Hierarchical Bayesian changepoint models for chemical properties inference

Amanda Koepke* Felix Jimenez† Kenneth Kroenlein‡ Chris Muzny§

Abstract

Even in scenarios where extensive historical data is available to inform estimates of chemical properties, rigorous approaches for combining these disparate but related data are unclear. We seek to provide defensible statistical methods that utilize all available data to inform both the estimate of the quantity of interest and its uncertainty. We estimate the triple point temperature for a well-studied chemical, naphthalene. Some of the literature provides information on temperature and pressure measurements, which can be understood using a changepoint model with one model for vapor pressure (liquid phase) and a different model for sublimation pressure (crystal phase). The changepoint occurs at the triple point temperature. Other works provide direct estimates and uncertainties for the triple point temperature. We develop a hierarchical Bayesian approach to estimate triple point temperature using both types of data. We implement model selection to choose between several different vapor pressure and sublimation pressure models from the literature and arrive at a comprehensive estimate of the uncertainty for the triple point temperature of naphthalene that incorporates all available information.

Key Words: Thermodynamics, Bayesian, change-point, random effects model, Hamiltonian Monte Carlo, measurement uncertainty

1. Introduction

In chemical properties research, there exists considerable historical data on quantities of interest for many chemicals. For example, for the well-studied chemical naphthalene, we have over 1000 measurements of the temperature and pressure of phase equilibrium from many studies, and we also have 172 independent estimates of the triple point temperature with uncertainties. All of this data can be used to inform estimates of certain chemical properties of naphthalene. However, it is not clear what is the best way to combine this information to obtain an estimate and a comprehensive uncertainty evaluation for that estimate.

Current methods for computing chemical properties can be *ad hoc* and are not well-defined. Rigorous, defensible statistical methods are needed for computing properties of interest and associated uncertainties. With an interdisciplinary team of scientists and engineers, we worked to develop a general statistical methodology and approach to this problem. We focus on the triple point temperature of naphthalene to describe the general approach. This study is intended to represent foundational work, where future studies will examine larger sets of related properties, tune complexity of functional forms, and address data paucity.

We propose a changepoint model for the historical measurements of pressure as a function of temperature along the phase boundary for naphthalene, with one model for vapor pressure (vapor in equilibrium with a liquid phase) and one model for sublimation pressure (vapor in equilibrium with a solid phase). Measurements at lower temperatures are on the boundary between naphthalene in vapor and solid (crystal) form, and these are modeled using sublimation pressure models. At higher temperatures, the boundary is between naphthalene in vapor and liquid form, and this boundary is modeled using vapor pressure models. The changepoint is the triple point temperature, the temperature at which the three phases can coexist. In addition to these measurements, we also have historical data on triple point measurements, measured in different ways and reported with uncertainties. The question is, what is the best way to combine data from disparate sources?

*National Institute of Standards and Technology, 325 Broadway MC898.03, Boulder, CO 80305
†University of Colorado, Department of Physics, 390 UCB, Boulder, CO 80309
‡Citrine Informatics, 702 Marshall Street, Suite 520, Redwood City, CA 94063
§National Institute of Standards and Technology, 325 Broadway MC647.07, Boulder, CO 80305

One could take a maximum-likelihood approach, fitting the parameters using a Lagrangian multiplier to enforce the constraint that the two models must meet at the changepoint. However, this method requires that the changepoint be estimated first and then fixed, ignoring its uncertainty, to estimate the rest of the parameters. We prefer to estimate all parameters simultaneously.

Instead, we develop a hierarchical Bayesian approach to incorporate data from these two very different sources. A hierarchical approach allows us to build up our model using submodels for the different data sources. We use a changepoint model for the pressure vs. temperature data, and a random effects model for the triple point temperature estimates and associated uncertainties. For the changepoint model, we use many different possible vapor pressure and sublimation pressure models from the literature.

In Section 2, we describe the historical data available for naphthalene and the various vapor pressure and sublimation models proposed in the literature. The proposed hierarchical Bayesian model is detailed in Section 3. We provide results in Section 4 for both simulated data and the historical data. We find that, for the simulated data, we are able to recover the true parameter values, but are unable to select the combination of models used to simulate the data. We discuss this in Section 5 and propose possible improvements to our approach.

2. Historical data

2.1 Pressure vs. temperature measurements

From the literature, we have pressure vs. temperature data from many different studies for both sublimation pressure and vapor pressure measurements. These measurements are shown in Figure 1. It is clear that there are some studies with data that are not consistent with the other points; these points are orange in the figure. Since the focus of this work is the development of the hierarchical model, for now we have excluded these points. For a careful analysis of this data, it will be important to explore these points more in the future.

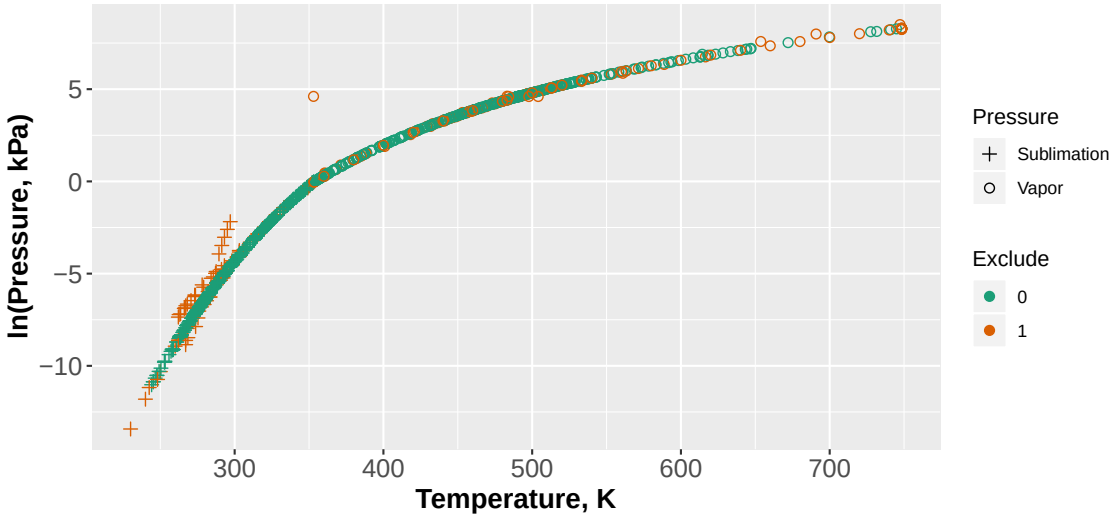

Figure 1: Pressure vs. temperature data from many different studies for both sublimation pressure and vapor pressure measurements. The transition between the type of measurement is the triple point temperature, which we do not know exactly. However, for the purpose of illustration we have roughly categorized the measurements. Sublimation pressure measurements correspond to the crosses, and vapor pressure measurements are denoted by the circles at the higher temperatures. The orange points (Exclude = 1) where excluded from the analysis.

2.1.1 Models from literature

Traditionally, a triple point temperature was evaluated independently, after which independent but consistent models were fit to the vapor pressure data and to the sublimation pressure data. The two models must meet at the triple point. There are seven candidate models to fit the vapor pressure data, and five candidate models to fit the sublimation pressure data. All models come from the literature and are semi-empirical extensions of the Clausius-Clapeyron relation. In this paper, we consider four from each set of models and will expand the analysis to include other models in future work.

The following are the vapor pressure models we considered from the literature [Reid et al., 1977, Diky et al., Yaws, 1977]. We let $y = ln(p/p^o)$ where $p^o = 1$ kPa.

- Antoine

  $y = A_1 + \frac{A_2}{T+A_3}$

- PV Expansion

  $y = A_1 + A_2/T + A_3 ln(T) + A_4 T + A_5 T^2 + A_6/T^2 + A_7 T^6 + A_8/T^4$

- DIPPR 115

  $y = A_1 + A_2/T + A_3 ln(T) + A_4 T^2 + A_5/T^2$

- Yaws Vapor Pressure

  $log_{10}(p/p^o) = A_1 + A_2/T + A_3 log_{10}(T) + A_4 T + A_5 T^2$

  $\Leftrightarrow y = ln(10)A_1 + ln(10)A_2/T + A_3 ln(T) + ln(10)A_4 T + ln(10)A_5 T^2$

The same four models are used for the sublimation pressure models, with one minor change. For sublimation pressure, the PV expansion model has one less term, so

- PV Expansion (Sublimation pressure)

  $y = a_1 + a_2/T + a_3 ln(T) + a_4 T + a_5 T^2 + a_6/T^2 + a_7 T^6.$

We use uppercase letters to denote coefficients for the vapor pressure models and lowercase letters for coefficients for the sublimation pressure models. Clearly, all of the linear models are very similar, sharing many of the exact same terms. Additionally, covariates within a linear model are highly correlated since they are all functions of temperature, suggesting multicollinearity. To deal with this, we implement a variable selection preprocessing step. We first split the data roughly at the triple point, using the mean value of the separate triple point measurements described in Section 2.2. Based on the measurements before the rough triple point temperature estimate, for each sublimation pressure model option we implement forward stepwise AIC for variable selection [Weisberg, 2005]. With the simulated and real datasets we have analyzed, usually one to three variables are selected for each model. We repeat this for the vapor pressure models and the data after the triple point. Since these models share many covariates, this usually results in the three linear models reducing to just one or two unique models.

2.2 Triple point temperature data

We have 172 estimates of triple point temperature, with uncertainties, taken from the literature. Again we exclude values from the dataset that the thermodynamics domain experts have marked as unusual, and one additional point with a suspiciously large uncertainty, leaving the 133 estimates of triple point temperature depicted in Figure 2.

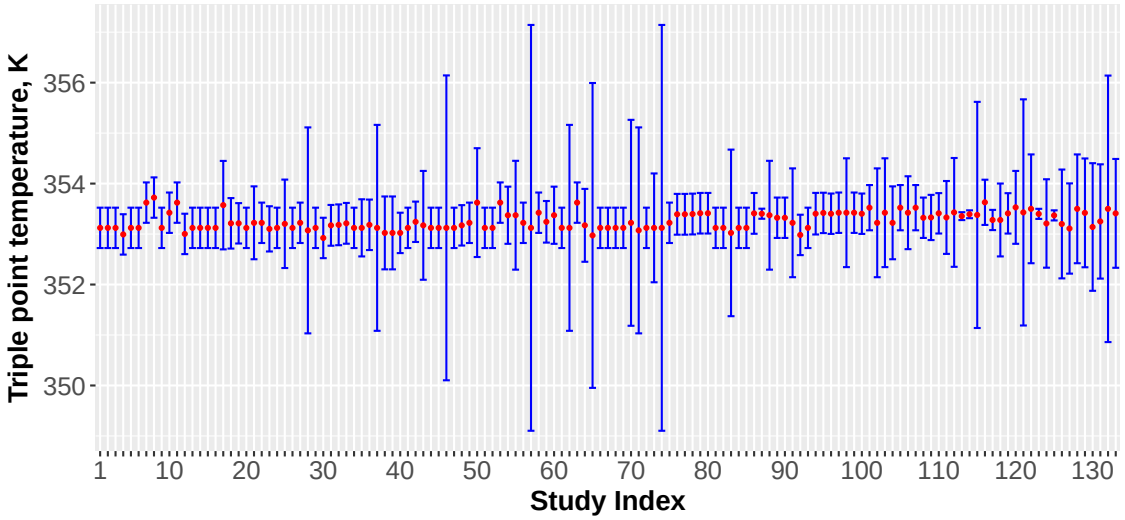

Figure 2: Estimates of triple point temperature, with uncertainties, taken from the literature. The red dot denotes the reported value of the triple point temperature. The lower edge of the blue bars denotes the reported value minus two times the reported uncertainty, and the upper edge marks the reported value plus two times the reported uncertainty.

3. Hierarchical Bayesian approach

We propose a hierarchical Bayesian approach that allows us to simultaneously incorporate data from different sources when estimating parameters. With this approach, we write our model in terms of submodels, one for each source of data, and simultaneously estimate our parameters while considering all possible sources of uncertainty. We first define the submodels, detailed in the following sections. With Hamiltonian Monte Carlo (HMC), implemented via `Stan` using the `R` package `rstan` [Carpenter et al., 2017, Stan Development Team, 2018], we sample from the posterior distribution of the parameters given the observed data.

3.1 Changepoint model for pressure vs. temperature measurements

We use a Bayesian changepoint analysis to model the pressure vs. temperature measurements. We assume that for a particular temperature value, T_j,

$$f(T_j) = \begin{cases} w_1(T_j) - w_1(\gamma) + TPP & T_j \geq \gamma \\ w_2(T_j) - w_2(\gamma) + TPP & T_j < \gamma. \end{cases} \quad (1)$$

This model and the relevant parameters are depicted in Figure 3. Here γ is the changepoint in the model (the triple point temperature) and TPP is the triple point pressure. Note that $f(\gamma) = TPP$, so this parameterization forces the two models to meet at the triple point. The functions $w_1(\cdot)$ and $w_2(\cdot)$ are chosen based on the models from the literature, but reparameterized so that they meet at the triple point rather than have individual intercept terms. For example, one model we consider for temperature values smaller than γ is $w_2(T) = b_1/T + b_2 ln(T)$, based on the reduced PV Expansion model for sublimation pressure. We have removed the intercept term from the PV expansion model, as defined earlier, and we have reduced the number of covariates using the variable selection data preprocessing step described in Section 2.1.1. Since this model is similar to the literature models defined in Section 2.1.1, but not the same, we use b instead of a to denote the coefficient parameters. We keep the convention of using uppercase letters to denote coefficients for the vapor pressure models and lowercase letters for coefficients for the sublimation pressure models. We assume that the observed log pressure values y_j have a normal distribution

centered at $f(T_j)$ and with variance τ^2 for $j = 1, \ldots, J$= 1198, where 1198 is the total number of pressure vs. temperature pairs from the literature.

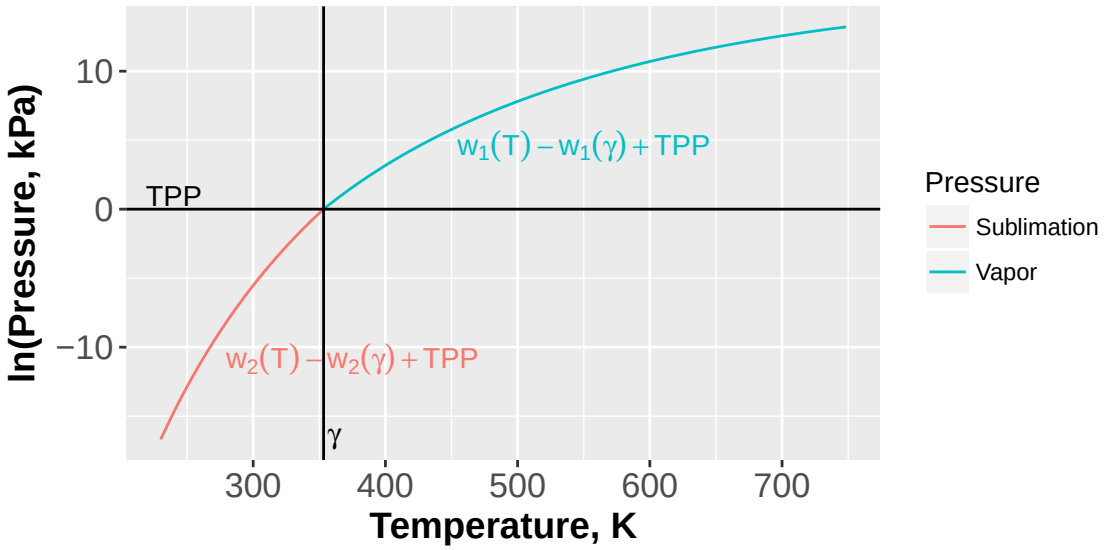

Figure 3: Changepoint model depiction. The triple point temperature, γ, is the changepoint, and the ln(Pressure) value at γ is the triple point pressure (TPP). In the absence of measurement error, the log pressure value y_j equals $f(T_j) = w_1(T_j) - w_1(\gamma) + TPP$ for temperatures above γ, and it equals $f(T_j) = w_1(T_j) - w_1(\gamma) + TPP$ for temperatures below γ.

3.2 Random effects model for triple point temperature measurements

In Section 2.2, we described the multiple measurements of triple point temperature, x_i for $i = 1, \ldots n$, available from the literature, each with a reported uncertainty, u_i. We assume these triple point temperature measurements follow a random effects model

$$x_i = \lambda_i + \epsilon_i.$$

We model the study effect as $\lambda_i \sim N(\gamma, \tau_g^2)$, where γ is the true triple point temperature and τ_g^2 is the between study variability. We model the measurement error as $\epsilon_i \sim N(0, u_i^2)$. The triple point temperature γ is shared between this random effects model and the changepoint model defined in (1).

3.3 Prior distributions

The triple point temperature, γ, has a uniform prior distribution between the minimum observed temperature, 230.034 K, and the maximum temperature, 748.332 K. Thus, *a priori* we assume that the changepoint must be in the range of observed data, but make no further assumptions. The standard deviation terms τ_g and τ are assigned diffuse prior distributions, using $\tau_g' \sim \mathrm{N}(0, 1)$ with $\tau_g = \exp(\tau_g')$ and $\tau' \sim \mathrm{N}(0, 1)$ with $\tau = \exp(\tau')$. We also assume *a priori* that the triple point pressure has a standard normal distribution; based on scientific judgement, we think value should be close to zero.

Prior distributions for the parameters in $w_1(\cdot)$ and $w_2(\cdot)$ were informed by computational challenges. For example, fitting our model was very difficult without standardizing the covariates for the linear models by subtracting the mean and dividing by the standard deviation. With the standardized covariates, we found that a weakly informative standard normal prior distribution worked well for all the coefficient parameters in the linear models, and then we use transformations in post-processing to obtain posterior samples for the parameters of our original models.

For the nonlinear model, we were not able to define general, weakly informative priors for all situations, so we used normal prior distributions with parameters informed by another data preprocessing step. Again we split the data roughly at the triple point, then fit the Antoine model, $y = a_1 + \frac{a_2}{T+a_3}$, to the data below the triple point using the function `nlsLM` [Elzhov et al., 2016] in R [R Core Team, 2018]. Similarly, we fit the model $y = A_1 + \frac{A_2}{T+A_3}$ to the data above the triple point. Coefficient values for a_2 and A_2 were around -5000 and for a_3 and A_3 were around -50. Thus for the corresponding parameters $\{b_1, B_1\}$ and $\{b_2, B_2\}$ in the changepoint model, we used the values returned by `nlsLM` as the prior mean for these coefficient parameters, and used standard deviations of 1000 and 100, respectively. Because of the large standard deviations, we do not think these priors will have a strong influence on the posterior.

3.4 Simulated data

To test our ability to estimate the parameters of this hierarchical Bayesian model, we simulate data similar to the data found in the literature. Using simulated data allows us to compare the results of the Bayesian analysis with the true parameter values.

We first simulate 133 observed triple point temperature measurements according to the random effects model described in Section 3.2, with $\gamma = 353.26$, $\tau_g = 0.055$, and the u_i equal to what is observed in the real data. These simulated triple point temperature measurements, with uncertainties, are shown in Figure 4

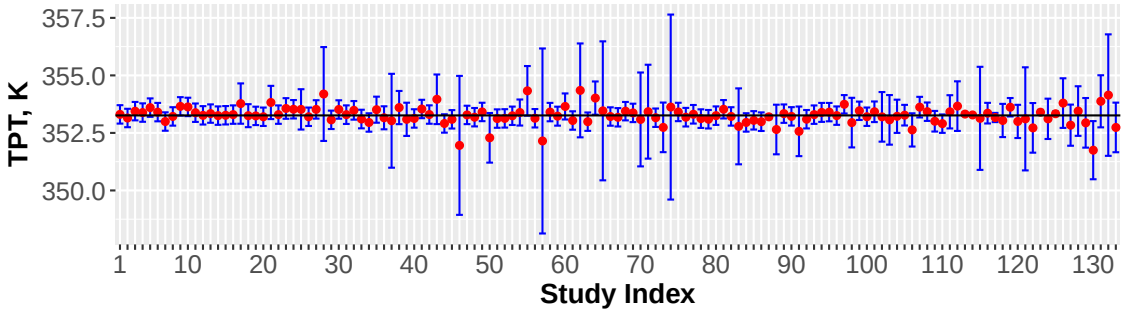

Figure 4: Simulated observed triple point temperature measurements according to the random effects model. The red dots denote the reported triple point temperature values, x_i. The blue bar marks $x_i \pm 2 \cdot u_i$, where u_i is the reported uncertainty.

We also simulate ln(pressure) data based on the observed temperature values. To generate the data below the triple point temperature, we assume $w_2(T) = b_1/T$ with $b_1 = -11000$. For the data above the triple point temperature, we assume $w_1(T) = B_1/T + B_2T^2$, with $B_1 = -10000$ and $B_2 = -4 \times 10^{-6}$. We set the standard deviation, τ, to 0.05. The simulated ln(pressure) vs. temperature measurements are shown in Figure 5.

4. Results

Analysis for each pair of models proceeds as follows. We first pick a pair of models from the literature, one for vapor pressure measurements and one for sublimation pressure. For example, the PV Expansion model for vapor pressure, $y = A_1 + A_2/T + A_3ln(T) + A_4T + A_5T^2 + A_6/T^2 + A_7T^6 + A_8/T^4$, and the DIPPR 115 model for sublimation pressure, $y = a_1 + a_2/T + a_3ln(T) + a_4T^2 + a_5/T^2$. We then implement forward stepwise AIC for variable selection. The selected covariates are then used in the hierarchical Bayesian changepoint analysis. For all pairs of models we use HMC to sample from the posterior distribution, using three chains, 500 burn-in iterations, and 500 iterations after burn-in.

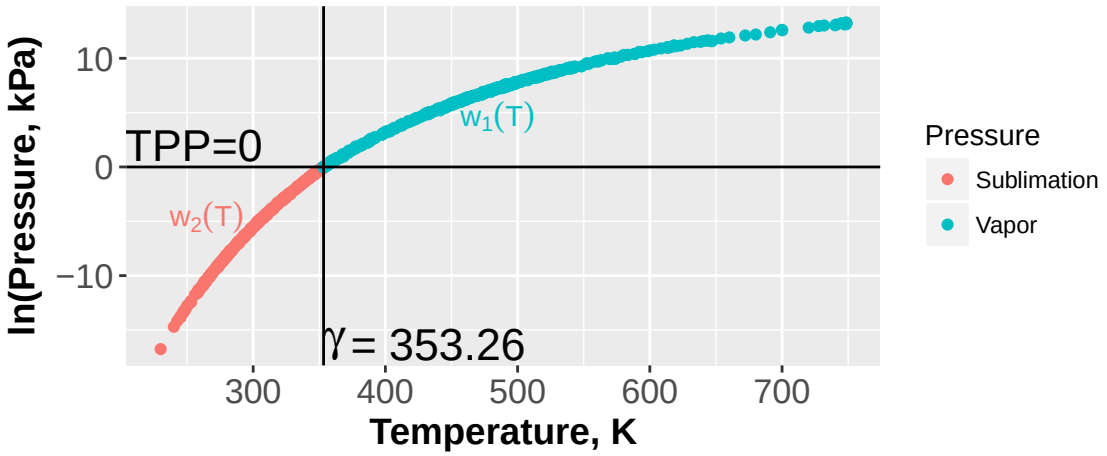

Figure 5: Simulated ln(pressure) vs. temperature measurements.

4.1 Simulated data

We first present results for the data simulated as detailed in Section 3.4. After implementing the variable selection preprocessing step for the linear models, the unique sublimation pressure models for data below the changepoint reduce to

- Model A: $w_2(T) = b_1/T$ and
- Model B: $w_2(T) = \frac{b_1}{T+b_2}$.

The unique vapor pressure models for data above the changepoint are

- Model C: $w_1(T) = B_1/T + B_2T^2$,
- Model D: $w_1(T) = B_1/T + B_2T + B_3ln(T)$, and
- Model E: $w_1(T) = \frac{B_1}{T+B_2}$.

Models B and E are Antoine models without the intercept. There are six models to fit in total, one for each combination of sublimation and vapor pressure models. Figure 6 shows histograms of the posterior samples for the parameters when Model A is used for $w_2(\cdot)$ and Model C is used for $w_1(\cdot)$. This is the set of models used to simulate the data. In Figure 6, the red line denotes the true value of the parameter used to simulated the data, the blue solid line denotes the posterior mean, and the dashed blue lines denote the 95 % credible intervals. For this simulation, the true value lies within the 95 % credible interval for all parameters. Trace plots are shown in Figure 10 in the Appendix.

To compare these models, we use the deviance information criterion (DIC) [Spiegelhalter et al., 2002, Gelman et al., 2013], which is defined as

$$\text{DIC} = -2 \log p(\boldsymbol{z}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) + 2p_{\text{DIC}},$$

where p_{DIC} is calculated as

$$p_{\text{DIC}} = 2\left(\log p(\boldsymbol{z}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S} \log p(\boldsymbol{z}|\boldsymbol{\theta}^s)\right).$$

For our example, $\boldsymbol{\theta} = (\gamma, \boldsymbol{\lambda}, \tau_g, \boldsymbol{B}, \boldsymbol{b}, \tau)$ is the vector of parameter values and $\hat{\boldsymbol{\theta}}_{\text{Bayes}}$ denotes the posterior means of these parameters. The data vector $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{T})$ is composed of the triple

point temperature measurements and uncertainties ($\boldsymbol{x}$ and $\boldsymbol{u}$), and the temperature and ln(pressure) measurements ($\boldsymbol{T}$ and $\boldsymbol{y}$)

In general, the likelihood for this hierarchical Bayesian model can be written as

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{T}|\boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{u}|\boldsymbol{\theta})p(\boldsymbol{y}, \boldsymbol{T}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\lambda_i, u_i) \prod_{j=1}^{J} p(y_j|T_j, \boldsymbol{B}, \boldsymbol{b}, \tau).$$

The probability $p(x_i|\lambda_i, u_i)$ is defined by a normal distribution with mean λ_i and variance u_i^2, and $p(y_j|T_j, \boldsymbol{B}, \boldsymbol{b}, \tau)$ is defined by a normal distribution with mean $f(T_j)$ (as defined in Equation 1) and variance τ^2. Using these distributions, we calculate the DIC values for each pair of models and display the results in Figure 7. The number in the figure denotes the rank of that DIC value, so using the Model B for the sublimation pressure ($w_2(T) = \frac{b_1}{T+b_2}$) and Model D for the vapor pressure measurements ($w_1(T) = B_1/T + B_2T + B_3ln(T)$) resulted in the lowest DIC value. Unfortunately, this model comparison approach did not suggest that the model used to simulate the data was the best model, since it only had the third lowest DIC value. However, the four models with the lowest DIC have very similar values. We comment more on the use of DIC and other possible model selection criteria in Section 5.

4.2 Historical data

We repeat the analysis using the historical data shown in Figures 1 and 2, excluding values as suggested by the subject matter experts. With the historical data, the possible models for sublimation pressure reduce to

- Model F: $w_2(T) = b_1/T + b_2T + b_3ln(T)$,
- Model G: $w_2(T) = b_1/T + b_2T^2$, and
- Model H: $w_2(T) = \frac{b_1}{T+b_2}$.

The vapor pressure models reduce to

- Model I: $w_1(T) = B_1/T + B_2/T^2 + B_3T^2$,
- Model J: $w_1(T) = B_1/T + B_2ln(T) + B_3T$,
- Model K: $w_1(T) = B_1/T + B_2ln(T) + B_3T^2$, and
- Model L: $w_1(T) = \frac{B_1}{T+B_2}$.

For this data, there are 12 possible model combinations. We again fit all of these using the hierarchical Bayesian approach and obtain DIC values for each possible combination, which are shown in Figure 8. The best model combination, according to DIC, is Model G for the sublimation pressure ($w_2(T) = b_1/T + b_2T^2$) and Model I for the vapor pressure measurements ($w_1(T) = B_1/T + B_2/T^2 + B_3T^2$). The posterior distributions for the parameters obtained using these models are depicted in Figure 9, and trace plots are shown in the Appendix.

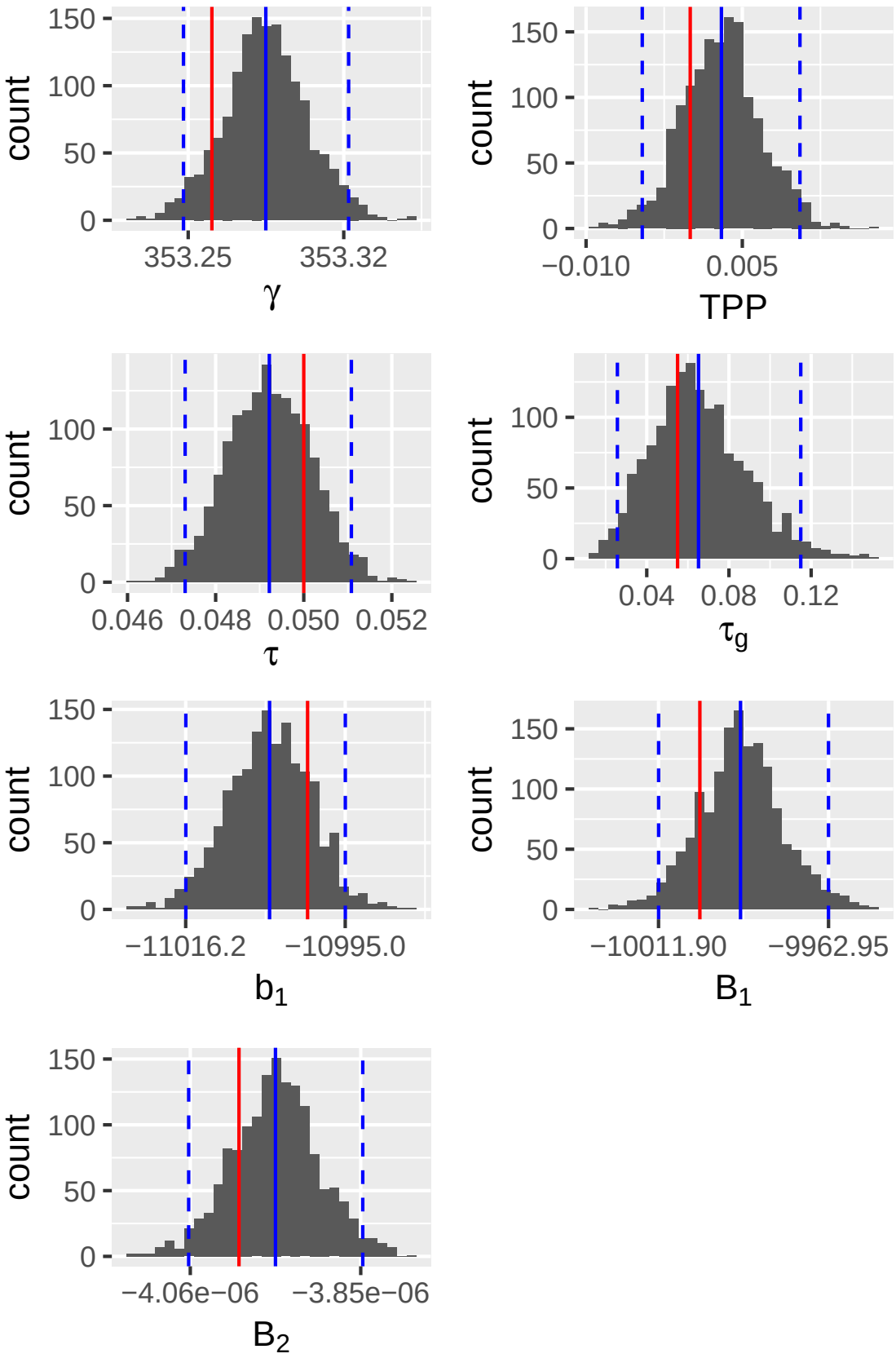

Figure 6: Simulated data results: Posterior distributions for the parameters. The solid blue line denotes the posterior mean and the dashed blue lines show the 95 % credible interval. The solid red line shows the true value of the parameter, used to simulate the data, and this value lies within the 95 % credible interval for all of the parameters.

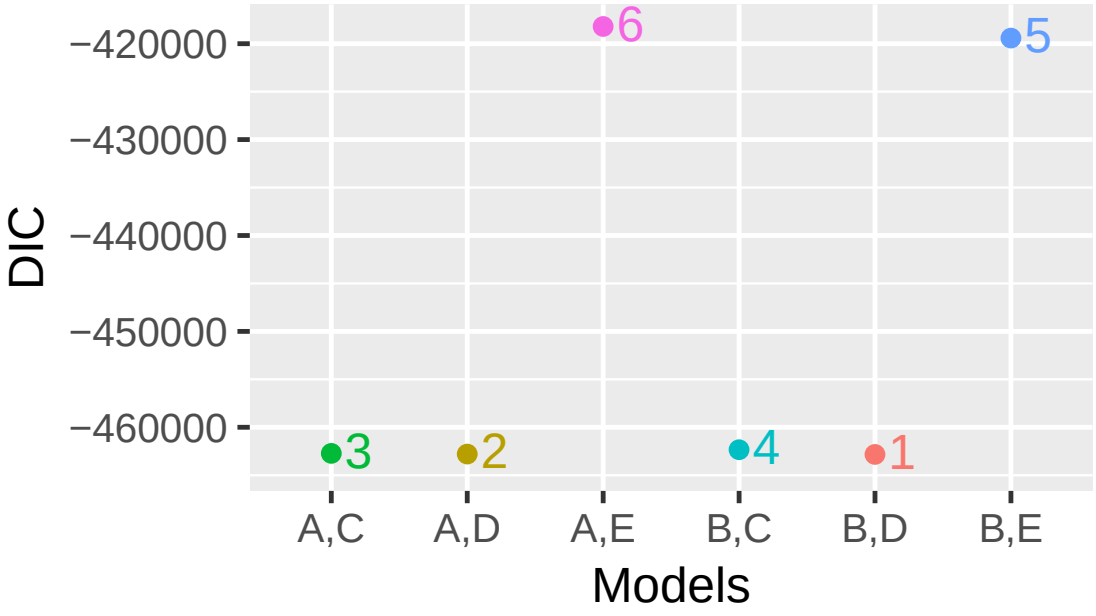

Figure 7: Simulated data results: DIC values for the six possible models. Model combinations are denoted on the x axis as (sublimation pressure model, vapor pressure model), so the label *A,C* corresponds to Model A for sublimation pressure measurements and Model C for vapor pressure measurements. The numbers denote the rank of the DIC value, so the combination with the lowest DIC value is Model B for sublimation pressure and Model D for vapor pressure.

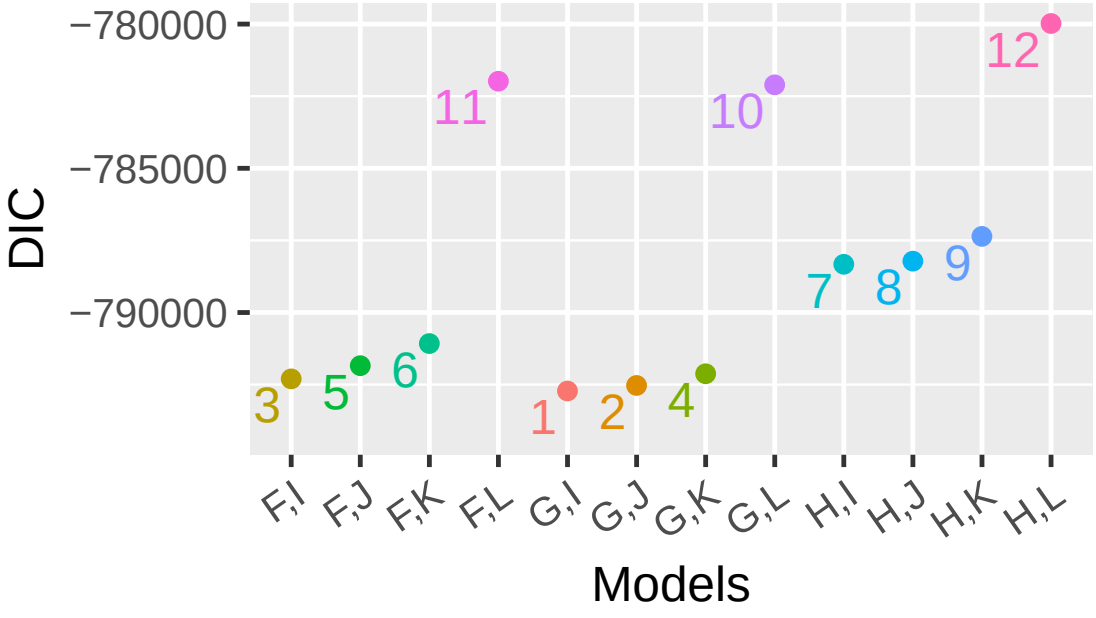

Figure 8: Historical data results: DIC values for the 12 possible models. Model combinations are denoted on the x axis as (sublimation pressure model, vapor pressure model), as detailed in Figure 7. The numbers denote the rank of the DIC value, so the combination with the lowest DIC value is Model G for sublimation pressure, and Model I for vapor pressure.

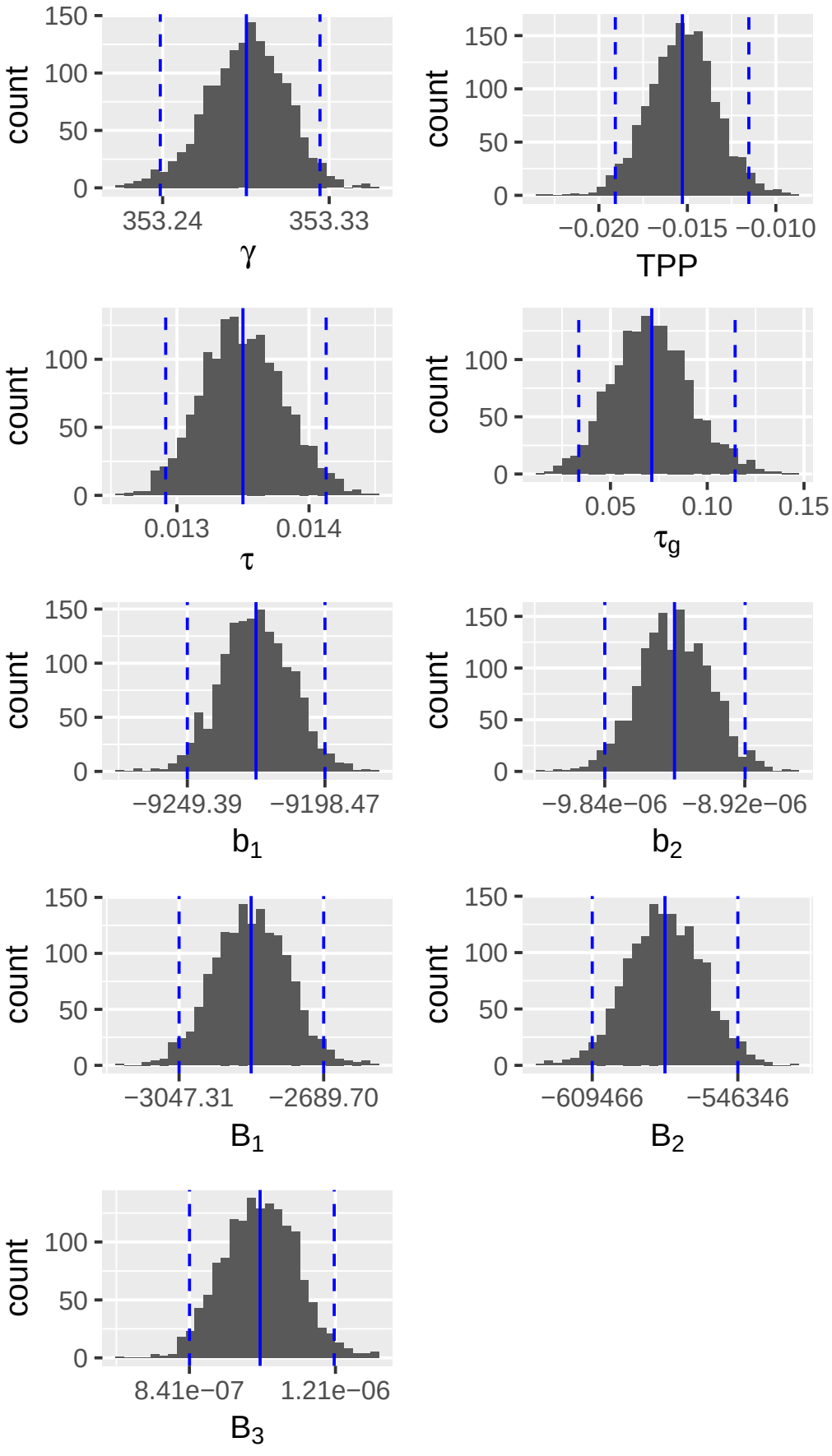

Figure 9: Historical data results: Posterior distributions for the parameters. The solid blue line denotes the posterior mean and the dashed blue lines show the 95 % credible interval.

5. Conclusions

We have proposed a hierarchical Bayesian approach for estimating chemical properties of interest for some well-behaved naphthalene data using disparate data sources. The hierarchical model allows for easy and natural incorporation of data from different sources, and also allows us to obtain parameter estimates while simultaneously accounting for all sources of uncertainty. The available naphthalene data consists of pressure vs. temperature measurements and triple point temperature measurements. We fit our hierarchical Bayesian model using both simulated data and historical data, and for the simulated data found that the true parameter values were all in their corresponding 95 % credible intervals.

Chemical properties literature offers many possible models for this data, so we also sought to select the best combination of models for a given set of data. We chose DIC for model comparison, and found that with the simulated data example we could not recover the true models used to simulate the data. It could be that these models are all so similar that they are practically indistinguishable, as is evidenced by the very similar DIC values for most model combinations. However, a better model selection criterion might be able to recover the true models. Vehtari et al. [2017] argue that DIC is not the best method for model comparison, and that leave-one-out cross-validation, the widely applicable information criterion, or K-fold cross-validation should be used instead. Leave-one-out cross-validation and the widely applicable information criterion approximations [Vehtari et al., 2017] did not work for our simulated data example, as diagnostics indicated that the approximation was not reliable. In the future we will implement the more computationally intensive K-fold cross-validation for more accurate model comparison.

Additionally, we will also work to include other nonlinear literature models in the changepoint analysis. One such model, the DIPPR 101 model, includes the term $A_4 \times T^{A_5}$ where A_4 and A_5 are both covariates to be estimated. Due to the multicollinearity issues encountered with the linear models with a similar form, it seemed ill-advised, using the current approach, to attempt to estimate these parameters in a model that also included other functions of temperature. Input from domain experts and more informative prior distributions will be necessary to estimate parameters in this more complicated model. Other models were also excluded from this analysis due to their more complicated nonlinear form, and future work will include these models in the analysis and model comparison. Once appropriate procedures have been fully developed for this well-behaved naphthalene data, we will expand this analysis to compute chemical properties of interest for some less well-behaved and well-studied chemicals. Finally, we remark that the various models explored here are not physically motivated; they were chosen at different times by different chemical properties research teams primarily for their numerical fit to the data collected in each team's study. Treating these disparate models collectively and attempting to statistically select among them may be unsound. The opportunity exists to create a well-defined nested family of models that would both ease and clarify the selection process and bring subject matter consistency to their usage.

Acknowledgements

The authors would like to thank reviewers Adam Pintar, Andrei Kazakov, Kevin Coakley, and Michael Frey for their very helpful comments.

Appendix

A. Trace plots

Figure 10 shows trace plots for the parameters fit using simulated data, and Figure 11 shows the same for the historical data analysis. Each plot shows samples drawn from three chains after 500 burn-in iterations. All chains appear to be mixing well.

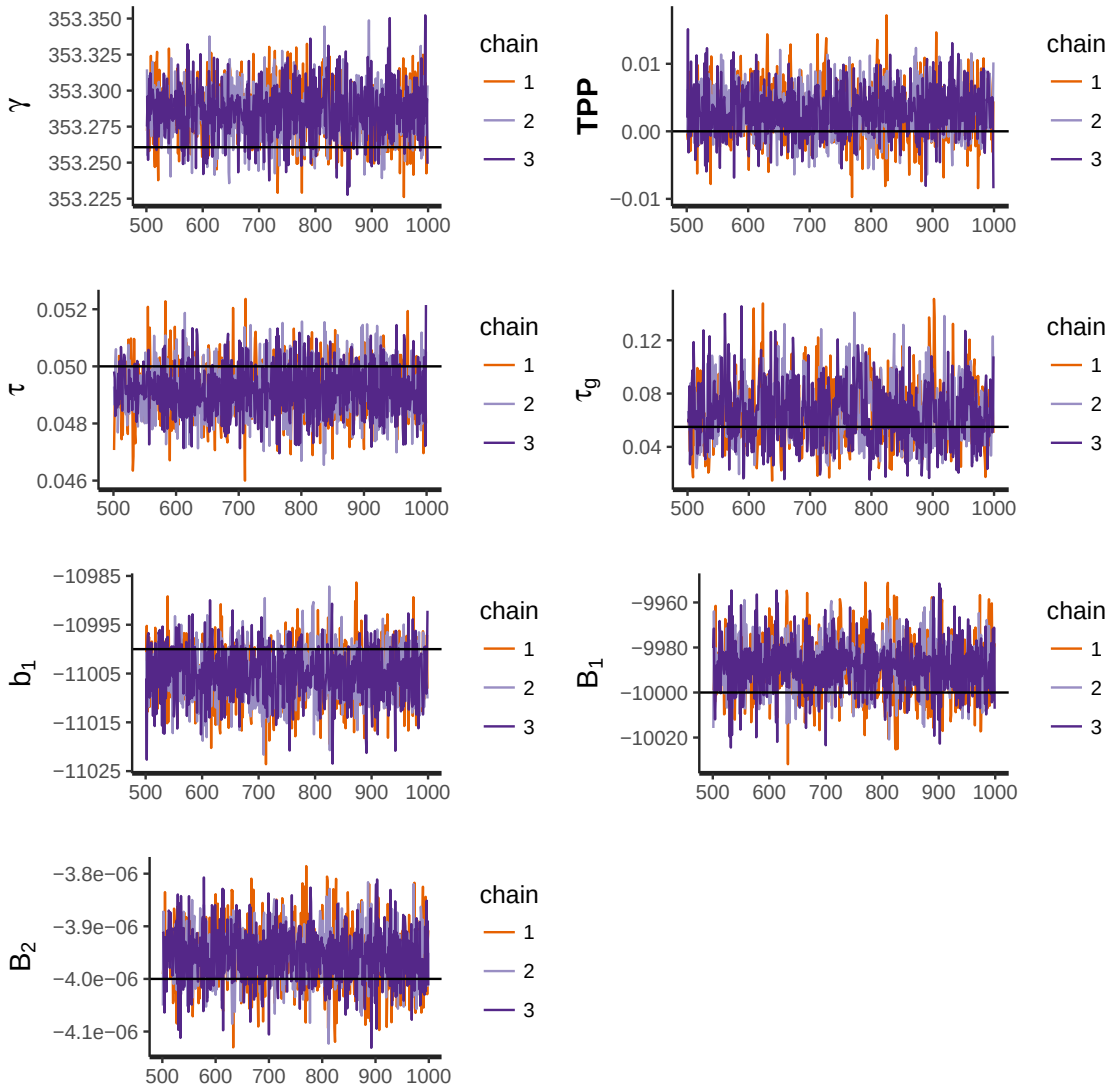

Figure 10: Trace plots for the parameters fit using the simulated data.

References

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Vladimir Diky, Robert D. Chirico, and Michael Frenkel et. al. NIST ThermoData Engine - Pure Compounds, Binary Mixtures and Reactions - SRD 103b, National Institute of Standards and Technology. Accessed: 2019-9-17.

Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess, and Ben Bolker. *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*, 2016. URL https://CRAN.R-project.org/package=minpack.lm. R package version 1.2-1.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

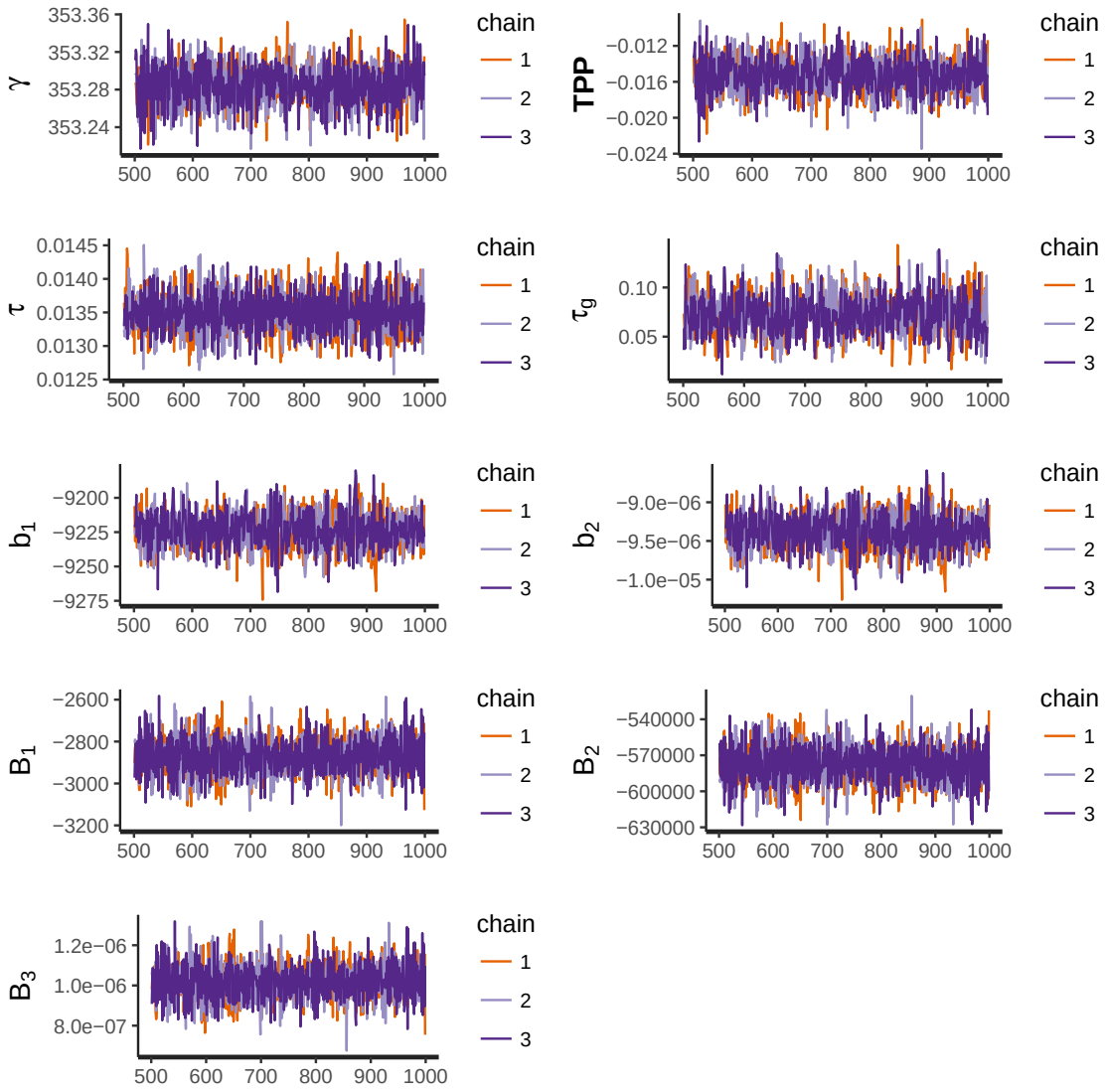

Figure 11: Trace plots for the parameters fit using the historical data.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Robert C Reid, John M Prausnitz, and Thomas K. Sherwood. *Properties of Gases and Liquids (3rd ed.)*. McGraw-Hill, New York, NY, 1977.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Stan Development Team. RStan: the R interface to Stan, 2018. URL http://mc-stan.org/. R package version 2.18.2.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. URL https://doi.org/10.1007/s11222-016-9696-4.

Sanford Weisberg. *Applied linear regression (3rd ed.)*. John Wiley & Sons, 2005.

Carl L. Yaws. *Physical Properties: A Guide to the Physical, Thermodynamics and Transport Property Data of Industrially Important Chemical Compounds*. McGraw-Hill, New York, NY, 1977.