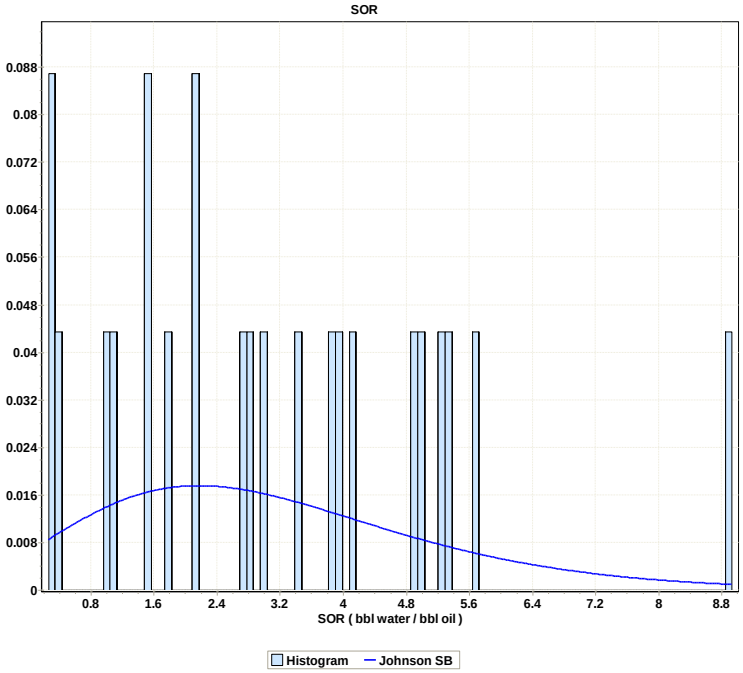


**Figure S8. SOR distribution in California. Best-fitting: Johnson SB distribution.**

The method of recovery in the Monte Carlo (MC) simulation is a non-continuous variable. It is either conventional or steam flooding. The probability of steam flooding was calculated based on the portion of crude oil produced within several API gravity ranges in California (See Table S2). The binary random numbers for selection of the method of recovery were generated anew for each simulation.

**Table S2. Probability table for use of steam flooding in California.**

| API gravity | Probability |
|---|---|
| $8 \leq API < 10$ | 0.423 |
| $10 \leq API < 15$ | 0.979 |
| $15 \leq API < 20$ | 0.657 |
| $20 \leq API < 28$ | 0.848 |
| $API \geq 28$ | 0.000 |

OPGEE version 1.1 Draft A[7] is used as the GHG calculating model. One minor change was made to OPGEE v1.1 Draft A: the standard temperature in the definition of standard volume (e.g., SCF) was changed from 32 °F to 60 °F.

The database of random numbers generated in *EasyFit* was incorporated into a new spreadsheet in OPGEE. OPGEE was modified with several Visual Basic for Application (VBA) scripts for the following:

- Calculation of the possible combinations of the input parameters.
- Removal of the inconsistent combinations from the table of the combinations.
- Populating OPGEE with random input parameters for 10000 runs for each acceptable combination.
- To remove odd results and combinations that result in errors (see below).

Some combinations of parameters resulted in "odd" results. For example, some atypical combinations can result in extremely high emissions rates, or even mass balance errors in the model. Based on available data for 152 oil fields,[6] OPGEE estimates of well-to-refinery (WTR) GHG emissions from production of California crude oils ranges between 1.7 to 31.4 $gCO_2$ eq./MJ

crude oil. Therefore results with WTR GHG emissions estimated as larger than 40 $gCO_2$ eq./MJ crude oil are discarded as outliers resulting from physically unlikely or impossible combinations of input parameters. These odd combinations of input variables should not be confused with the inconsistent combinations removed in the multi-factor sensitivity analysis (described in main paper). Additionally, some combinations of parameters caused model errors (for example due to inconsistency of the productivity index with the required production rate). The percentage of these purged cases out of 10,000 runs for each case study are given in Table S3.

**Table S3. Percentage of purged run-with-errors and results with WTR GHG emissions beyond 40 $gCO_2$eq./MJ from 10000 runs for each combination.**

| | Min (%) | Max (%) | Avg, (%) | SD (%) | Number of combinations |
|---|---|---|---|---|---|
| Wilmington | 0 | 4.6 | 2.2 | 1.2 | 38 |
| Midway Sunset | 0 | 3.6 | 1.6 | 1.0 | 86 |
| Beverly Hills | 0 | 4.2 | 1.6 | 1.1 | 38 |

The 8 simulation input variables are shown in Figure S9. They are grouped in three levels. There is a simple rule to distinguish the inconsistent combination of parameters. Once a parameter is selected, then all the dependent variables in higher levels should be selected as well. For example, we can select method of recovery only when we have already selected API gravity. This is done to prevent physically unrealistic combinations. For example: if we choose steam injection as the method of recovery and let API gravity change freely through its range, then the instances in which we combine a very light crude production and steam flooding as the recovery method are unrealistic.

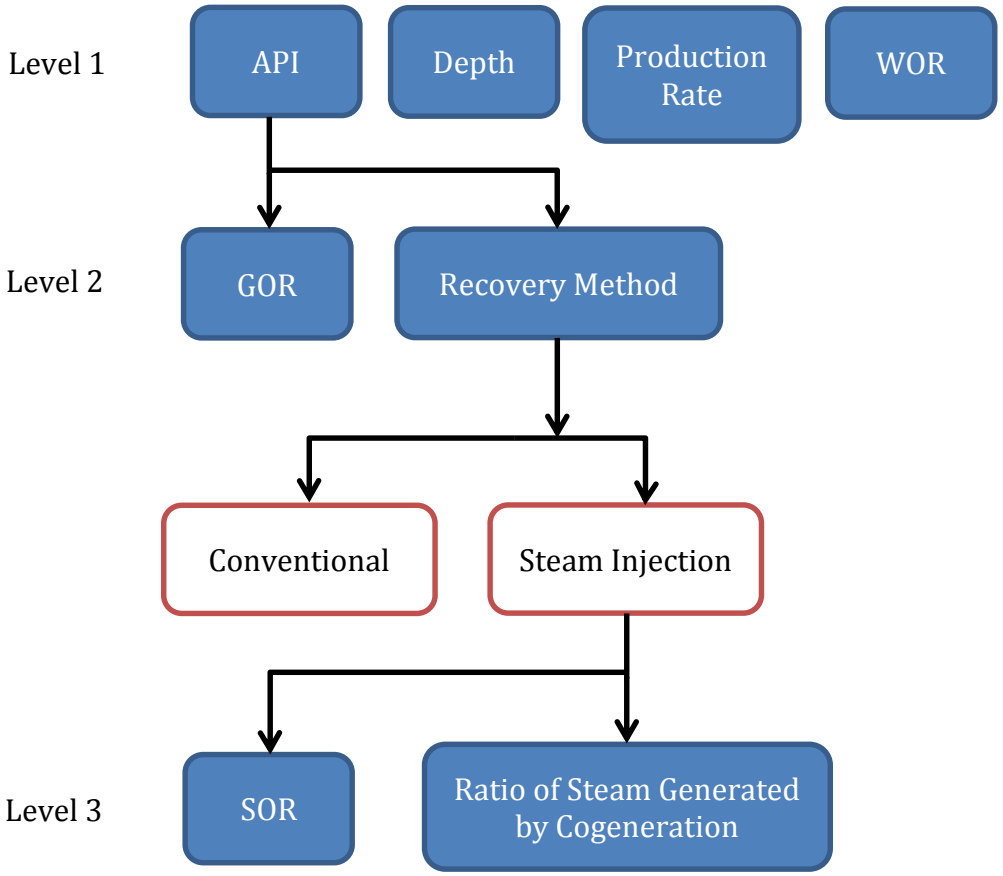


**Figure S9. Causal relationship between fixing of dependent simulation parameters.**

**S2: SD, CV, and Error diagrams**

Figure 3 shows the WTR GHG emissions from the three oilfields versus number of input variables which are fixed (fixed means known as a real number). Figures S10 to S18 show the corresponding diagrams for SD, CV, and error (bias) versus the number of fixed variables. Error, or MC mean bias, is defined by equation S1:

$$Error = \left(\frac{True\ WTR\ GHG - Dist.mean}{True\ WTR\ GHG}\right) \times 100 \qquad \text{(S1)}$$

We assume that the model calculates true WTR GHG value when all the input variables are known (fixed). When we know the value of none of the input variables, we use the regional probability distribution. We name what the model calculates with such probability distributions the base line. CV is calculated by dividing SD by the distribution mean.

From Figures S10 to S18 we can say before learning is complete, for all cases, there are combinations of pieces of information that increase SD, CV, and the absolute error above the baseline value. However, we found that that a combination of pieces of information that increase the SD did not necessarily increase CV.

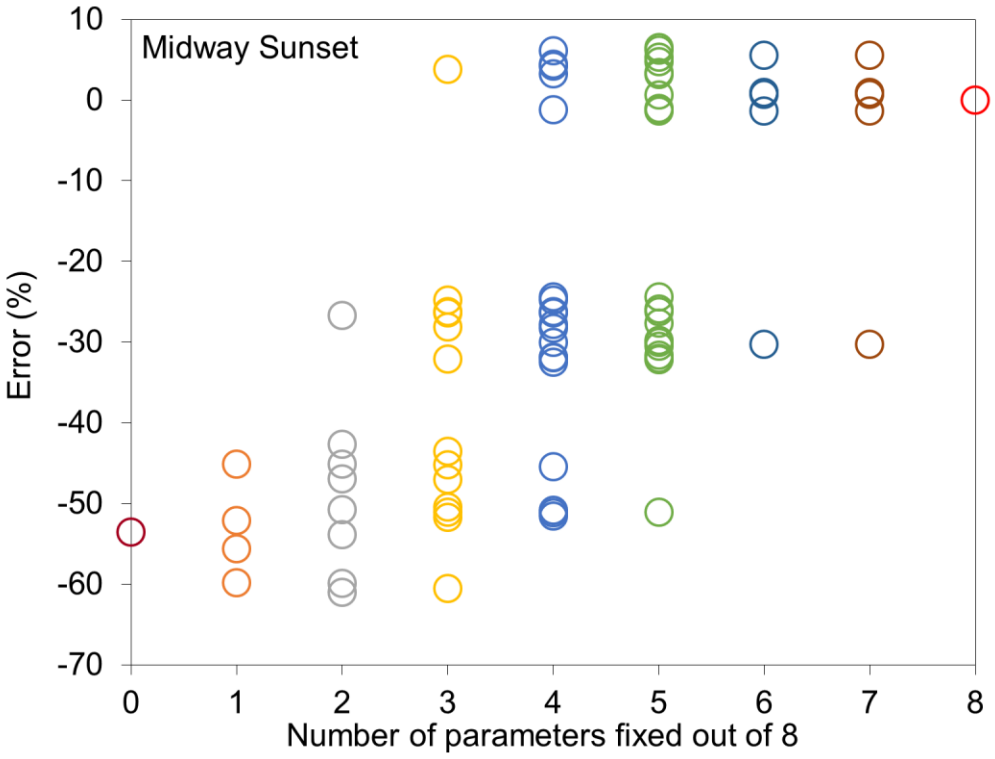


**Figure S10: Change in mean value error (bias) in Midway-Sunset field as information is learned.**

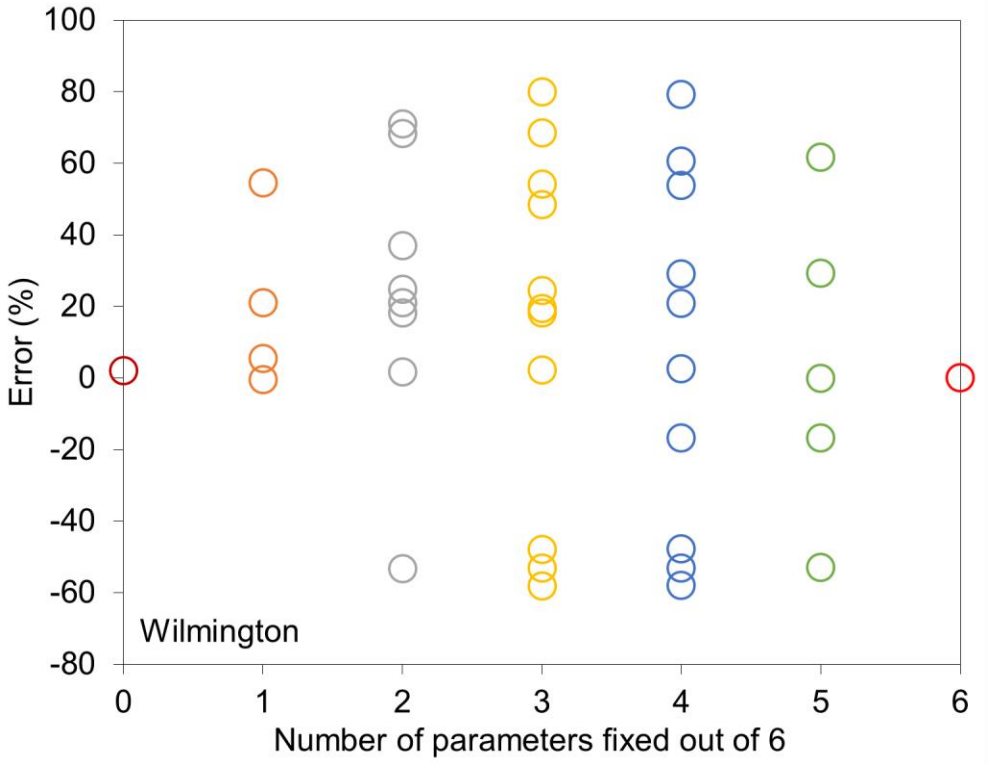


**Figure S11: Change in mean value error (bias) in Wilmington field as information is learned.**

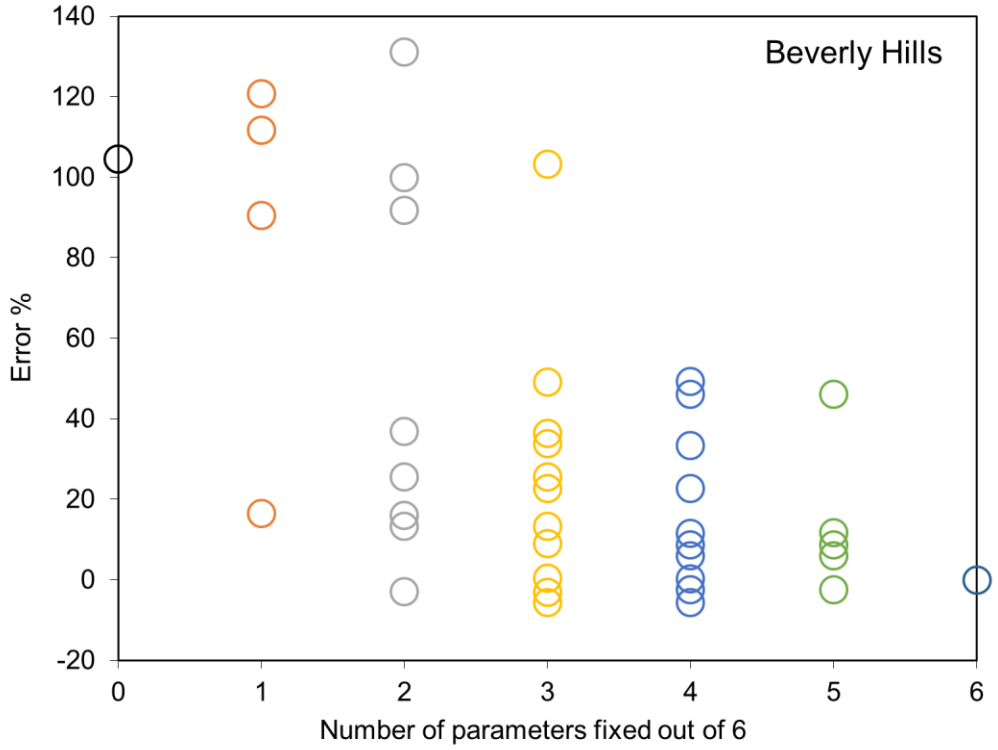


**Figure S12: Change in mean value error (bias) in Beverly Hills field as information is learned.**

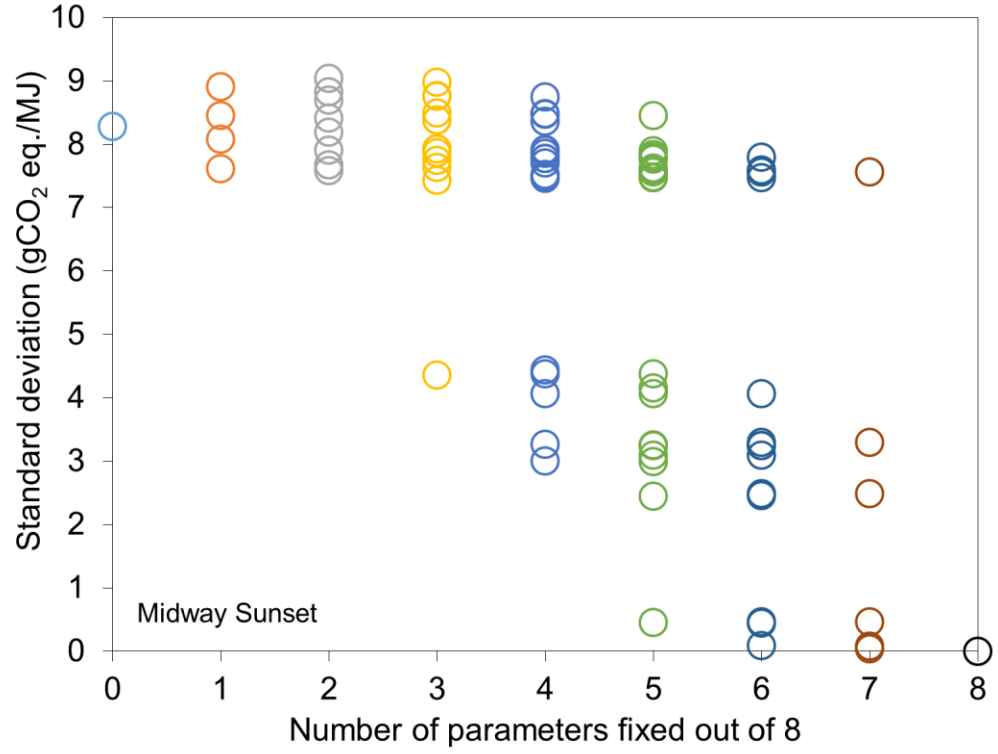


**Figure S13: Change in standard deviation in Midway-Sunset field as information is learned.**

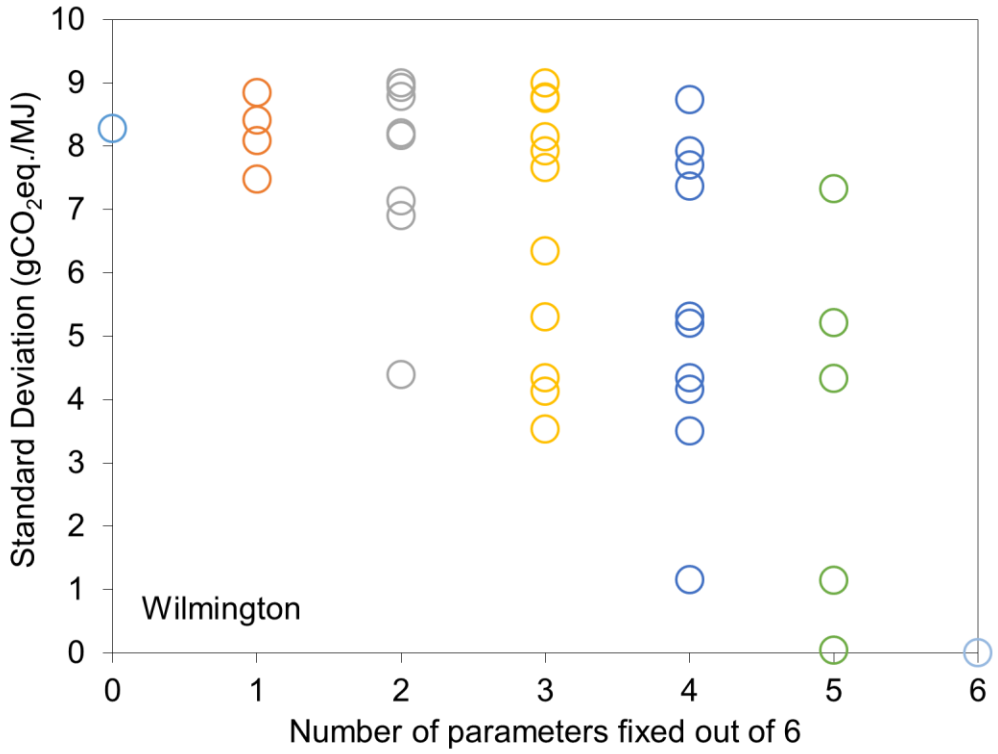


**Figure S14: Change in standard deviation in Wilmington field as information is learned.**

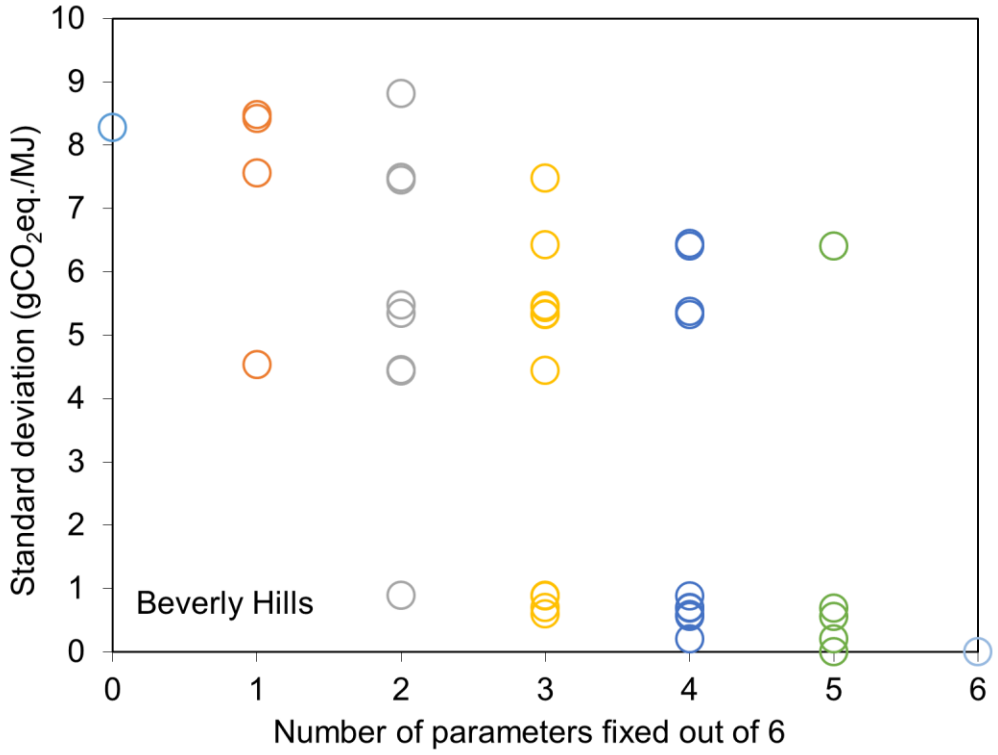


**Figure S15: Change in standard deviation in Beverly Hills field as information is learned.**

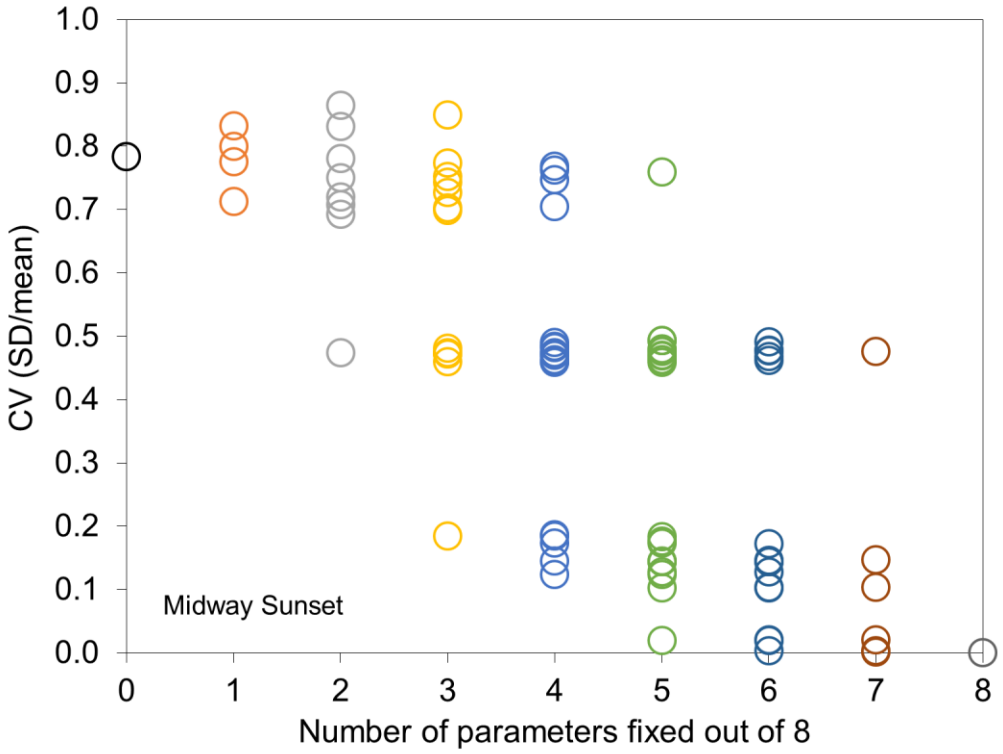


**Figure S16: Change in coefficient of variation in Midway Sunset field as information is learned.**

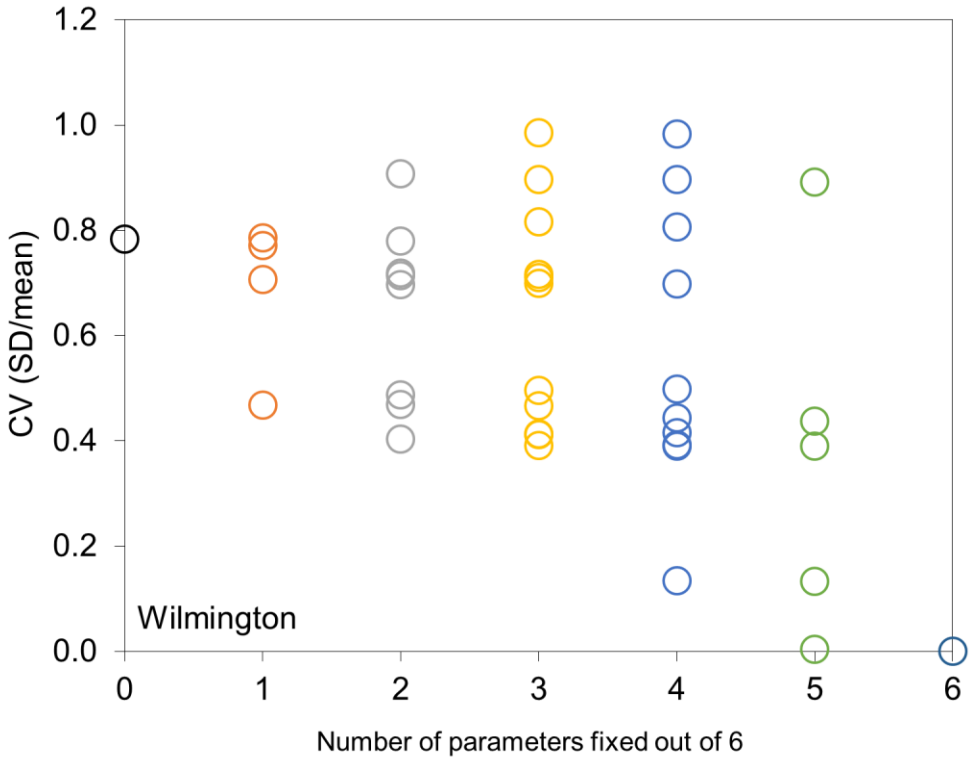


**Figure S17: Change in coefficient of variation in Wilmington field as information is learned.**

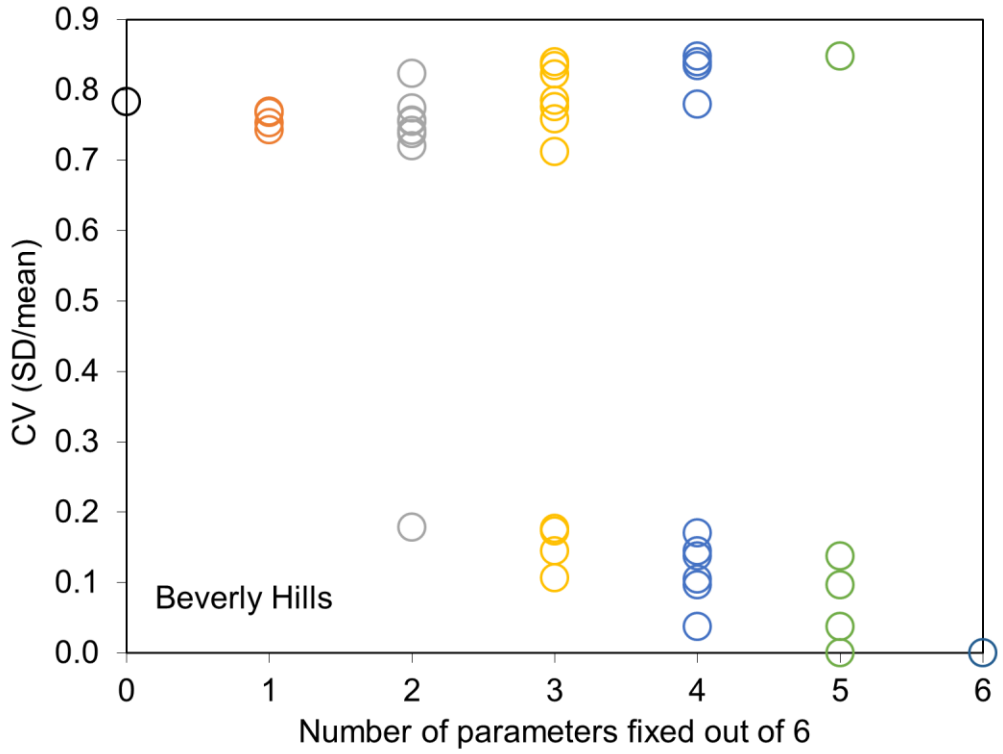


**Figure S18: Change in coefficient of variation in Beverly Hills field as information is learned.**

## S3. Convergence analysis

We investigated the required number of runs in Monte Carlo simulation which guarantees a convergence. We chose the case with all 8 input variables drawn from probability distributions. The results are depicted in Figures S19 and S20. Figure S19 shows that when MC simulation count goes above 2000, the change in mean value and SD decreases significantly. We define divergence as relative change of the mean or SD compared to its value at 10000 runs. Figure S20 shows that for number of runs of 2000 and more the divergence is less than 2%. The divergence falls below 1% when the number of runs is 6000 or more.

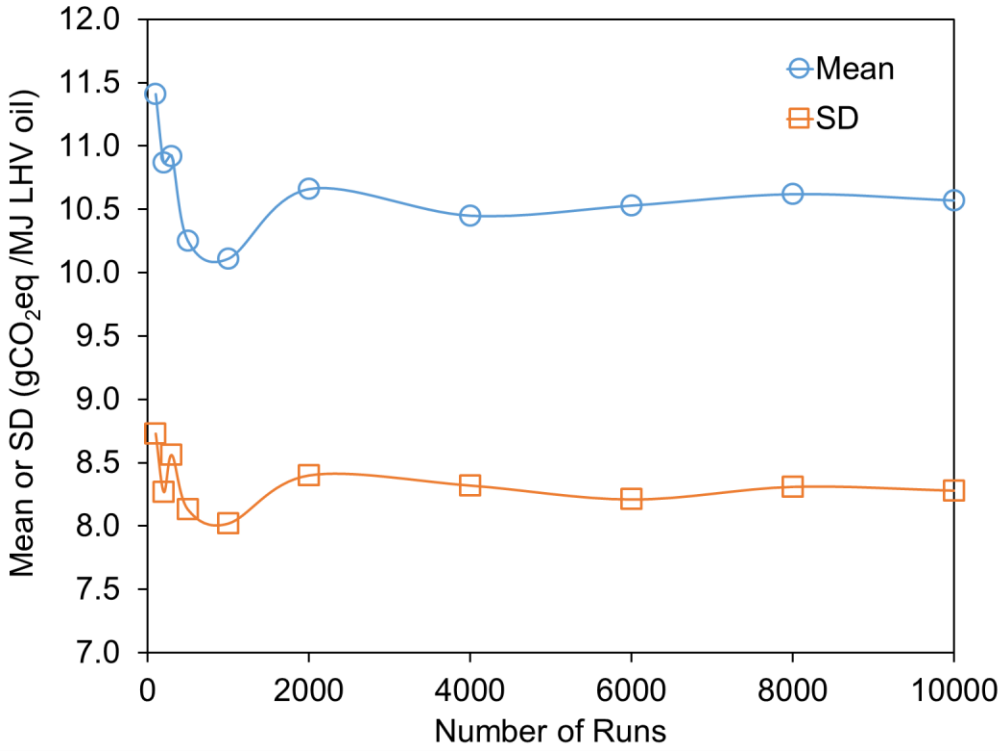


**Figure S19. Convergence of mean and SD as a function of number of Monte Carlo runs.**

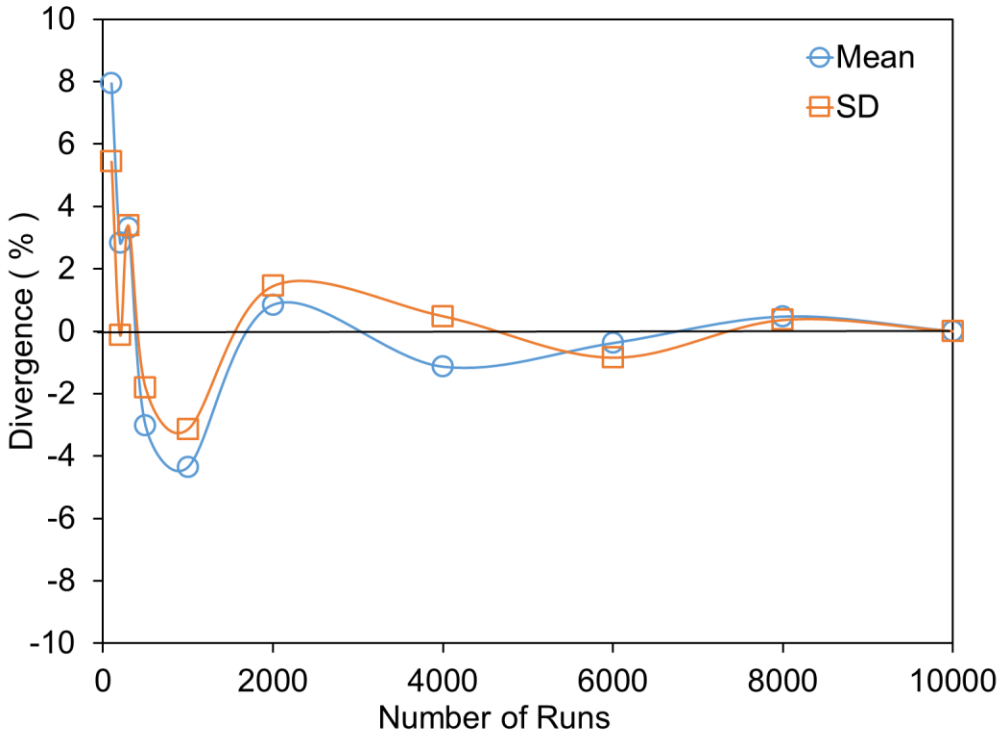


**Figure S20. Convergence of relative error in mean and SD, measured as divergence from case with 10,000 Monte Carlo runs.**

**S4. Can learning information about a system increase its uncertainty?**

It may seem unlikely that knowing a particular variable of a multivariable model can lead us to results with greater uncertainty than when the variable was not known. In this work we modeled uncertainty of model parameters by probability distributions. Naturally any mathematical description of the dispersion of data around a central tendency can represent uncertainty. In this study we generally use the standard deviation as an indicator of the level of dispersion and hence a metric to describe the level of uncertainty. We found that in some cases knowing more about an oil field (i.e., replacing probability distributions with fixed values for model parameters) can actually increase the standard deviation of the results distribution. Above, we showed that using the coefficient of variation (CV = SD/mean) results in similar patterns (though for different combinations of variables). For mathematical completeness of this study, we prove that this observation is not a mathematical inconsistency by proving the following theorem:

*Theorem:*

> *Consider a real function* $h \in R$ *of real variables* $\boldsymbol{X} \in \boldsymbol{R}$. *If* $\boldsymbol{X}$ *have probability distributions with standard deviations* $\boldsymbol{S_X}$, *then reduction of the standard deviations of the probability distributions of a subset of the variables* $\boldsymbol{X}$ *can increase the standard deviation of the resulting distribution of function h (*$\boldsymbol{X}$).

We prove this theorem by contradicting the opposite statement: "It is impossible to increase the standard deviation of the resulting distribution of function *h (****X***) by decreasing the standard deviation of any of the variables **X**.

Assume:

$$h(x, y) = \frac{x^2+1}{|y|+0.01}, \qquad \text{(S2)}$$

in which *x* and *y* are real and are normally distributed with values given in Table S4. The mean value and the standard deviation for the resulting probability distribution for *h*(*x*,*y*), based on sampling *x* and *y* 10 times is also shown in Table S4.

**Table S4. Initial parameters of normal distribution functions.**

| | ***x*** | ***y*** | ***h*** |
|---|---|---|---|
| Mean | 1 | 0.1 | 11.0 |
| Standard deviation | 1 | 1 | 32.0 |

Now we decrease standard deviation of the normal distribution that generates *y* from 1 to 0.1. Consequently, standard deviation of the model predictions, $h(x, y)$, increases from 32.0 to 73.7 (Table S5). This proves the theorem.

**Table S5. Change in probability distribution of h(x,y) due to change of $S_y$.**

| | ***x*** | ***y*** | ***h*** |
|---|---|---|---|
| Mean | 1 | 0.1 | 43.6 |
| Standard deviation | 1 | 0.1 | 73.7 |

**Table S6. Change in probability distribution of h(x,y) due to change of $S_y$.**

| | ***x*** | ***y*** | ***h*** |
|---|---|---|---|
| Mean | 1 | 0.1 | 37.6 |
| Standard deviation | 0.75 | 0.1 | 55.7 |

Figures S10 and S11 illustrates the results. Now if we further change the standard deviation of *x* from 1 to 0.75 keeping the mean value of x unchanged, the standard deviation and mean value of h(x,y) will change to 55.7 and 37.6 respectively. See Table S6 (this case is not shown in the figures). We see that a notable improvement on our knowledge about y results in significant uncertainty in the model predictions, $h(x,y)$. We see that becoming more certain about variable *x* can improve on the certainty of $h(x,y)$ yet comparing with the initial level of uncertainty (Table S4), improvement of our knowledge about both variables do not necessarily improve the certainty of $h(x,y)$. We conclude this study by demonstrating that increase in dispersion of the predictions of a model upon decreasing the dispersion of data (model inputs) is not necessarily an outcome of a mathematical inconsistency.

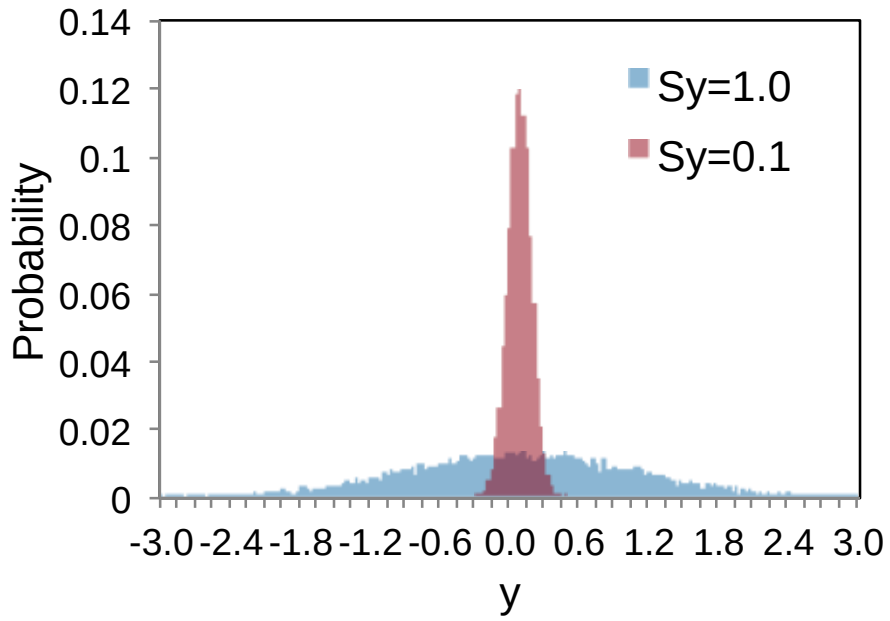


**Figure S21. As our knowledge about variable *y* of the model *h*(*x*,*y*) improves the dispersion around the central tendency decreases.**

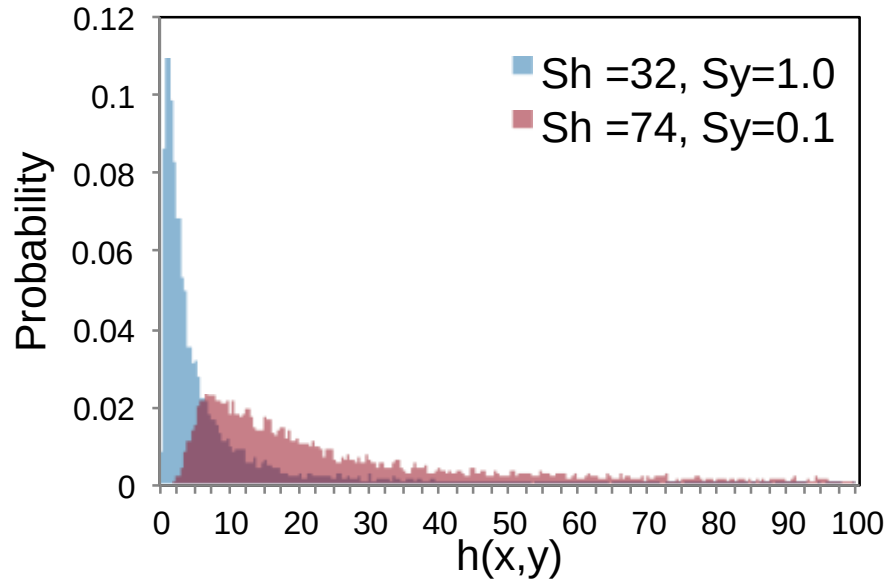


**Figure S22. As our knowledge of *y* improves the dispersion of the model predictions, *h*(*x*,*y*), around the central tendency increases.**

## S5. Best and worst paths for Wilmington and Beverly Hills oil fields

The best and worst paths for learning of information for the Wilmington and Beverly Hills fields are shown below in Figures S23 and S24.

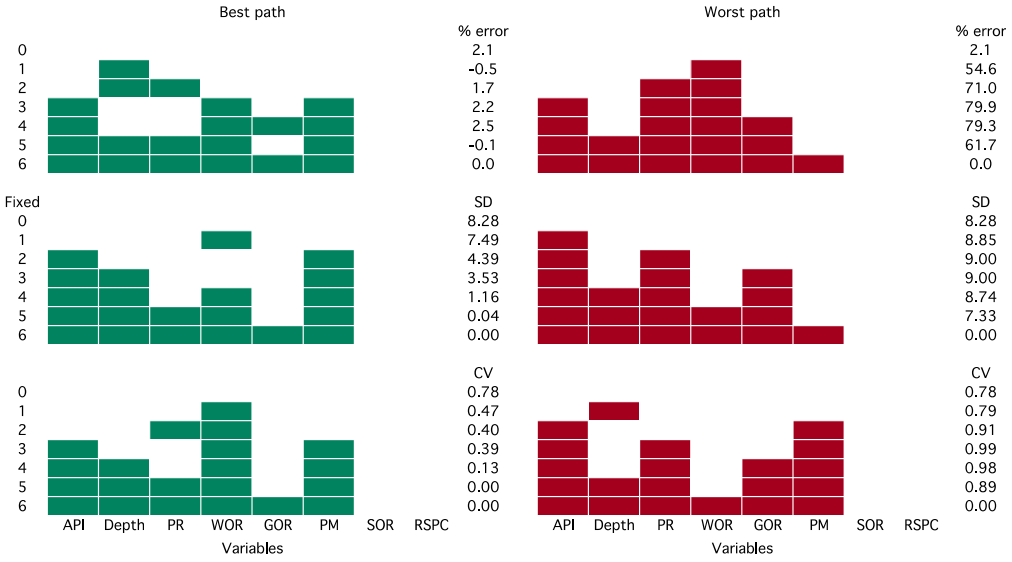


**Figure S23. Best and worst paths for the Wilmington oil field.**

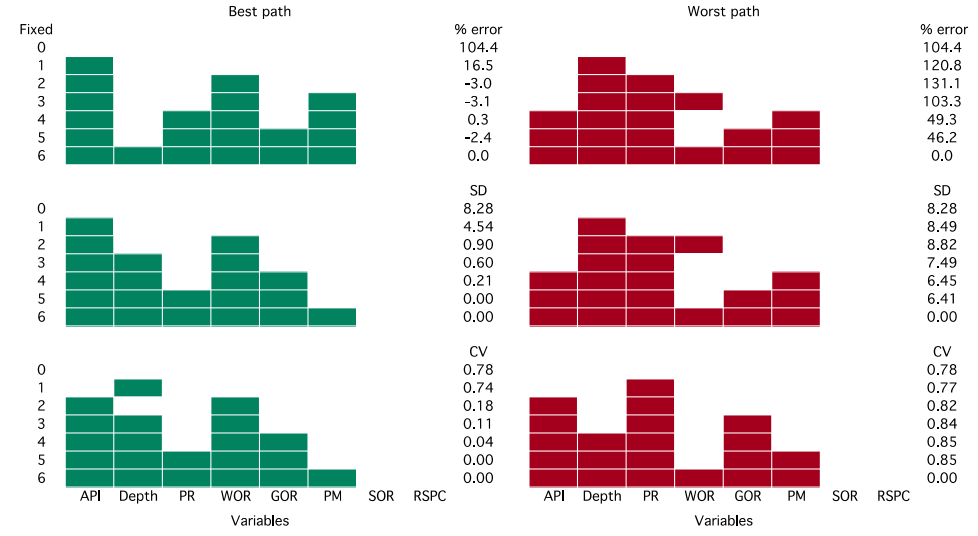


**Figure S24. Best and worst paths for the Beverly Hills oil field.**

## S6. Binary variables and bifurcation of resulting probability distributions

The use of binary variables results in bifurcation of probability distributions depending on the choice of binary variable. For example, in the Midway-Sunset field, if the variable for production method is chosen (correctly) to be thermal recovery, a very different distribution results compared to the case where the production method is chosen (incorrectly) to be non-thermal recovery. This bifurcation of distributions is illustrated upon learning the binary production method in Figure S25.

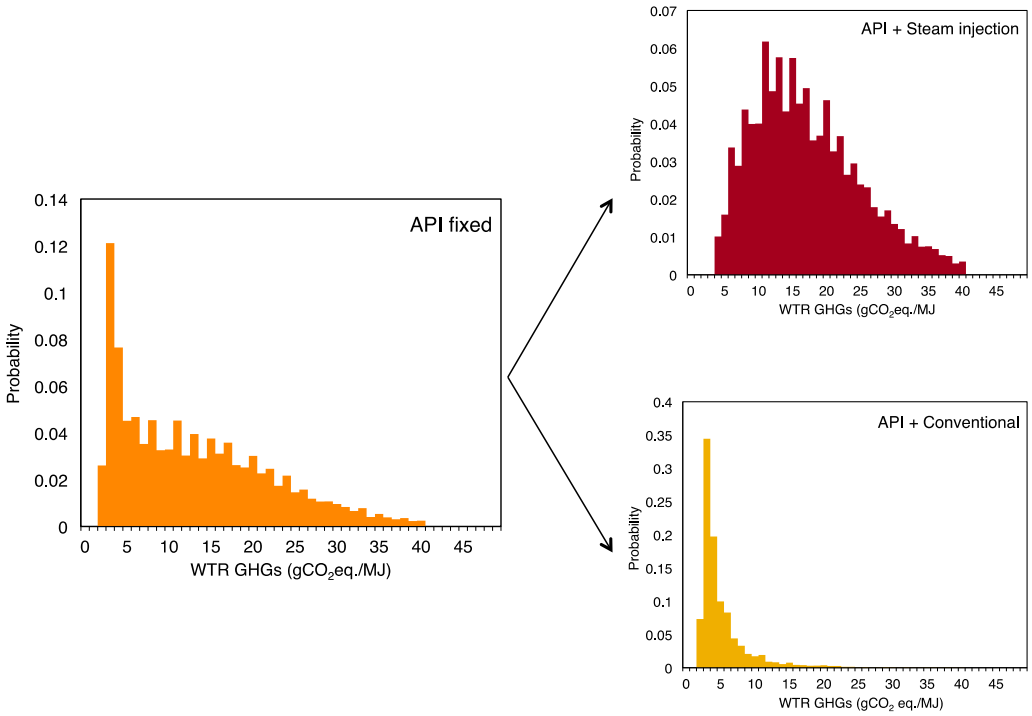


**Figure S25. Bifurcation of results distribution upon learning the method of recovery for the Midway-Sunset field. On left, only API gravity is known. On right, the field is determined to either use thermal recovery (top) or conventional recovery (bottom).**

## S7. *EasyFit* Error (Exponential Power) Distribution[8]

Parameters:
k- Shape parameter
σ- scale parameter (σ > 0)
μ- location parameter

Domain: $-\infty < x < +\infty$

Probability Density Function (PDF):

$$f(x) = c_1\, \sigma^{-1} \exp(-|c_0 z|^k) \quad \text{(S3)}$$

$$c_0 = \left(\frac{\Gamma(3/k)}{\Gamma(1/k)}\right)^{1/2}$$

$$c_1 = \frac{k\, c_0}{2\Gamma(1/k)}$$

$$z \equiv \frac{x-\mu}{\sigma}$$

## S8. Glossary

*API gravity*

API gravity is a measure of specific gravity (SG) for petroleum fluids at 60°F. The relation between API gravity and SG is given by[9]:

$$API\ gravity = \frac{141.5}{SG} - 131.5 \quad \text{(S4)}$$

*Depth*

Average vertical reservoir depth in feet.

*Gas oil ratio (GOR)*

The ratio of the produced gas volume to produced oil at standard condition (standard cubic feet of gas per barrel of oil).

*WTR GHG emissions*

Greenhouse gas (GHG) emissions from production of oil when the boundary of the analysis is from the wells to the inlet gates (WTR) of the refinery.

*Production rate*

The rate of production of oil in barrel oil per day

*Water oil ratio (WOR)*

The ratio of produced water to produced oil in barrel of water per barrel of oil

*Specific gravity (SG)*

Specific gravity is the ratio of the density of the petroleum liquid to the density of water

*Steam oil ratio (SOR)*

SOR is the water equivalent volume of steam required to produce one unit volume of oil. It is a metric of the efficiency of the oil production processes based on steam injection. Typical values of SOR for cyclic steam stimulation are often in the range of 3 to 8, while typical SOR values for steam flooding or steam assisted gravity drainage (SAGD) are often in the range of 2 to 5. The lower the SOR, the more efficiently the steam is utilized and the lower the GHG emissions.[10]

*Ratio of steam produced by cogeneration (RSPC)*

In some heavy oilfields the steam is cogenerated by the electricity which is generated by the gas turbines. The exhaust combustion gases from the gas turbine still carry recoverable heat. The recovery can be accomplished by passing the exhaust gases through a water cooled heat exchanger. This heat exchanger is called heat recovery steam generator (HRSG). The steam generated in this process is called steam by cogeneration of electricity. The cogenerated steam, however, may not amount to the total required steam in the oilfield and the rest may be generated using steam boilers. RSPC is the ratio of the steam cogenerated by electricity to the total produced steam in the oilfield.

## S9. References


(1) El-Houjeiri, H. M.; Brandt, A. R.; Duffy, J. E. Open-source LCA tool for estimating greenhouse gas emissions from crude oil production using field characteristics. *Environmental science & technology*, **2013**,*47*(11), 5998-6006.

(2) Mathwave website : http://www.mathwave.com/en/home.html

visited in April 2014

(3) Department of Conservation: California Oil and Gas Fields, Volume I,II,III http://www.conservation.ca.gov/dog/pubs_stats/Pages/technical_reports.aspx visited in April 2013

(4) California Oil and Gas Fields, Volumes I-III. California Department of Conservation, Division of Oil, Gas, and Geothermal Resources, 1982, 1992, 1998 , (GOR data collected from this source and processed by K. Clearly, California Air Resources Board)

(5) Monthly Oil and Gas Production and Injection Report, California Department of Conservation. January to December (**2010)** http://opi.consrv.ca.gov/opi/opi.dll (API data collected from this source by Duffy, J. and processed for API –GOR correlation) visited in April 2014

(6) Low Carbon Fuel Standard Program Meetings, MCON Inputs, ARB meeting on March 5th 2013, California Resources Board, Sacramento, California:
http://www.arb.ca.gov/fuels/lcfs/lcfs_meetings/lcfs_meetings.htm
visited in April 2014

(7) Hassan M. El-Houjeiri, Scott McNally, Adam R. Brandt , Oil Production Greenhouse Gas Emissions Estimator OPGEE v1.1 DRAFT A, User guide and Technical documentation, 2013.

(8) Mathwave website :
http://www.mathwave.com/help/easyfit/html/analyses/distributions/error.html
visited in August 2014

(9) Schlumberger – Oil field glossary – API gravity
http://www.glossary.oilfield.slb.com/en/Terms/a/api_gravity.aspx
visited in August 2014

(10) Schlumberger – Oil field glossary – Steam oil ration (SOR)

http://www.glossary.oilfield.slb.com/en/Terms.aspx?LookIn=term%20name&filter=steam+oil+ratio
visited in August 2014