[Geophysical Research Letters]

Supporting Information for

Obtaining the Equation of State for Multiphase Iron under Earth's Core Conditions using Bayesian Statistics

Run Wu[1], Shikai Xiang[2*],YiSun[2*], Yunting Xian[3], Yin Luo[4] , Feifan Dai[5]

National Key Laboratory of Shock Wave and Detonation Physics, Institute of Fluid Physics

China Academy of Engineering Physics, Mianyang 621999, China

Contents of this file

Text S1. Bayesian Theory and Sampling

The two main differences between Bayesian and classical statistical methods are, firstly, that Bayesian methods consider the parameters in the model as random variables, and the random distribution of the parameters can be calculated by Bayesian formulas, and secondly, that Bayesian methods can take into account not only the sample information, but also the subjective a priori information of the parameters. Under the Bayesian framework, given the data and physical model, the probability of the parameters in the model can be expressed as:

$$p(\theta|data) = \frac{p(\theta)p(data|\theta)}{p(data)} \quad (1)$$

In the Bayesian framework, $p(\theta)$ represents the prior probability distribution of the parameter θ, reflecting the initial beliefs about the parameter before any

experimental data is obtained. When there is little knowledge about the parameter, an uninformative prior such as a uniform distribution can be used. This type of prior assumes that within a specified interval, the likelihood of the parameter θ taking any value is the same. $p(data)$ denotes the probability distribution of the data, which is a normalization constant similar to the partition function in physics. This constant is necessary for sampling from the posterior distribution $p(\theta|data)$, but it does not directly affect the sampling process and therefore does not require special attention. $p(\theta|data)$ is the posterior probability distribution, describing the probability of the parameter θ after observing the experimental data. This distribution is obtained by updating the beliefs about the parameter through the combination of the prior distribution $p(\theta)$ and the likelihood function $p(\theta|data)$The likelihood function $p(data|\theta)$ represents the probability of observing the experimental data given the model parameters θ. It is commonly assumed that the error for a single data point follows a normal distribution, which means the likelihood function can be written as:

$$p(y_i|\widetilde{\mu_i};\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y_i-\widetilde{\mu_i})^2}{2{\sigma_i}^2}} \tag{2}$$

$$p(data|\theta) = \prod\prod p(y_i|\widetilde{\mu_i};\sigma_i) \tag{3}$$

The uncertainty of a parameter is associated with the experimental data and its error and model. y_i is a single experimental data point, $\widetilde{\mu_i}$ represents the physical true value, and σ_i is the corresponding error. $\widetilde{\mu_i}$ is usually replaced by the parameter-containing physical model, $\mu(\theta)$, in performing the uncertainty quantification, and thus σ_i should contain the model error, the experimental measurement error, and the random error. The model error we consider negligible if the model is sufficiently correct and reasonable. Under the assumption that the experimental data points are all considered to be independent, the total calibration data likelihood function $p(data|\theta)$ is the product of the likelihood functions of a series of individual calibration data points.

The above describes the calculation of the posterior distribution of the parameters in the model, i.e., the parameter uncertainty, in the case of a physical model with given calibration data. Usually for parameter uncertainty quantification, it is implemented by sampling the Markov-chain Monte Carlo (MCMC) method. In this work, we use the

python-emcee package to sample from the posterior distribution $p(\theta|data)$, which has been used many times in published projects in many astrophysical neighborhoods. emcee is computationally much more efficient and convergent than the standard MCMC method (Metropolis–Hastings) for sampling complex distorted probability distribution functions, due to the use of affine transformation model, which also conveniently allows multi-core CPUs to execute it in parallel.

Text S2. Treatment of Phase Boundaries

When dealing with multi-phase equations of state, the best way is to use the piece of data on the phase boundary to constrain the parameters in the model. Since the experimental data we generally measure are the two experimental variables of pressure P, temperature T, the usual practice is to use the inverse solved $P(T)$ or $T(P)$ function based on the equality of Gibbs free energies of the two phases on the boundary, or the equality of pressures and temperatures, which need to be solved inversely, and then further adding this constraint to the calculation. Of course, there are other approaches, Beth A. Lindquista and Ryan B. Jadrich were doing a parametric uncertainty analysis of the equation of state for carbon, and they derived a probability from Boltzmann statistics that could turn the points on the phase boundary into a classification problem, again with good results. In our operation, since there are more phases of iron, we still used the coexistence line model, but we did not invert the solution to solve for the functional relationship between pressure and temperature, which would also reduce the time spent. Instead, we changed our vision and added the equation constraints to the Bayesian approach to perform it.

Admitting that the experimental data all obey a Gaussian distribution, the great likelihood estimation, weighted least squares, and the Bayesian maximum probability estimation without prior information to obtain the optimal parameters should be the same from the point of view of the calibration data for the model parameters. In computing the weighted least squares.

$$\prod p_{yi} = \prod \frac{1}{\sqrt{2\pi}\sigma_{yi}} e^{-\frac{\left(y_i-\mu(\theta,x_i)\right)^2}{2{\sigma_{yi}}^2}} \tag{4}$$

$$\sum \frac{\left(y_i-\mu(\theta,x_i)\right)^2}{2{\sigma_{yi}}^2} \tag{5}$$

x_i， y_i experimental measurements corresponding to the independent variable and the dependent variable measurements corresponding to it， the experimental data and the model if they can be perfectly matched, there is no error, the second equation above the experimental measurements substituted into the calculation should be 0. That is, if we write down $y_i - \mu(\theta, x_i) = 0$, then f_i= 0, if we consider that there error, then it is not 0. In fact, this means that the magnitude of the second equation above is able to reflect the magnitude of the difference between the computed value of the model (which can also mean or the magnitude of the difference between f_i and 0) and the experimental value. If we do some complicated mathematical deformation or calculation of the equation $y_i - \mu(\vec{\theta}, x_i) = 0$, to get another equation for example written as $F_i(y_i, \vec{\theta}, x_i) = 0$, similarly F_i will be constant equal to 0 when the experimental measurements are substituted into the calculation without taking any error into account, and if there is any more error, the experimental measurements substituted into the calculation F_i is not 0. Similarly the size of the difference between F_i and 0 reflects the size of the difference between the experimental calibration data and the model calculation. That is, the constraints on the parameters in the second equation above are equivalent to the weighted least squares between the lower $F_i$and 0, as follows:

$$\sum \frac{(F_i(y_i, \theta, x_i) - 0)^2}{2{\sigma_{Fi}}^2} \tag{6}$$

Analogously the likelihood function can be obtained as:

$$p_{Fi} = \frac{1}{\sqrt{2\pi}\sigma_{Fi}} e^{-\frac{\left(F_i(y_i, \vec{\theta}, x_i) - 0\right)^2}{2{\sigma_{Fi}}^2}} \tag{7}$$

For the consideration of which σ_{Fi}, to give a special example, $F_i(y_i, \vec{\theta}, x_i) = f_i = y_i - \mu(\theta, x_i)$, that is, there is no mathematical manipulation (identity operation) of the above equation of $y_i - \mu(\theta，\ x_i) = 0$, and obviously we only need to calculate σ_{Fi} by means of error transmission: $\sigma_{Fi} = \sigma_{fi} = \sqrt{(\frac{\partial f_i}{\partial y_i}\Delta y_i)^2} = \sigma_{yi}$.Substituting these into p_{Fi}, we find that p_{Fi}=p_{yi}. We therefore generalize the idea a bit to fit the deformation of the equations, to compute σ_{Fi} by means of error transfer. in a way that is sufficient. This idea is equivalent to considering $F_i(y_i, \vec{\theta}, x_i)$ as still obeying a normal distribution, treating it as an indirectly

measured quantity, and 0 as a model theoretical computational value, and so again correctly taking into account the uncertainty introduced by the experimental measurements.

The mathematical form of the physical model is deformed and still retains the important information before the deformation. The mathematical essence of this is that when performing Monte Carlo sampling, or weighted least squares, and given the parameter $\vec{\theta}$ (at this point you can think of $F_i(y_i, \theta, x_i)$ as an indirect measure of y_i) we are going to go ahead and calculate equation (4) or equation (5). However, our approach is to use equation (6) and equation (7) to replace the computation of equation (5) and equation (4). Probabilistically, the direct and indirect measures correspond to the same value of the random variable in the sample space, and must have $p_{Fi} = p_{yi}$, so this substitution is possible. However, we must be clear that for which $\sigma_{Fi}$is considered it is estimated by error transmission, strictly speaking $\sigma_{Fi} = \left|\frac{F_i(y_i,\theta,x_i)}{y_i-\mu(\theta,x_i)}\sigma_{yi}\right|$, so this estimate is quite conservative. Another point is that this is itself an optimization tool, and after the mathematical form is morphed, the objective function is transformed from f_i to F_i, and the problem of finding the extremes of equation (4) and equation (5) is transformed into the problem of finding the extremes of equation (7) and equation (6), resulting in the optimal parameters to be different from the original due to the fact that the estimation of σ_{Fi} is passed through the error, but is not rigorous (with respect to the specific mathematical form). Experimentally, however, this practice is common and still gives good estimates of σ_{Fi}.

When dealing with the EOS boundary problem, based on the equality of the two-phase Gibbs free energies, we do not have to invert the solution to obtain the $P(T)$ or $T(P)$ function. $F_i\left(y_i, \vec{\theta}, x_i\right)$ the corresponding function is the difference between the Gibbs free energies of the two neighboring phases $G_a\ (P_i, T_i)\ - G_b\ (P_i, T_i)$, with a, b marking the two neighboring phase regions, y_i, x_i corresponds to P_i, T_i. From $F_i\left(y_i, \vec{\theta}, x_i\right)$, y_i, x_i are of comparable status, and it might be possible to consider the error in both the independent and dependent variables by means of error transfer, but we only considered the error in the pressure data.

Additionally, concerning the constraints on the liquid shock temperature, one can resort to the Rankin-Hugoniot equation:

$$E_H(V_H, T_H) - E_0(V_0, T_0) - \frac{1}{2}(P_H + P_0)(V_0 - V_H) = 0$$

Here, $E, P, V, and\ T$ represent internal energy, pressure, volume, and temperature, respectively. Subscript H denotes a point along the Hugoniot curve, while subscript 0 indicates the initial state. However, the starting point for iron is the bcc structure, and we did not directly address the internal energy of bcc iron but started with the internal energy of liquid iron at the melting point, then subtracted the experimentally measured enthalpy difference (1.3 kJ/g) (Anderson & Ahrens, 1994) to determine $E_0(V_0, T_0)$ for bcc iron.

$$E_0(V_0, T_0) = E_{bcc}\left(\frac{1}{7.85\ g/cm^3}, 300\ K\right) = E_{liqiud}\left(\frac{1}{7.019\ g/cm^3}\ , 1811\ K\right) - 1.3\ kj/g$$

Substituting into the above Rankin-Hugoniot equation:

$$E_{liqiud}(V_H, T_H) - E_{liqiud}(\frac{1}{7.019}, 1811) + 1.3 - \frac{1}{2}(P_H + P_0)(V_0 - V_H) = 0$$

The left-hand side of the above equation can be considered an indirect measured quantity, and its error can be estimated using error propagation methods. Alternatively, without solving for the temperature explicitly, one can incorporate probabilistic constraints directly, thereby accelerating calculation speed.

Text S3. Quantitative Details of Implementation

There are some details that we must elucidate when sampling and quantifying the parameters in the equation of state of iron in a Bayesian framework:

First, In our research, we employed the equation of state model put forth by Dorogokupets et al., with the detailed aspects of this model accessible in pertinent literature. For body-centered cubic (bcc) structured iron, we specified a Curie temperature of 1043 K and assigned an average magnetic moment per atom of $B_0 = 2.22$; these values were considered fixed parameters and not subject to optimization within the model. Within the solid phase, we characterized the thermodynamic properties of each atom using a set of ten parameters. Recognizing the entropy change that occurs between the solid and liquid states, we incorporated an extra parameter when describing the liquid phase, thus necessitating the use of eleven parameters for the liquid phase representation. With the aim of ensuring that the model could relatively accurately describe the thermodynamic behavior of iron under high-temperature and high-pressure conditions, we designated the hexagonal close-

packed (hcp) structure as the reference phase region where the potential energy is zero. Throughout the entire model development process, we refrained from introducing any additional empirical parameters. Consequently, the modeling endeavor encompassed a total of 40 parameters in aggregate. By synergistically leveraging these parameters, we aimed to construct a model that would precisely reflect the thermodynamic characteristics of iron, particularly under extreme conditions.

Second, In our study, due to the absence of specific prior knowledge regarding the model parameters, we opted for a general prior distribution—a uniform distribution—which served as a preliminary assumption for these parameters. To accelerate the sampling process and swiftly enter the burn-in period, we initially utilized the Python-emcee package to conduct sampling estimates on individual phases. This initial step provided us with a rough outline of the plausible parameter ranges. Subsequently, we took the high-probability sampled values obtained from this first stage as the starting inputs for the parameter chains across all four phase regions, thereby conducting joint quantitative sampling for all phases. Additionally, we also considered employing the parameter values derived from previous experimental research conducted by Dorogokupets et al. as the starting points for our sampling, further enhancing the effectiveness and reasonableness of the sampling procedure.

Third, In this work, we use the python-emcee package to sample from the posterior distribution $p(\theta|date)$, which has been used many times in published projects in many astrophysical neighborhoods. Emcee (Foreman-Mackey et al., 2013) is computationally much more efficient and convergent than the standard MCMC method (Metropolis–Hastings) for sampling complex distorted probability distribution functions, due to the use of affine transformation models, which also conveniently allows multi-core CPUs to execute it in parallel. I used the mixed sampling from the Python-emcee package for DEMove, and DIMEMove (Boehl, 2022) the ratio corresponding to the two types of moves is (0.5:0.5), because the hybrid moves are much better than the default ones. Regarding the convergence analysis of the sample chain, emcee authors give a conservative estimate of about greater than 50 times the autocorrelation time step, we sampled the samples obtained to calculate the autocorrelation time of 4,000 steps, a total of 200,000 steps of sampling.

Fifth, for the case on the boundary, our data for the hcp-liquid boundary comes from the article (Li et al., 2020), which has more accurate thermometry data relative to the others. For the fcc-hcp-liquid boundary the data comes from the article (Morard et al., n.d.). For the bcc-fcc, bcc-hcp, and bcc-liquid boundary data read from articles (O. L. Anderson, 1986; Kaufman et al., 1963; Johnson et al., 1962).

Text S4. Probability Distribution of Parameters and Correlation Coefficients Between Parameters

After obtaining the simulation results, the direct visualization of a 40-dimensional posterior distribution is inherently challenging; consequently, we leveraged the Python library Seaborn to plot kernel density estimates for each individual parameter's marginalized distribution, thereby depicting their probability density functions in the Fig.1 . Fig.2 this plot represents a symmetric correlation matrix of 40 parameters within a multiphase equation of state, where red signifies positive correlation and blue indicates negative correlation; the darker the color, the stronger the correlation. Most pairs of strongly correlated parameters are found within the same phase, as evidenced by the diagonal blocks, for instance, in the bcc phase, V_0 and K exhibit strong positive correlation, while in the hcp phase, K and V_0, as well as K' show marked negative correlations. However, there also exist noteworthy inter-phase relationships where some parameters display significant correlations across different phases. For example, it can be observed that the Einstein characteristic temperature parameter Θ_0 and the reference energy U_0share a positive correlation between the bcc and fcc phases. This could imply that data at phase boundaries link these parameters across phases. This scenario suggests that the model's parameters are not mutually independent. The dependencies among the parameters must be taken into account to accurately reflect the underlying relationships in the model. However, we obtained sample values through sampling. After plugging in 10,000 samples into the posterior probability function, we found the parameter values corresponding to the maximum value of the posterior function, which are treated as the Maximum Posterior Probability (MPP) estimates. These estimated values are listed in the following Table 1.

Table 1. Maximum Posterior Probability (MPP) estimate of 1000 sets of parameters.

	bcc	fcc	hcp	$liquid$
V_0 (cm^3/g)	0.1267473544	0.1239622495	0.1210535874	0.1424706171
$K_0(Gpa)$	163.66348004	147.24377837	156.07389919	78.950276587
K_0'	5.5060150605	4.5688650309	5.6782779899	6.0465579669
$\Theta_0(K)$	283.60896417	199.17877870	217.61535001	229.14705084
β	1.1348028698	-0.1632751972	-0.0509793421	0.3357194740
γ_0	1.6041671999	2.1364013187	2.0599424899	2.1744112631
γ_∞	-0.364118620	-0.4110305217	0.18389759162	-2.8865018184
$e_0(10^{-6}K^{-1})$	170.81443981	143.716252101	65.0590856591	172.22209539
m	1.9272464163	1.39424465306	-0.9874161768	2.0973388146
$U_0(kJ/g)$	-0.0960101266	-0.0060216531	0	2.0383567906
α				-1.9126697751

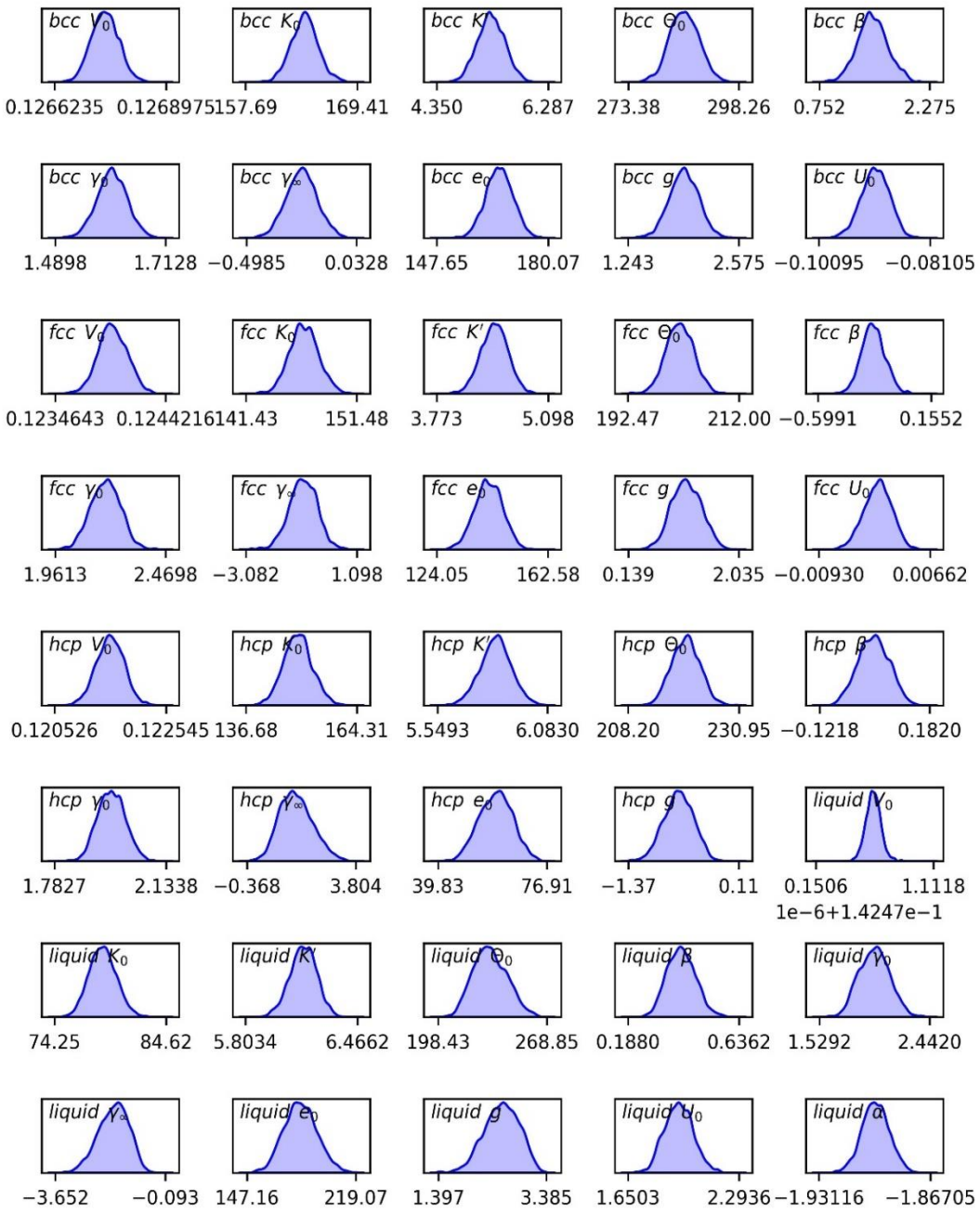

Figure 1. Plot kernel density estimates for each individual parameter's marginalized distribution for 40-dimensional parameters.

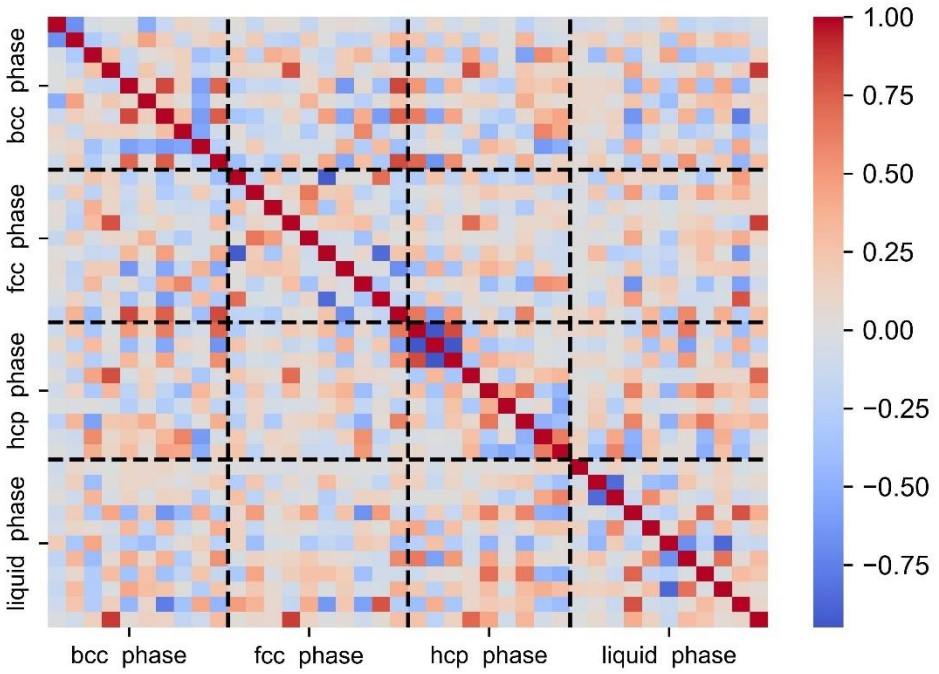

Figure 2. Correlation coefficients between 40-dimensional parameters in the multiphase equation of state for iron.

Text S5. Comparison of Computed and Experimental Values of Relevant Thermodynamic Quantities

The Fig.3 shows comparison of the calculated curve of heat capacity as a function of temperature under 0.1 MPa conditions with experimental data (Desai, 1986). It can be seen that our calculated results for the bcc structure are basically consistent with the experimental data , but the experimental data are significantly higher than our calculated data at the Curie temperature of 1043 K, which may be due to the fact that the mathematical model that describes the process of the ferromagnetic transition is still not precise enough. The calculated hot melt of the Fcc structure is in good agreement with the experimental data. of the heat capacity is in better agreement with the experimental data. The Fig.4 calculates the thermal expansion coefficients of iron in both bcc and fcc structures at a pressure of 0.1 MPa. The represent reference data (Novikova, 1974; Lu et al., 2005) from the article (Dorogokupets, 2017). As observed from the graph, the experimental data slightly exceed the calculated results, which may be due to the lack of experimental constraints on the thermal expansion coefficient for the fcc structure during the simulation process. The Fig.5 shows the comparison of the isothermal pressure lines calculated using

100 sets of parameters for solid phases with the corresponding experimental data. It can be seen that the calculation results can well reproduce the data for bcc-Fe, hcp-Fe at 15 K, bcc-Fe at 300 K, hcp-Fe at 300 K, fcc-Fe at 1073 and 1273 K (Nishihara et al., 2012). And like the melting curve, the uncertainty range of the calculated isothermal pressure lines is very small.

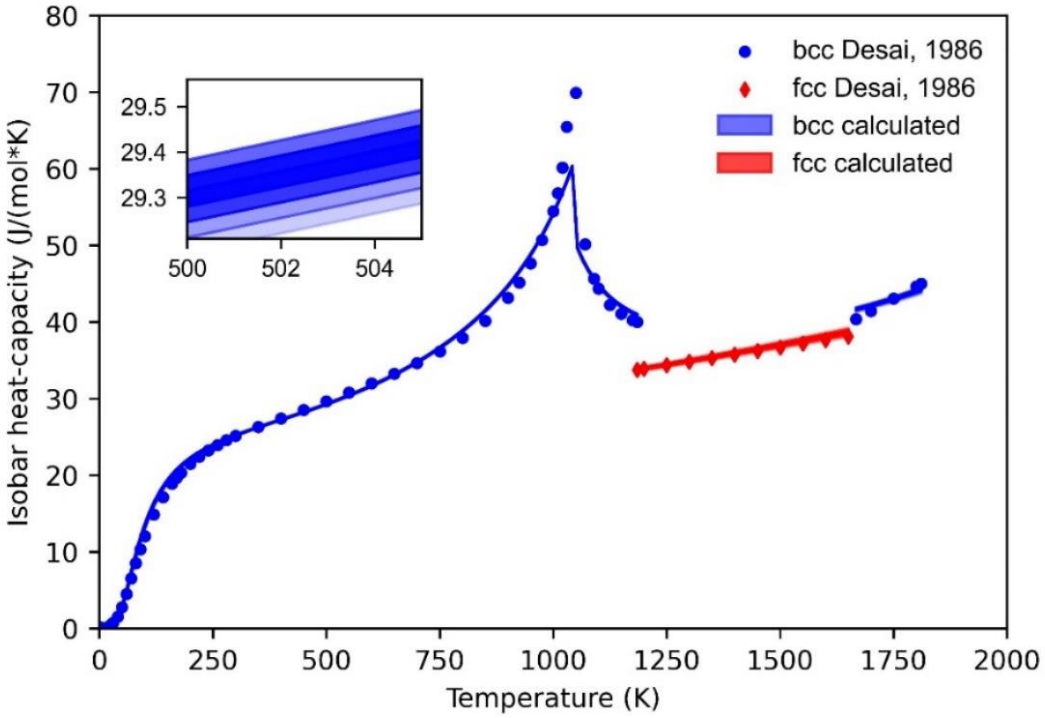

Figure 3. The figure illustrates the comparison of the calculated curve of heat capacity as a function of temperature at 0.1 MPa conditions using 100 sets of sample parameters against experimental data (Desai, 1986); only a portion of the experimental data is presented here.

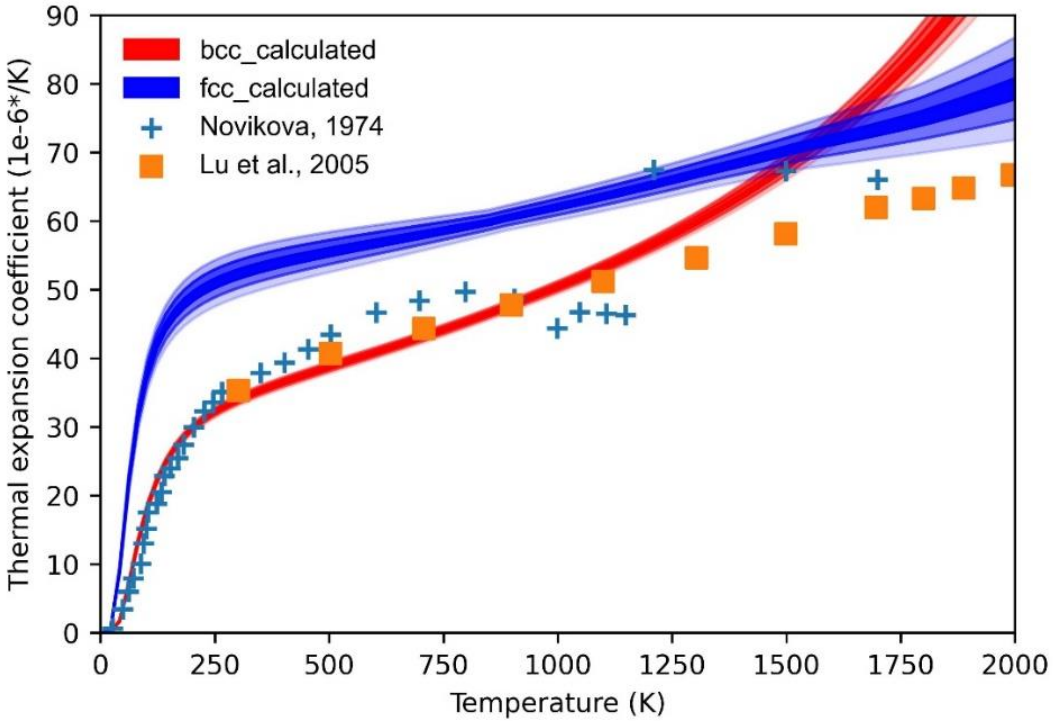

Figure 4. In this study, under 0.1 MPa conditions, the thermal expansion coefficients for bcc-Fe, fcc-Fe, and hcp-Fe structures have been calculated utilizing 100 sets of sample parameters. These computed results are intricately compared with the reference data furnished by Novikova (1974) and Lu et al. (2005), as illustrated within this figure.

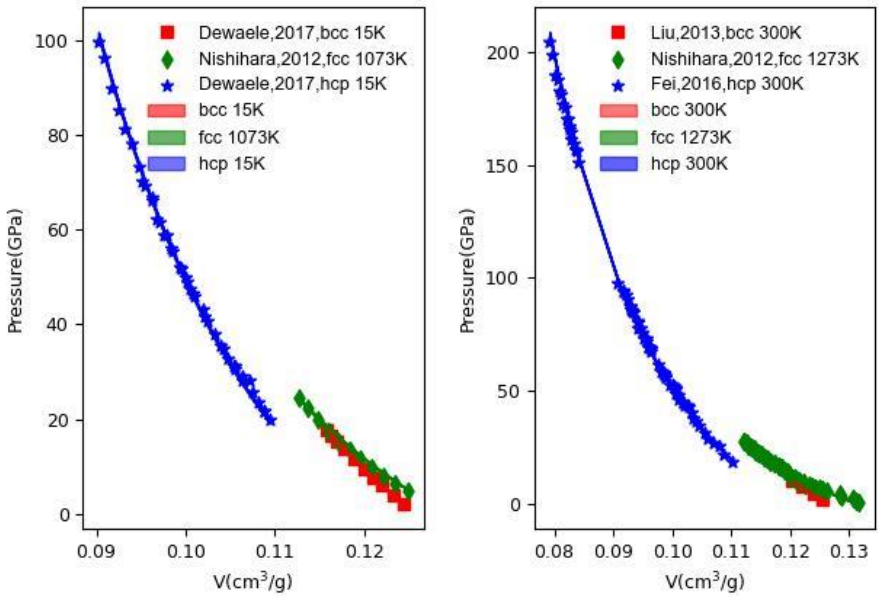

Figure 5. Comparison of the isothermal pressure lines calculated using 100 sets of parameters for solid phases with the corresponding experimental data (Dewaele & Garbarino, 2017; Liu et al., 2013; Nishihara et al., 2012; Fei et al., 2016)