university of groningen

faculty of science and engineering

artificial intelligence

Disentangling Uncertainty in Regression for Out-of-Distribution Data: A Comparative Study of Methods and Applications to Real-world Scenarios

Bachelor's Project Thesis

Muhammad Amir Bin Mohd Azman, s4013948, m.a.mohd.azman@student.rug.nl,
Supervisors: Dr M. Valdenegro-Toro

Abstract: The objective of this study is to compare and evaluate previously studied uncertainty quantification methods in out-of-distribution (OOD) data using real-world scenario regression tasks, focusing on their ability to distinguish between aleatoric and epistemic uncertainties. A comprehensive analysis is conducted using four uncertainty methods: MC-Dropout, MC-DropConnect, Flipout and Ensembles, in both numerical and visual datasets to determine their effectiveness in disentangling uncertainty. The findings from the numerical dataset indicates that variations of OOD inputs on individual features produces different relationships between ID and OOD data when using the same uncertainty method. Uncertainty methods such as Dropout and Dropconnect show positive signs of increasing epistemic uncertainty in both types of datasets. However, it is inconclusive to title them as the best general disentangling method when applied on the respective datasets. Yet, methods such as Flipout show to be unreliable.

1 Introduction

In neural networks, the primary objective is to create a representation that models the intricacies of a scenario. Thus, while the primary focus is on creating an accurate model, minimizing potential uncertainty works in conjunction with this goal. Since the role of a Deep Neural Network (DNN) requires the ability to make decisions that are reasonable in the context of a stochastic environment (Y. Gao & Su, 2021), the training data, also known as In-distribution data (ID), will have to make unforeseen predictions that may result in using data that has never been encountered, known as Out-Of-Distribution data (OOD) (Cui & Wang, 2022). Within real-world scenarios, solely using ID data is considered to be a close-world assumption. Therefore, in order to produce a practical model in an open-world scenario, it is crucial to understand the uncertainty the model makes in its predictions (Drummond & Shearer, 2006). With that in mind, there are various methods to limit uncertainty, as disentangling uncertainty aims to clarify the best possible interpretation of the source of these uncertainties.

The definition of uncertainty can be fully explained through the idea of predictive uncertainty, which can be broken down into two components for further analysis in estimating the confidence of the model (Abdar et al., 2021). The first component is aleatoric uncertainty, an irreducible uncertainty due to the non-deterministic nature of measurement errors or the inherent variability in the underlying process being modeled; randomness or noise. Consider a game of dice, you roll a fair die to determine one of the six values. In this context, aleatoric uncertainty is presented as the unpredictability of the specific number that will be rolled. The second component is Epistemic uncertainty, a reducible uncertainty due to a lack of knowledge about the perfect predictor which can be reduced by collecting more data or refining the parameters of the model. For example, if prompted to predict the meaning of the word “kichwa” in Swahili, and given the choices of ”head” or ”tails”, your initial probability might be the same or lower, however, gaining knowledge in the language can limit this uncertainty (Hüllermeier & Waegeman, 2021).

In many scenarios where regression tasks are most commonly applied, the confidence interval is equally important as the prediction itself. Since the confidence interval is a measure of reliability, it is a crucial factor in safety-critical situations (Yang et al., 2022) (Siegel, 2012), consequently, it is a necessity for the model's estimations to be properly aligned with its prediction errors (Gustafsson et al., 2023). For example, when applying a DNN on retinal photographs it notes that certain features were significantly more prominent to be correctly predicted such as gender (0.97) compared to other features with figures around 0.70. Nevertheless, the research indicates that all features are relevant yet it cannot be fully justified independently for making their prediction (Topol, 2019). In other lower-risk domain scenarios, the housing neural network applied to the California mansions depicts the evident errors that occur in the application of OOD data (Anonymous, 2020). From the perspective of a human, these errors seem nonsensical, yet the network has decided to misinterpret the information. Thus indicating that there are situations where the uncertainty of the machine reaches a limit in its knowledge and underscores the capabilities of human intelligence. In the case of the house prediction network, the features fail to capture the essential indicators to determine the actual price of the house, the result of the presence of unorthodox objects such as palm trees to the side of the house, or completely utilizing an image of a non-house object presented an absurd prediction.

Networks tend to have the most difficulty to generalize OOD data that stems from the distributional shift (Anonymous, 2020) (Gustafsson et al., 2023), these models tend to exhibit heightened uncertainty and generate potentially unreliable outputs, emphasizing the critical role of handling model selection in OOD data (J. Gao et al., 2023). To simply train a deep neural network with target $\hat{y}$ = f(x) does not encapsulate a natural dynamic environment (Gustafsson et al., 2023). Although, accounting for the inherent unpredictability in data is inevitable, and while it is challenging to ensure that the training data does not lack information, nor obtain noise, methods such as calibration are used in practical application face. The use of calibration in scaling are one step in regression tasks to reduce these issues, however, it is only a method in reducing the magnitude effects (Gal, 2016).

In order to analyze the role of the uncertainties in the OOD data, it is important to consider that epistemic uncertainty is most useful, whereas aleatoric uncertainty is not usable. Thus, this paper will probe into whether the disentanglement of uncertainty can validate the behavior of aleatoric and epistemic uncertainty when applying Out-Of-Distribution data in regression tasks. Secondly, this paper will investigate the application of different uncertainty methods to individual features impact and provide interpretation for the generation of OOD data. Finally, when assessing OOD data across both data types (numerical and visual), is it possible to generalize the uncertainty methods to disentangle the uncertainties? This research will offer a novel approach to visualizing the application of existing uncertainty methods on regression tasks, while also tackling the integration of OOD data in the disentanglement of uncertainty. To do this, the paper will proceed with an explanation of the literature background including the understanding of regression, the method to disentangle aleatoric and epistemic, the necessary uncertainty methods and previously established findings. This will be followed by the setup of our regression tasks, a deep-analysis with an interpretation of the results, concluding with the discussion and limitations that were found.

2 Literature Review

2.1 Related Work

The separation of epistemic and aleatory uncertainty have already been well documented in regression problems when attempting to evaluate a trained model solely on the data it was given. According to Depeweg et al. (2018) research, the aleatoric component of uncertainty may be distinguished from the total amount of uncertainty, leaving the epistemic uncertainty as the only remaining component. The type of data he works with involves examining specific patterns in the distribution of the data, including heteroscedasticity and bimodality. These specific patterns additions to the existing data helped achieve extracting the necessary information to disentangle uncertainty. On the other hand, using a technique called MC-Dropout, Kendall & Gal (2017) have shown they can separately measure both uncertainty components. These two methods in particular do not consider the application to OOD, therefore, research such as Valdenegro-Toro & Mori (2022) delves further into the investigation of OOD, wherein he utilizes the presence of heteroscedasticity through the use of a toy-regression dataset. To expand the disentanglement of uncertainty, Valdenegro-Toro's research employs different types of uncertainty methods to gain a more comprehensive understanding of the diverse uncertainties involved. With the current knowledge on methods in decreasing uncertainty estimates in OOD, the opportunity to take these methods into consideration to improve the validity of regression uncertainty estimation methods to real-world distribution shifts has also been recognised. Gustafsson et al. (2023) assesses multiple uncertainty estimation approaches to discover that no approach is entirely calibrated across all datasets. Some approaches become overconfident despite having good performance on baseline variations without distribution shifts. Thus, it can be argued that OOD's absolute performance through the use of disentangling uncertainty is still insufficient. In consideration of the struggles to consider overconfidence in model prediction in OOD and realizing the information gained from disentangling uncertainty, this research hopes to gain knowledge in disentangling uncertainty by applying the uncertainty methods extended by Valdenegro-Toro including Monte-Carlo Dropout, Monte-Carlo DropConnect, Flipout and Ensembles to mimic a datasets with real-world potential distribution shift by implementing OOD data (Valdenegro-Toro & Mori, 2022). The conclusion from the paper presents that Dropout is the best disentangling method while Flipout and Ensembles prove to be good indicators for epistemic uncertainty in the OOD area.

2.2 Disentangling Uncertainty

2.2.1 Regression

In the case of regression for ID data, the total uncertainty is quantified through the use of variance. Specifically, the aleatoric uncertainty uses variance of the observation error while the epistemic uncertainty uses the variance created by parameter uncertainty ($\sigma^2\epsilon$). The conditional mean by itself is not sufficient as it is only a point estimation of the target variable given the input, therefore disregarding inherent variations. In order to distinguish the residual error, the total uncertainty is deconstructed by considering the heteroscedastic aleatoric uncertainty as loss attenuation. The existence of heteroscedastic behavior occurs when variance is presented as a function of $x \in X$ (Hüllermeier & Waegeman, 2021). Kendall & Gal (2017) theory is a result of considering to penalize the prediction errors for points with high residual variance less . Thus, the difference between the calculated loss attenuation and total uncertainty results in the epistemic uncertainty.

2.2.2 Bayesian Predictive posterior

When measuring the uncertainty of a model, the weights are sampled from a distribution $p(\theta|x, y)$. For every sample of the set of weights that is produced, the model makes a prediction that creates a mean µ and variance $\sigma^2\epsilon$. The prediction is a sample of the predictive posterior distribution $p(y|x, \theta)$.

The Bayesian predictive posterior is presented as the distribution of the probability over the outputs (predictions) can be determined based on the given inputs and observed data. Given the distribution of the weights, which is also represented as the uncertainty in the parameters, the Bayesian predictive posterior can be calculated.

$$p(y|x) = \int_w P(y|w, x)P(w|D)\, dw \tag{2.1}$$

Equation 2.1 presents the theoretical interpretation of calculating the predictive posterior in a bayesian neural network (Valdenegro, 2023). The calculation involves taking the integration of the prediction of a forward pass,which considers a probability distribution of the weights, and using the probability distribution of those weights over all weighted values. However, in consideration of bayes rule, in the attempt to calculate the probability distribution of the whole data, it is intractable to take the integral over all possible weight configurations, especially when the data's dimensionality is high.

Thus, the practical method is to use the Monte-Carlo (MC) approximation.

$$P(y|x) \approx M^{-1} \int_i^M P_i(y|w,x)\,dw \qquad (2.2)$$

Equation 2.2 presents a simpler method to calculate the predictive posterior that results in an approximation, by taking M samples of forward passes. The finite number of samples will inherently result in a worst approximation, however, by taking the summation of M number of forward passes, we are able to create an approximation using the posterior distribution over weights and averaging the predictions. This will be useful in obtaining the predictive variance and predictive mean and in explaining the uncertainty methods in the next section.

$$\mu^*(x) = M^{-1} \sum_i \mu_i(x) \qquad (2.3)$$

$$\sigma_*^2(x) = M^{-1} \sum_i \left(\sigma_i^2(x) + \mu_i^2(x) - \mu_*^2(x)\right) \qquad (2.4)$$

The predictive variance can be further deconstructed into aleatoric and epistemic uncertainty. The aleatoric uncertainty can be represented as the mean of the variance while epistemic uncertainty is the variance of the means.

$$\sigma_*^2(x) = M^{-1} \sum_i \sigma_i^2(x) + M^{-1} \sum_i \mu_i^2(x) - \mu_*^2(x) \qquad (2.5)$$

2.3 Loss Functions

In probabilistic models, the Negative Log-Likelihood (NLL) loss function (equation 2.6) is frequently employed, especially for Gaussian distributions Seitzer et al. (2022). In consideration of this, the Gaussian Negative Log-Likelihood can be described through this equation:

$$L_{\text{NLL}}(y_n, x_n) = \frac{1}{2}\log(\sigma_i^2(x_n)) + \frac{(\mu_i(x_n) - y_n)^2}{2\sigma_i^2(x_n)} \qquad (2.6)$$

However, it tends to underestimate variance but trains the model's "variance heads" to quantify aleatoric uncertainty. Beta-Negative Log-Likelihood (β-NLL), shown in equation 2.7, is used to reduce this. By adding a weighted element that is responsive to variance, β-NLL alters NLL by increasing the weight on variances that are larger. Therefore, the stop() is a gradient operation that stop the backpropagation of gradient in the operations in the parentheses (Valdenegro-Toro & Mori, 2022).

$$L_{\beta-\text{NLL}}(y_n, x_n) = \text{stop}(\epsilon^{2\beta}) L_{\text{NLL}}(y_n, x_n) \qquad (2.7)$$

We can efficiently estimate the aleatoric uncertainty by training models using these loss functions, and we can then estimate the epistemic uncertainty by estimating the variance of the output of the model, giving us a complete picture of model uncertainty.

2.4 Uncertainty methods

2.4.1 Monte-Carlo Dropout

Dropout is a technique mainly used for preventing overfitting, it is usually performed only during training. During the Dropout process, each neuron in a layer has a certain probability of being "dropped out", with the exception of last layer. Monte-Carlo Dropout utilizes the Dropout in inference testing, this allows for varying predictions of the predictive mean and predictive variance for each forward pass (Gal & Zoubin, 2016). Thus, these samples are a result of the approximation of Bayesian predictive posterior (Hüllermeier & Waegeman, 2021). In this research we used the Dropout probability that produced optimized loss values (found in Appendix A.3) , we use probability values of 0.3 for air quality regression and 0.5 for age regression.

2.4.2 Monte-Carlo DropConnect

DropConnect is a regularization technique that is an extension of the Dropout method. Unlike switching off neurons in dropout, DropConnect randomly 'drops" or sets the weights of each connection between the neurons in a neural network to zero. Each weight has a probability 'p' of being set to zero (Valdenegro-Toro & Mori, 2022). Compared to dropout, this technique affects more parameters as it produces a stronger regularization effect due to being able to sample from a larger set of models using different subsets of model parameters (Mobiny et al., 2021). We use a probability value of 0.5 for air quality regression and 0.5 for the age regression.

2.4.3 Flipout

Flipout is another regularization technique that is a variant of the dropout technique, as it aims to approximate inference that introduces stochasticity through the sampling method of weights during the forward pass (Valdenegro-Toro & Mori, 2022). The technique applies perturbation to the mean weights that are modeled by the Gaussian distribution. The perturbations are affected, by reducing the variance of gradient estimates (Lee et al., 2023). The kl_weight controls the KL divergence term, thus, we set the kl_weight to a very low value in both tasks to not have a strong affect on the training process.

2.4.4 Ensembles

Ensembles implies training multiple neural networks, and when it makes a prediction, in our case, it combines the predictions of all trained based models since different instances of a random weight is initialized (Valdenegro-Toro & Mori, 2022). It assumes that these models' predictions follow a Gaussian distribution, the assumption is used to combine the predictions in a way that estimates the uncertainty of the prediction. This can be very useful in many applications where not only the prediction, but also an estimate of its reliability, is required. An Ensemble of M = 5 neural networks is used.

3 Experimental Setup

3.1 Keras Uncertainty

This paper will use a modification of the repository keras_uncertainty by Valdenegro-Toro & Mori (2022), the repository consists of utilities and models that perform Uncertainty Quantification on Keras.

3.2 Datasets

3.2.1 Age recognition dataset

The dataset consists of 2,000 images of faces that are labeled according to their ages. The ages included a range from 1 to 110 years, the distribution of the ages varies as the higher ages include less data compared to the younger ages (Rabbi, 2018). However, each age has at least one image within the range. The distribution should be not be of high importance due to the use of the transfer learning is used in the model architecture.

3.2.2 Air quality dataset

The dataset contains 9,357 hourly averaged responses recorded by a gas multisensor device that are placed in an Italian city between March 2004 and February 2005 (Learning, 2020). The dataset contains missing values in the dataset which are represented by a -200 value. The device specifically captures five important continuous features that contribute to air quality: CO, Non Methanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2) concentrations. Therefore, the target label that will be used is Total Nitrogen Oxides (NOx) as it has been stated to be a precursor to many of the other gasses (Bereitschaft, 2011). The dataset contains evidence of uncertainty in the form of cross-sensitivities, concept drifts, and sensor drifts, which may impact the sensors' concentration estimation capabilities.

3.3 Air Quality

The air quality data collection is unique as it captures many areas where uncertainty may arise. The complexity of a multi-sensor device allows for misreadings as evident errors in the recording of all the features was visible. Given that the target variable is a precursor to the other features, allows for a potentially interesting relationship in epistemic uncertainty. Thus, due to the temporal behaviour of the data, we can expect real-world changes applied in the data.

3.3.1 Model Architecture

We use an optimized classic model as the base model for the use of the uncertainty method. The model consists of three densely connected layers. From the given dataset we only considered 8 of the 13 features. The model is trained on 80% of the data and 20% is used for the inference data. When applying the uncertainty methods to the model, we see very limited changes in loss values. Therefore, stable performances may prove to be a resourceful benchmark for further analysis of the uncertainty methods (more details can be found in Appendix A.3.1 - A.3.4).

3.3.2 Out-of-Distribution data using AUC from ROC cruve

The OOD data is created by using the top 10% of the highest values of a feature and manipulating them to be larger than the largest value in that feature, this is done by adding a random value between 1 and 2. Since the normalization method ensures that each features are all values between 0 to 1. A neutral setting was created by focusing on each feature individually, only one feature was manipulated at a time.

To test if the OOD data can be discriminated by the models, Area Under the Curve (AUC) values from the Receiver Operating Characteristic (ROC) is utilized. Two seperate ROC curves are created based on the predictions of aleatoric and epistemic gained from the different uncertainty models using the two different loss functions. The ID data used were those of the test inputs. AUC values closer to 1 indicate that the model can distinguish the data types better, values closer to 0.5 indicates random guessing, lower values indicated the models inability to distinguish the data types.

3.3.3 Visualize Aleatoric and Epistemic Uncertainty

To analyze the uncertainty of the OOD data of the model, we visualize it by including 3 domains: the training, inference (ID data) and the OOD data. This is so that the graphs are able to depict the total of aleatoric and epistemic and showcase it within one graph and two separate graphs which then displays aleatoric and epistemic uncertainty. We found that for this scenario, placing the prediction of the features' mean values in increasing order allows for the visualization of the uncertainty to be clearer. In the interest of clarity, the OOD data from the ID data is differentiated by a dotted separation and is set to the largest value of the ID value in the feature. Additionally, by only analyzing the standard deviation without the mean predictions, presents us with a more comprehensible visualization to analyze the relationship between ID and OOD.

3.4 Age Recognition

Age recognition has been a challenging task since the human face contains a variety of features that make it difficult for even the human eye to recognize. For example, features such as a change in expressions have been noticed to affect the prediction significantly (Meghana et al., 2020), while more content unrelated effects such as head pose, image quality, or image resolution (ELKaraze et al., 2022) pose problems to optimizing an efficient model.

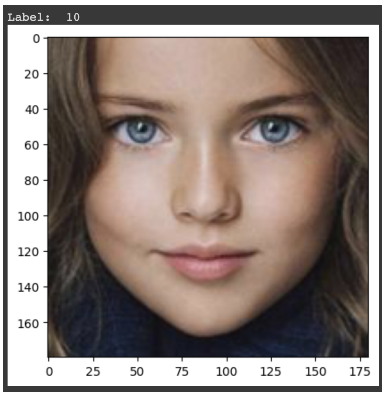


Figure 3.1: Sample Image of training (ID) data, labelled 10 years

3.4.1 Model Architecture

Residual Network (ResNet) is a deep learning model that has been developed to train extremely deep networks, specifically tasks that involve classifying images. The idea proposed by He et al. (2015) aims to directly mitigate the vanishing gradient problem by reformulating layers as learning residual functions opposed to the use of unreferenced functions. Through the use of "shortcut connections", passing the input directly to the output of a layer allows weights of early layers to be updated less frequently (Iqbal et al., 2021). Resnet-50, built with 50 layers, is most commonly utilized in classification tasks. However, due to the sophisticated nature of this dataset, applying transfer learning can provide insights to successfully train the model needed for our regression task. Therefore, utilizing a pre-trained model, specifically on the imagenet dataset, allows us to decrease the amount of data required while being able to fulfill our objective in disentangling uncertainty (Meghana et al., 2020).

Referring to Appendix A.4, it is evident that the loss is relatively high to the most efficient trained models that exist (ELKaraze et al., 2022). However, the training and test loss present decent confidence in providing information on the uncertainty based on the R^2, MSE and MAE values.

3.4.2 Out-of-Distribution data

The OOD data utilized for this task considers semantically unrelated images, specifically cars. The dataset is taken of 15 images of Audi cars from a car image dataset by Kumar (2022). Due to the inherent concept that the content of an image is the most important characteristic of a prediction by a successful model. The robustness of the model will be the focal point of the uncertainty. The analysis is visualized through uncertainty bars representing each image, we compare 15 ID images (eg. prediction for age 19) with the 15 OOD images, unlike numerical dataset it is difficult to recognize the extent of each value membership to the OOD dataset. Thus the mean prediction with the variation in uncertainty in aleatoric and epistemic of each image will be of importance.

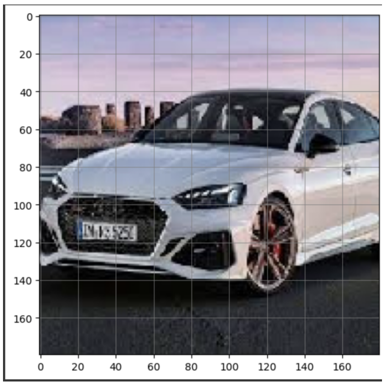


Figure 3.2: Sample Image of OOD data used in the domain

4 Air Quality Analysis

4.1 AUC values from ROC curve

Refer to Appendix B for table of values.

4.1.1 Uncertainty Methods

As expected, the classic model does not rely on aleatoric uncertainty as it inherently does not model uncertainty. Therefore, we see that the application of either loss function does not show any abrupt changes. Similarly, since Ensembles consider the predictions from multiple models, it may be able to capture more complex relationships resulting in the very high AUC value for the data in epistemic uncertainty, whereas the combination of multiple classic models does not translate in the possibility of capturing aleatoric uncertainty.

4.1.2 Loss Function

The effect of the loss function is evident in the Dropout and Flipout methods, the AUC values using NLL are significantly higher than those of the β-NLL in these two methods. The result of this clarifies the ability for these models to use aleatoric uncertainty as an identification method to distinguish the OOD and ID data. However, compared to Flipout, Dropout utilizes epistemic uncertainty substantially better to distinguish the two data types. Therefore, the models are efficiently using both aleatoric and epistemic uncertainty as a means to distinguish the OOD and ID data. Moreover, the aleatoric uncertainty can be seen to significantly increase when using DropConnect, thus, overall NLL performs as a better loss function to ensure that the model is able to distinguish between the data types.

4.1.3 Feature

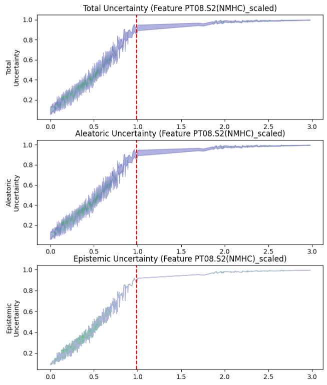
Additionally, it is important to note that there are distinct features such as PT08.S3(NOx), that show as outliers to the common trends of other features. The uncertainty methods cause no change to AUC values when using both loss functions indicating that the manipulation done to the features inputs are not sufficient enough to use aleatoric uncertainty as a distinguishing feature. This provides insights into efficiently creating OOD data while using other uncertainty methods and loss functions as a comparison tool. Therefore, to ensure that the use of the OOD data is valid, interpreting each feature by the use of either epistemic and aleatoric can be seen as beneficial to distinguish the data types. In consideration of this information, there is not a single feature that provides full reliability on both aleatoric and epistemic uncertainty using all uncertainty methods and both loss functions. However, the OOD input data for the feature CO(GT) using the NLL loss function provides promising results for both epistemic and aleatoric distinction in the creation of the OOD data. Even though, although choosing a single feature may provide a better understanding on creating OOD data using the predictions of either uncertainties, it can not be a conclusive representation when assessing the best disentangling method for the data.

4.2 ID and OOD comparison

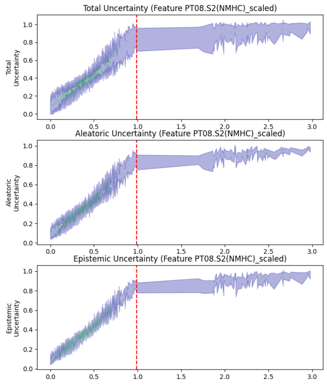
In reference to Figure 4.2, when comparing the uncertainty in the ID and OOD data per individual feature, in the classic model when utilizing the NLL and β-NLL loss functions, the total uncertainty decreases, indicating that the model is highly confident in its predictions and considers the observed data points to be similar to the original data trained on. The Dropout model when utilizing the NLL and β-NLL vary significantly, the Dropout using NLL show constant aleatoric and epistemic uncertainty while the β-NLL is evidently able to distinguish the aleatoric and epistemic uncertainty. The DropConnect model for both NLL and β-NLL do not show much difference, they both produce similar results with a difficult distinction between aleatoric and epistemic uncertainty. The Flipout model is very unreliable, in both the ID and OOD areas the uncertainty is unstable as it varies, we can see that although the epistemic uncertainty is very low, the aleatoric uncertainty is very high. Therefore, there is a clear distinction between the two uncertainties, the total uncertainty does not capture the inherent uncertainty in the original data. The impact of the β-NLL loss function did not seem to affect the uncertainty in the model. The Ensemble model in both loss functions seem to be the most clear in presenting the inability to generalize to the OOD data as an increase in epistemic uncertainty is shown.

4.2.1 Between Different Uncertainty methods

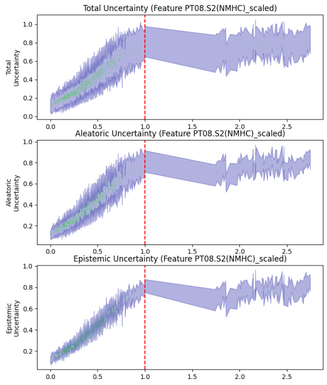
In consideration of both loss functions and each uncertainty method, arguably Dropout with the utilization of the β-NLL loss function and both Ensemble methods using β-NLL and NLL may be the best method for epistemic uncertainty. We are unable to select a single combination for the ideal disentanglement since, based on all the combinations of models and loss functions, the aleatoric uncertainty in the ID and OOD domains is, for the most part, constant.



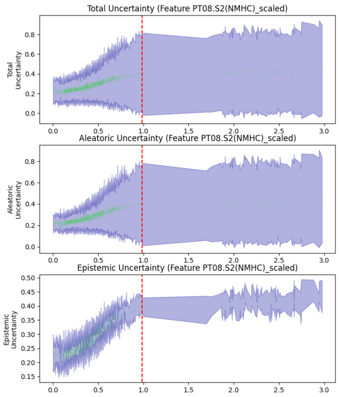
Classic Model



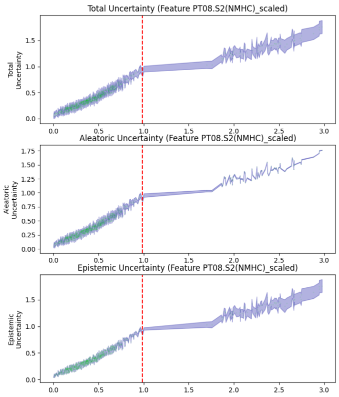
Dropout Model



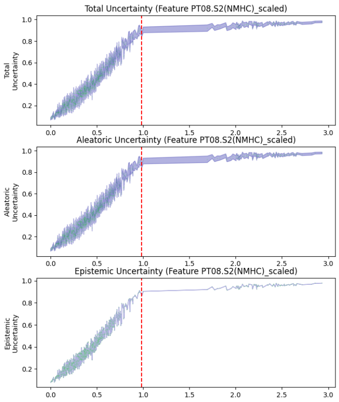
Dropconnect Model



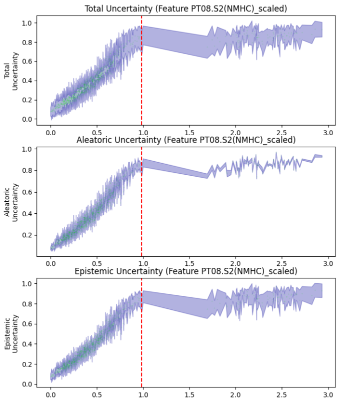
Flipout Model



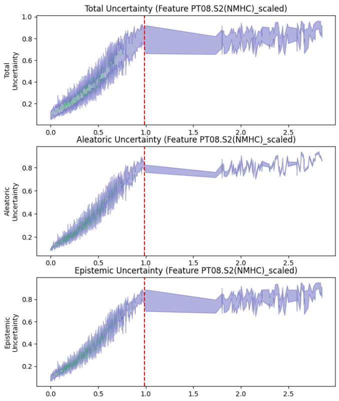
Ensemble Model



Classic Model

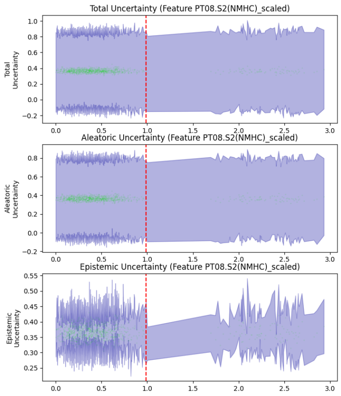


Dropout Model

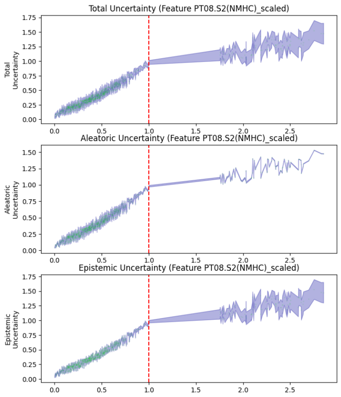


Dropconnect Model

Figure 4.2: Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.

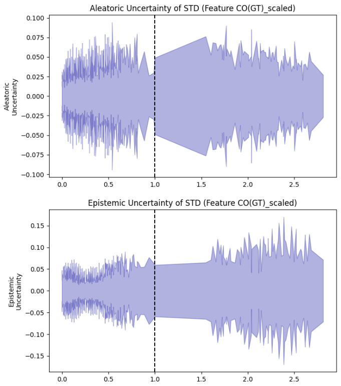


Flipout Model

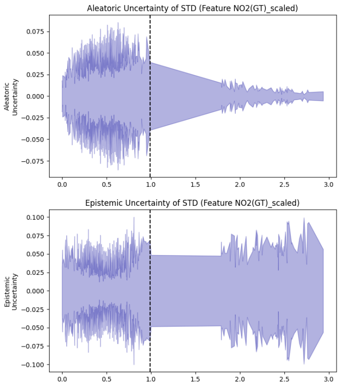


Ensemble Model

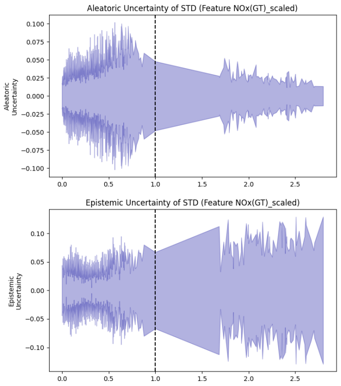
Figure 4.4: Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using β-NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD. The x-axis represents the domain



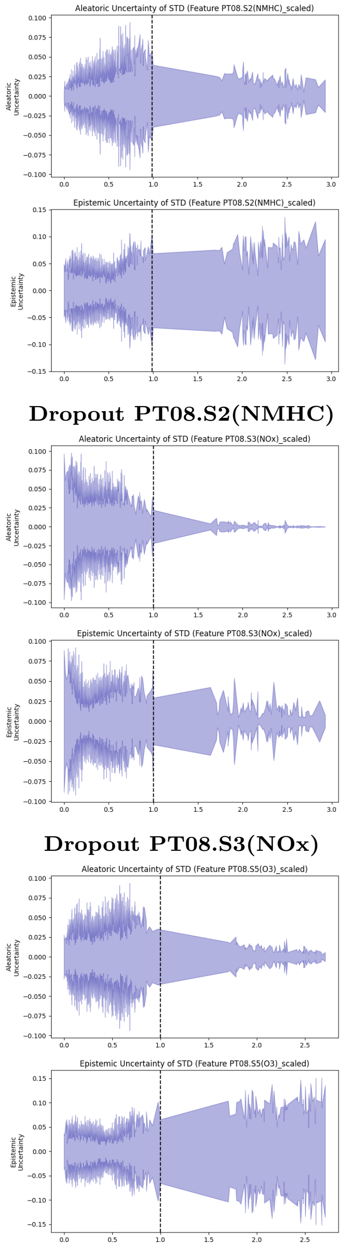
Dropout CO(GT)



Dropout NO2(GT)



Dropout NOx(GT)



Dropout PT08.S2(NMHC)

Dropout PT08.S3(NOx)

Dropout PT08.S5(O3)

Figure 4.6: Comparison of aleatoric and epistemic uncertainty on Dropout without the mean predictions using β-NLL loss function for 6 of 8 feature, all values to the left of the vertical black line are ID and the rest are OOD. The x-axis represents the domain.

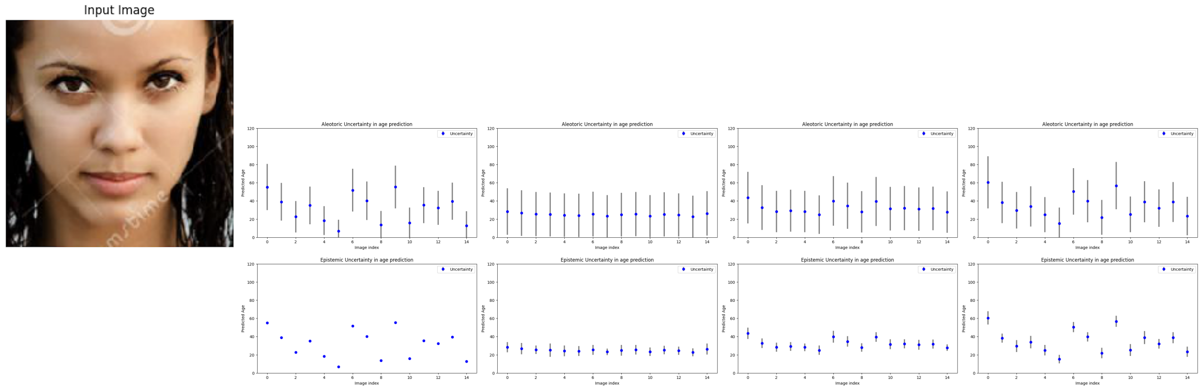
4.2.2 Between Different Features

Due to the possibility of the relationship of the out-of-distribution data to differ between each individual feature, through figures 4.6 we can visualize different types of variations of the same method. The figure presents the Dropout method without the predicted mean, as mentioned earlier, the method is reliable to find epistemic uncertainty. However, there are different types of relationships between ID and OOD, for example, all the individual features show an increase in epistemic uncertainty in the OOD area except for PT08.S3(NOx). Additionally, when looking at aleatoric uncertainty, CO(GT) shows that there is an increase instead of a decrease in uncertainty in the OOD. We expect all of the relationships in the epistemic and aleatory uncertainty of the same method to show the same pattern in each individual feature, however, clearly in certain cases there is a difference (Refer to Appendix C.4 for other Dropout graphs). These insights could prove that the OOD data may possess unique characteristics or patterns that differ from the training data, which is not inherent to the other OOD data.

Unlike Dropout, when visualizing Ensembles (graphs can be found in Appendix C.4), it has an evident difference in ID and OOD when it comes to all features, making it a better method for epistemic uncertainty. Therefore, when looking closer into each feature, we can visualize the uncertainties that may have a greater influence on the model.

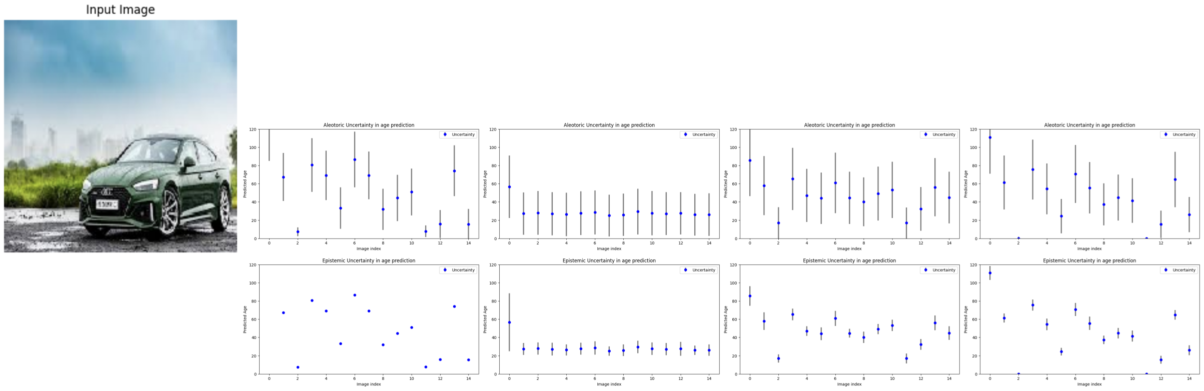
4.3 Face Age Regression Analysis

When comparing the method of prediction of the model, there seems to be a general pattern in it's mean prediction that is evident when comparing between the ID and OOD datasets. Through figure 4.7 & 4.8, as expected, the aleatoric uncertainty is extremely high in all models due the earlier explanation as the labels are very noisy due to the difficulty in finding age is inherently difficult even for humans. Moreover, the use of an ROC would not be valid in this case and is justified by the varying predictions between ID and OOD results. However, in the ID case, it seems that Dropout and Flipout have the closest predictions to the actual age (19). While the predictions made on the OOD are confident, it is clear that the range of predictions are bigger than



Classic Model **Dropout Model** **DropConnect Model** **Flipout Model**

Figure 4.7: Comparison of aleatoric and epistemic uncertainty on all uncertainty methods with the mean predictions using β-NLL loss function on ID data (Face images labelled 19). The error bars represent the corresponding uncertainty of the predictions. The x-axis represents the domain



Classic Model **Dropout Model** **DropConnect Model** **Flipout Model**

Figure 4.8: Comparison of aleatoric and epistemic uncertainty on all uncertainty methods with the mean predictions using β-NLL loss function on OOD data (Cars) . The error bars represent the corresponding uncertainty of the predictions.

the ID data. When comparing the uncertainty, as expected, the classic model in the ID results in no epistemic uncertainty while the OOD data causes the aleatoric uncertainty to increase. DropConnect shows very little to no variation in epistemic uncertainty and aleatoric uncertainty between the inputs in the ID. In the OOD, we see that aleatoric uncertainty is still consistent with the same range as the ID data, however, with a larger uncertainty. While it is clear that the epistemic has variation in it's uncertainty values that are larger than in the ID data. Similarly to the ID uncertainty in DropConnect, epistemic and aleatoric uncertainty are consistent. However, in the OOD condition, Dropout shows no change in aleatoric and epistemic uncertainty. In the OOD, the range of predictions in Flipout differ substantially compared to all the methods, the aleatoric uncertainty clearly increases while the epistemic uncertainty is constant in between the ID and OOD areas.

Based on the findings, classic and Flipout both show more sensitivity to the data as there is a clear indication of lower confidence in the range of the aleatoric uncertainty and predictions. However, the relationship of the uncertainty between ID and OOD is not sufficient to categorize them as an important disentangling method. Dropout produces the best predictions for the ID inputs, however, it clearly shows a lack of generalization in the OOD area. However, we can conclude that Dropconnect is the best in identifying epistemic uncertainty. The method shows consistency in the uncertainty in the aleatoric uncertainty yet has enough inputs in the OOD to show a variation in epistemic uncertainty compared to all other methods.

5 Conclusions & Discussion

In conclusion, we have been able to test and further research the applications of the various uncertainty methods of previous research on OOD data.

Based on the results of the air quality dataset, the model is overconfident in its predictions for the OOD values as it clearly is not substantial against the uncertainty in the ID data. However, the uncertainty methods prove to impact the model's epistemic and aleatoric uncertainty differently based on each individual features difference in creation of the OOD data. Given that the loss values provide information that the model is well trained to the ID data, the most evident difference in uncertainty between the two loss functions is in Dropout. We did not expect the method to have such a prominent difference on the epistemic uncertainty given that the model had already performed well on the ID data. In reference to Valdenegro-Toro & Mori (2022) results, once again Flipout proves to be an unreliable source to disentangle uncertainty in both types of datasets in the same manner, the application of Flipout seems to rely on aleatoric while identifying very low epistemic uncertainty. As expected ensembles provides a good indication of epistemic uncertainty in the OOD areas. On the other hand, although DropConnect and Dropout were seen to be heavily influenced by β-NLL loss like in previous studies, the impact of Dropout seems to improve aleatoric uncertainty estimation opposed to DropConnect. Despite this, based on this current research it may be difficult to conclude that DropConnect is a better general disentangling method when comparing it to similar datasets.

5.1 Future Research

An implementation of the uncertainty methods with the analysis of specific individuals on a scenario that follows a heteroscedastic pattern in the OOD inputs may provide a different indication to the best disentangling uncertainty method. As we attempt to mimic a real-world scenario dataset, the air quality dataset lacks the variation in feature relationships which results in the predictions to be significantly simple. Despite this, I argue that it has provided insightful results on the analysis of features with the addition of OOD data. On the other hand, the age recognition dataset may be more insightful with more determinant features that could distinguish different importance in features. For example, including certain face expressions may influence the models ability to distinguish uncertainty with the methods.

5.2 Conclusion

Hopefully these findings provide more insights and detail on different approaches where the uncertainty methods can be utilized in testing models on open-world scenarios. Although it may be difficult to create a realistic input of OOD data possible for a model, this research aims to provide insights on possible analysis that may be beneficial in the creation of synthetic data.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Ghavamzadeh, M., Fieguth, P., ... Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243-297.

Anonymous. (2020). Neural network out-of-distribution detection for regression tasks. *conference paper at ICLR 2020*, 1-14.

Bereitschaft, B. J. F. a. (2011). *Urban form and air quality in u.s. metropolitan and megapolitan areas (dissertation).* University of North Carolina at Greensboro.

Cui, P., & Wang, J. (2022). Out-of-distribution (ood) detection based on deep learning: A review. *Electronics*, *11(21)*, 1–19.

Depeweg, S., Hernandez-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. , 1-15.

Drummond, N., & Shearer, R. (2006). *The open world assumption.* (The University of Manchester)

ELKaraze, K., Raman, V., & Then, P. (2022). Facial age estimation using machine learning techniques: An overview. *Big Data Cogn*, *6(4)*, 128.

Gal, Y. (2016). *Uncertainty in deep learning* (Unpublished master's thesis). University of Cambridge.

Gal, Y., & Zoubin, G. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *University of Cambridge*, 1–12.

Gao, J., Bai, H., Zhu, L., & Ye, N. (2023). Rebenchmarking out-of-distribution detection in deep neural networks. , *12(145)*, 1–11.

Gao, Y., & Su, Q. (2021). Out-of-distribution detection with uncertainty enhanced attention maps. In *2021 international joint conference on neural networks (ijcnn)* (p. 1-8).

Gustafsson, F. K., Danelljan, M., & Schön, T. B. (2023). How reliable is your regression model's uncertainty under real-world distribution shifts? , 1-29.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *Microsoft Research*, 1-12.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*, *110*, 457–506.

Iqbal, U., Barthelemy, J., Perez, P., & Li, W. (2021). Regression on deep visual features using artificial neural networks (anns) to predict hydraulic blockage at culverts. , 1-11.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? , 1-12.

Kumar, K. (2022). *Car images dataset.* Kaggle. Retrieved from `www.kaggle.com/datasets/kshitij192/cars-image-dataset`

Learning, U. M. (2020). *Air quality dataset.* Kaggle. Retrieved from `www.kaggle.com/datasets/fedesoriano/air-quality-data-set`

Lee, J., Park, S., & Lee, J. (2023). Estimation of uncertainty for technology evaluation factors via bayesian neural networks. , *12(145)*, 1–18.

Meghana, A. S., Sudhakar, S., Arumugam, G., Srinivasan, P., & Prakash, K. B. (2020). Age and gender prediction using convolution, resnet50 and inception resnetv2. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 1328-1334.

Mobiny, A., Nguyen, H. V., Moulik, S., Garg, N., & Wu, C. C. (2021). Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific Reports*, *11*, 1–12.

Rabbi, F. (2018). *Facial age.* Kaggle. Retrieved from www.kaggle.com/datasets/frabbisw/facial-age

Seitzer, M., Tavakoli1, A., Antic, D., & Martius1, G. (2022). On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *Conference paper at ICLR 2022*, 1-24.

Siegel, A. F. (2012). Chapter 9 - confidence intervals: Admitting that estimates are not exact. In A. F. Siegel (Ed.), *Practical business statistics (sixth edition)* (p. 219-247). Academic Press.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*, 44–56.

Valdenegro, M. (2023). *Evaluation of uncertainty quantification.* (Lecture slides from European Summer School in AI, University of Groningen.)

Valdenegro-Toro, M., & Mori, D. S. (2022). A deeper look into aleatoric and epistemic uncertainty disentanglement. , 1-12.

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2022). Generalized out-of-distribution detection: A survey. , 1–22.

A Architecture and Optimization of Models

A.1 Configurations for Air Quality Models

The stochastic function replaces includes the final layer, Dense(1), in order to produce the uncertainties

- Number of epochs: 100.
- Batch size: 10.
- Optimizer: Adam (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$).
- **Classic**:
  - Dense(128, ReLU)
  - Dense(128, ReLU)
  - Dense(64, ReLU).
- **Dropout**:
  - Dense(128, ReLU)
  - Dropout(0.3)
  - Dense(128, ReLU)
  - Dropout(0.3)
  - Dense(64, ReLU)
  - Dropout(0.3).
- **DropConnect**:
  - DropConnectDense(128, ReLU, p = 0.5)
  - DropConnectDense(128, ReLU, p = 0.5)
  - DropConnectDense(64, ReLU, p = 0.5).
- **Flipout**:
  - FlipoutDense(128, ReLU)
  - FlipoutDense(128, ReLU)
  - FlipoutDense(64, ReLU)
  - (Prior is disabled)
- **Ensembles**:
  - 5 copies of the classic model trained with different random weight initializations.

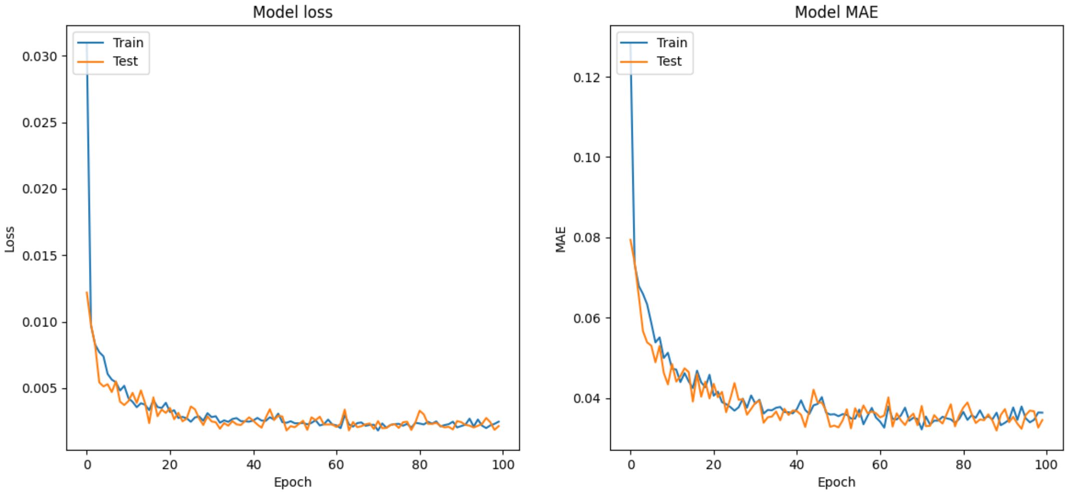
A.2 Configurations for Age Regression Models

The stochastic function replaces includes the final layer, Dense(1), in order to produce the uncertainties

- Number of epochs: 150.
- Batch size: 32.
- Optimizer: Adam (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$).
- **Resnet 50 Configuration**
  - pooling='avg'
  - weights='imagenet'
- **Classic:**
  - Resenet50
  - Dense(128, ReLU)
  - Dense(64, ReLU)
  - Dense(64, ReLU)
  - Dense(32, ReLU)
- **Dropout:**
  - Resenet50
  - Dense(128, ReLU)
  - Dropout(0.5)
  - Dense(64, ReLU)
  - Dropout(0.5)
  - Dense(64, ReLU)
  - Dropout(0.5).
  - Dense(32, ReLU)
  - Dropout(0.5).
- **DropConnect:**
  - Resenet50
  - DropConnectDense(128, ReLU, p = 0.5)
  - DropConnectDense(64, ReLU, p = 0.5)
  - DropConnectDense(64, ReLU, p = 0.5)
  - DropConnectDense(32, ReLU, p = 0.5).
- **Flipout:**
  - Resenet50
  - FlipoutDense(128, ReLU)
  - FlipoutDense(64, ReLU)
  - FlipoutDense(64, ReLU)
  - FlipoutDense(32, ReLU)
  - (Prior is disabled)

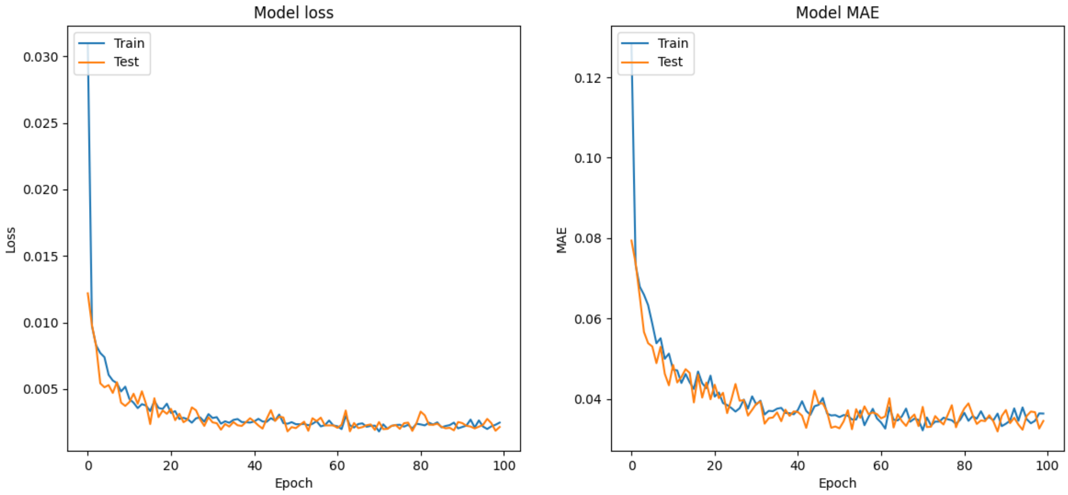
A.3 Loss Curves of Air Quality Models

A.3.1 Classic Model Loss Curve



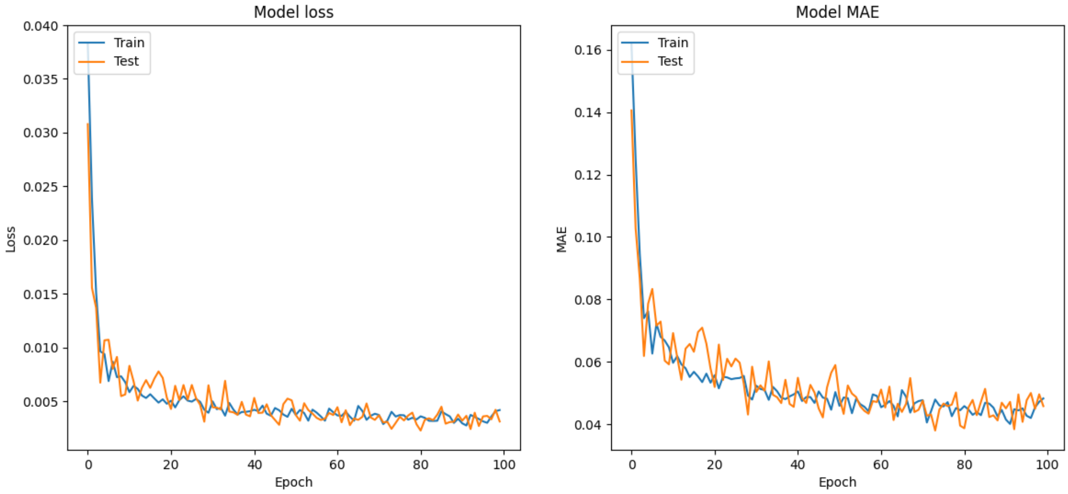
Loss : [MSE, MAE]
Train loss: [0.0006776957307010889, 0.018691718578338623]
Test loss: [0.0007112125167623162, 0.020568329840898514]
Train R2: 0.9776055570634924
Test R2: 0.9780257693869515

A.3.2 Dropout Model Loss Curve



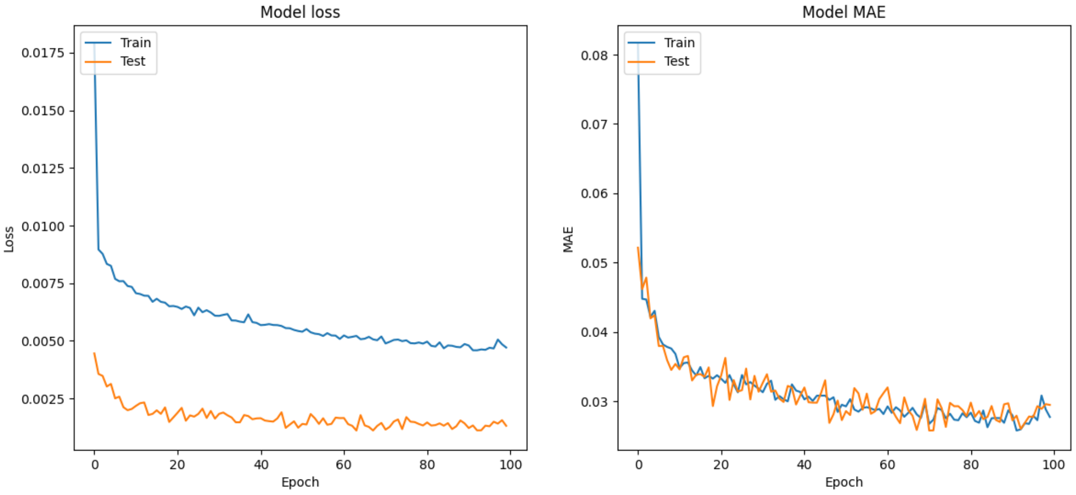
Loss : [MSE, MAE]
Train loss: [0.0021760028321295977, 0.034687817096710205]
Test loss: [0.0018724644323810935, 0.03415553644299507]
Train R2: 0.9313477120821618
Test R2: 0.9313402800606485

A.3.3 Dropconnect Model Loss Curve



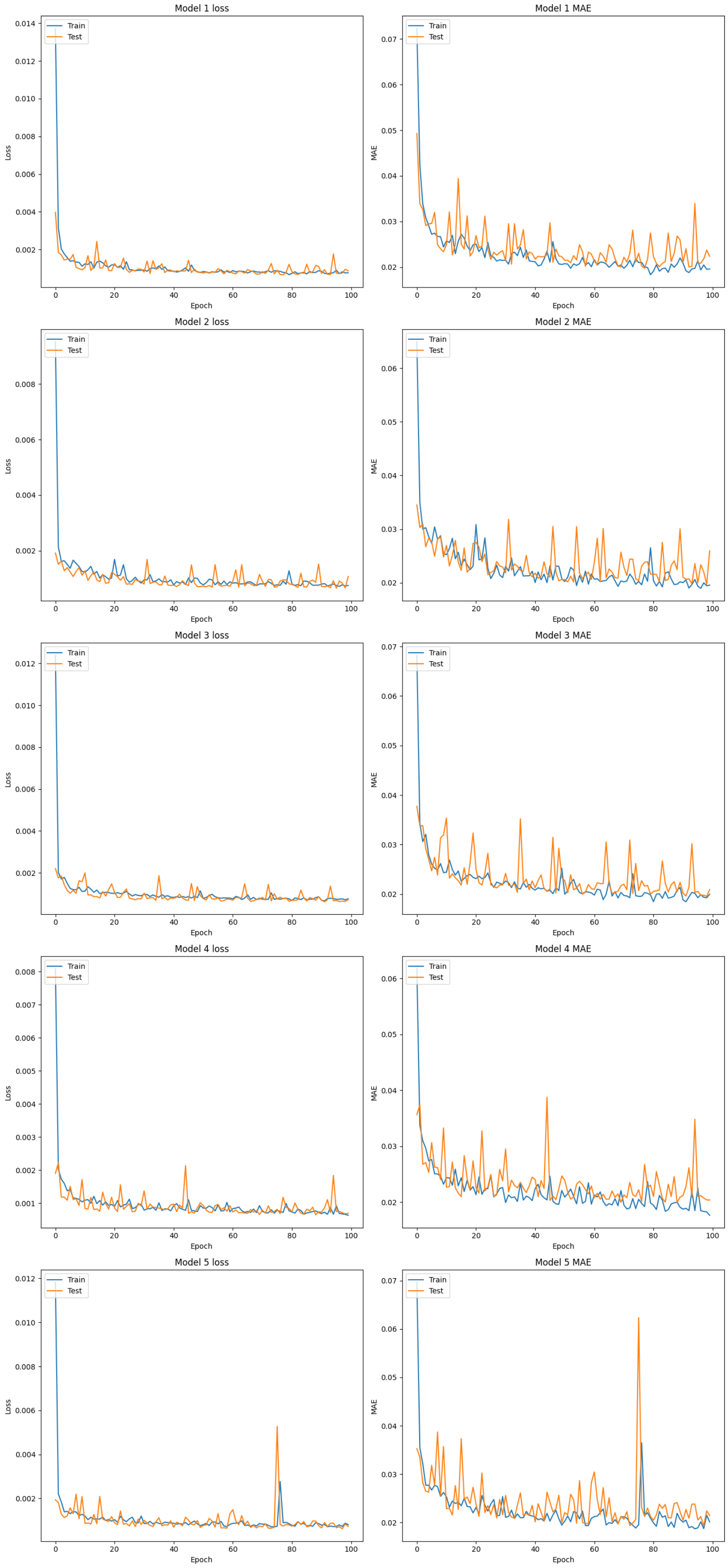
Loss : [MSE, MAE]
Train loss: [0.003306058468297124, 0.04279094934463501]
Test loss: [0.00324618024751544, 0.04616352915763855]
Train R2: 0.9244076023494844
Test R2: 0.9414050531525713

A.3.4 Flipout Model Loss Curve



Loss : [MSE, MAE]
Train loss: [0.001307631959207356, 0.02685621567070484]
Test loss: [0.001243737991899252, 0.027157479897141457]
Train R2: 0.9565270387391661
Test R2: 0.962132440518198

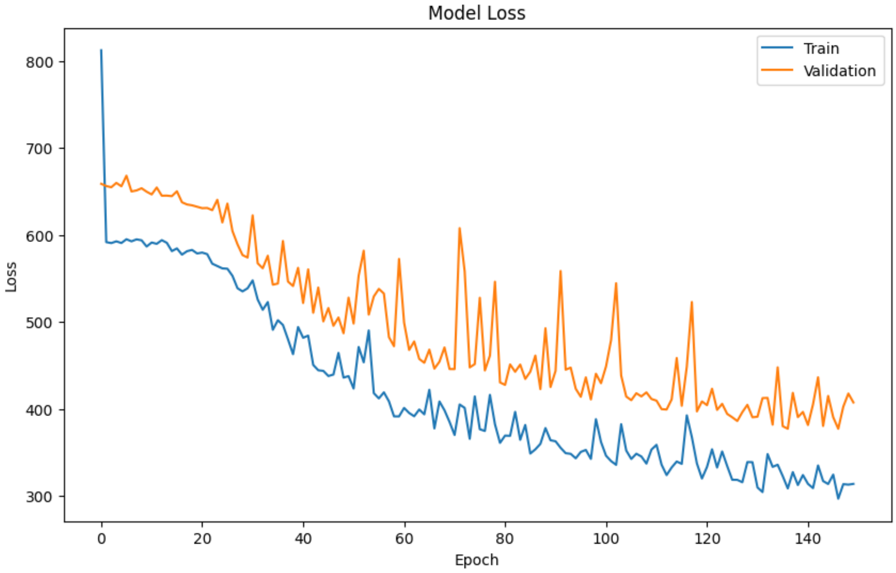
A.3.5 Ensemble Model Loss Curve



Loss : [MSE, MAE]
Train loss: [0.0006230067199714686 0.017634724918639474]
Test loss: [0.0006879925336423829 0.02061510332010165]
Ensemble R^2: [0.9794127516445735 0.9787431869505839]

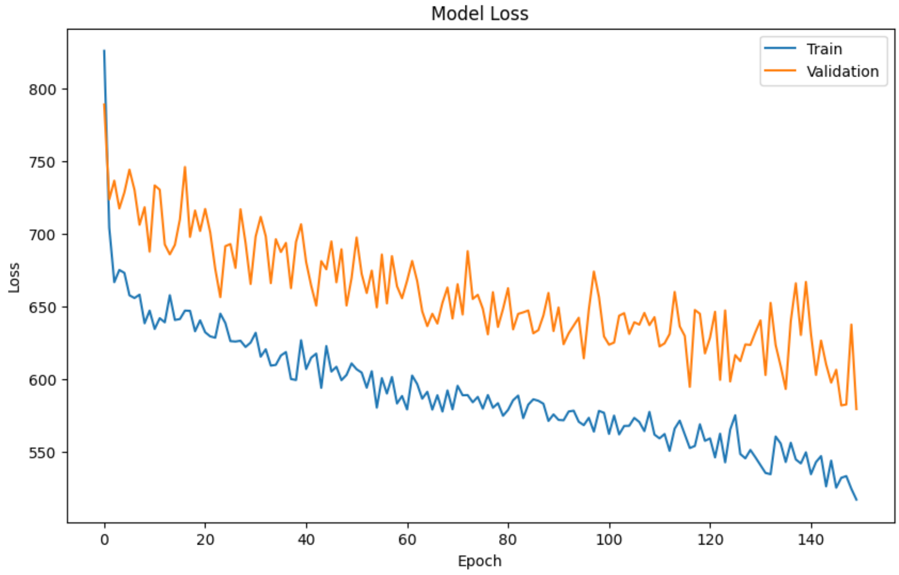
A.4 Loss Curves of Age Regressoin Models

A.4.1 Classic Model Loss Curve



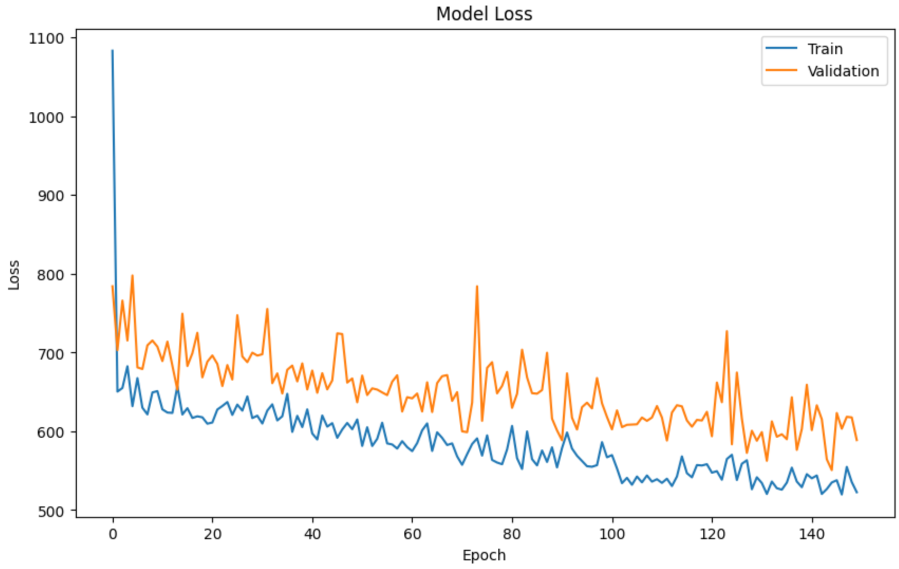
Loss : [MSE, MAE]
Train loss: [327.8652038574219, 14.865017890930176]
Test loss: [407.4013977050781, 16.871294021606445]
Train R2: 0.4462479753478099
Test R2: 0.38005652315256466

A.4.2 Dropout Model Loss Curve



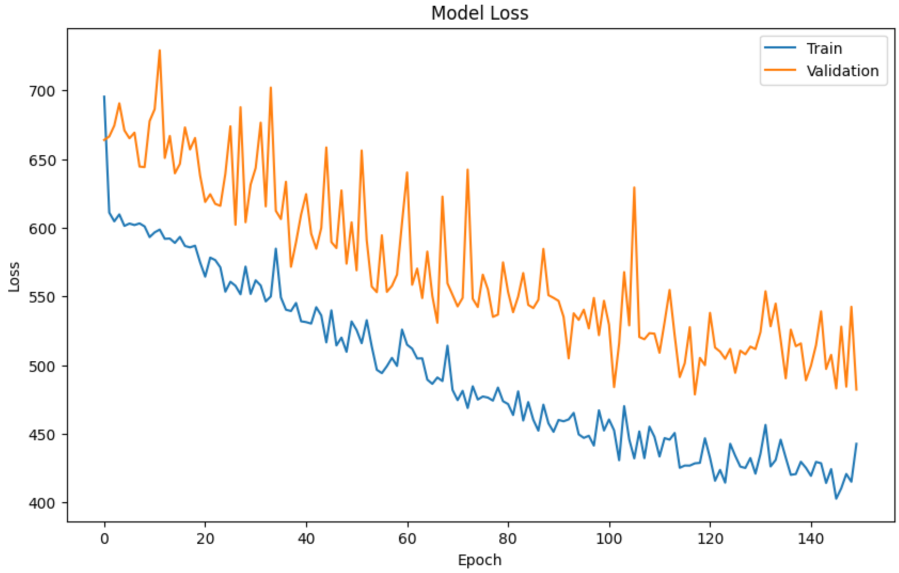
Loss : [MSE, MAE]
Train loss: [533.34521484375, 19.018558502197266]
Test loss: [599.1786499023438, 20.541803359985 35]
Train R2: 0.12777903676563573
Test R2: 0.1010953880354678

A.4.3 Dropconnect Model Loss Curve



Loss : [MSE, MAE]
Train loss: [548.69482421875, 18.952190399169922]
Test loss: [585.801513671875, 20.485567092895508]
Train R2: 0.12008096685632286
Test R2: 0.059550019013146493

A.4.4 Flipout Model Loss Curve



Loss : [MSE, MAE]
Train loss: [430.3692321777344, 17.077608108520508]
Test loss: [496.2757873535156, 18.633813858032227]
Train R2: 0.28674907615052336
Test R2: 0.2836386445386654

B Receiver Operating Curve Information

B.1 AUC values of Uncertainty Methods and Loss functions

Table B.1: AUC values from the ROC curve using each uncertainty method, NLL as the loss function and using the predictions from Aleatoric uncertainty

Feature	Classic	Dropout	Dropconnect	Flipout	Ensembles
PT08.S1(CO)	0.14	0.10	0.05	1.00	0.03
NOx(GT)	0.02	0.73	0.35	0.98	0.00
PT08.S3(NOx)	0.00	0.00	0.00	0.25	0.00
CO(GT)	0.81	0.80	0.69	1.00	0.92
PT08.S2(NHMC)	0.00	0.14	0.14	1.00	0.00
NO2(GT)	0.00	0.16	0.00	0.95	0.00
PT08.S5(O3)	0.00	0.88	0.77	0.99	0.00
NMHC(GT)	0.10	0.69	0.24	0.99	0.01

Table B.2: AUC values from the ROC curve using each uncertainty method, NLL as the loss function and using the predictions from Epistemic uncertainty

Feature	Classic	Dropout	Dropconnect	Flipout	Ensembles
PT08.S1(CO)	0.59	0.85	0.88	0.51	1.00
NOx(GT)	0.64	0.97	0.90	0.45	1.00
PT08.S3(NOx)	0.28	0.41	0.23	0.83	1.00
CO(GT)	0.59	0.91	0.74	0.36	1.00
PT08.S2(NHMC)	0.67	0.68	0.84	0.46	1.00
NO2(GT)	0.56	0.90	0.97	0.41	1.00
PT08.S5(O3)	0.57	0.92	0.77	0.43	1.00
NMHC(GT)	0.45	0.94	0.85	0.34	1.00

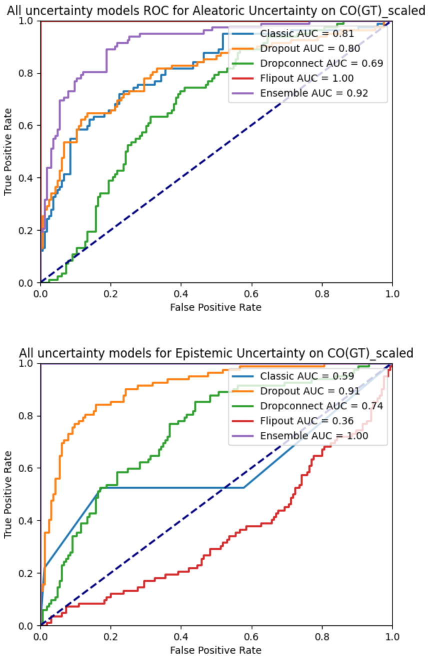
Table B.3: AUC values from the ROC curve using each uncertainty method, β-NLL as the loss function and and using the predictions Aleatoric uncertainty

Feature	Classic	Dropout	Dropconnect	Flipout	Ensembles
PT08.S1(CO)	0.00	0.15	0.08	0.30	0.00
NOx(GT)	0.01	0.13	0.12	0.74	0.00
PT08.S3(NOx)	0.00	0.00	0.00	0.43	0.00
CO(GT)	0.96	0.71	0.18	0.42	0.03
PT08.S2(NHMC)	0.01	0.05	0.17	0.31	0.00
NO2(GT)	0.00	0.00	0.02	0.73	0.00
PT08.S5(O3)	0.00	0.01	0.13	0.23	0.00
NMHC(GT)	0.12	0.42	0.38	0.42	0.11

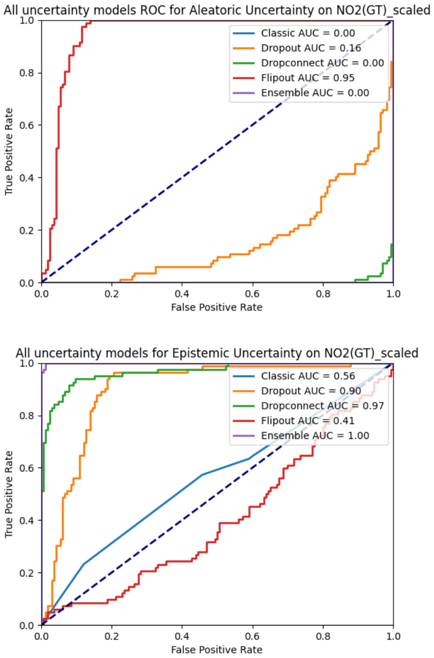
Table B.4: AUC values from the ROC curve using each uncertainty method, β-NLL as the loss function and and using the predictions Epistemic uncertainty

Feature	Classic	Dropout	Dropconnect	Flipout	Ensembles
PT08.S1(CO)	0.61	0.84	0.79	0.40	1.00
NOx(GT)	0.65	0.97	0.79	0.81	1.00
PT08.S3(NOx)	0.34	0.02	0.17	0.41	1.00
CO(GT)	0.64	0.91	0.73	0.67	1.00
PT08.S2(NHMC)	0.57	0.91	0.85	0.69	1.00
NO2(GT)	0.41	0.92	0.89	0.75	1.00
PT08.S5(O3)	0.63	0.93	0.82	0.66	1.00
NMHC(GT)	0.54	0.94	0.83	0.74	1.00

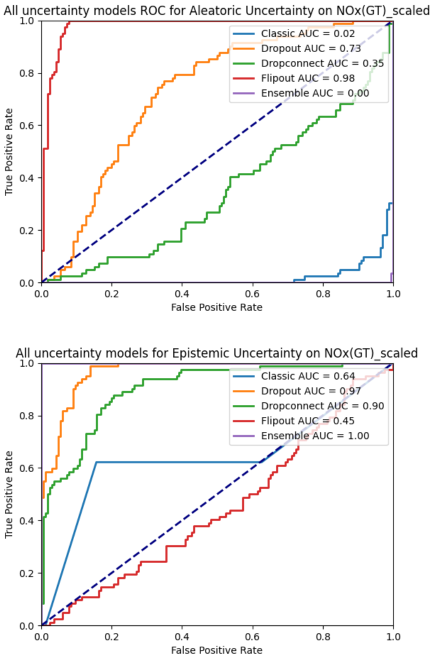
B.2 Sample Graphs of ROC using NLL



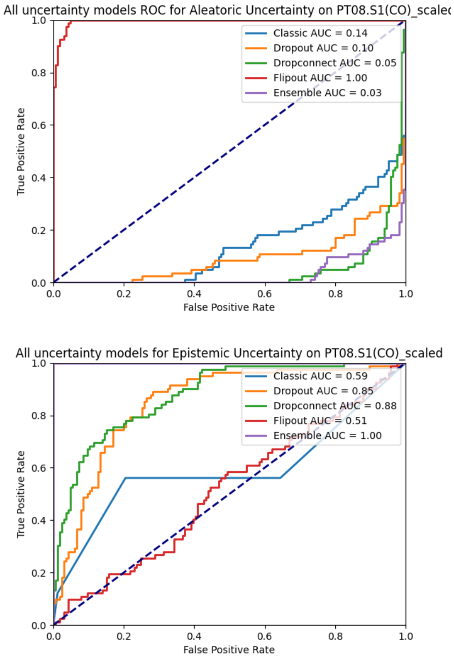
ROC curve of CO(GT)



ROC curve of NO2(GT)

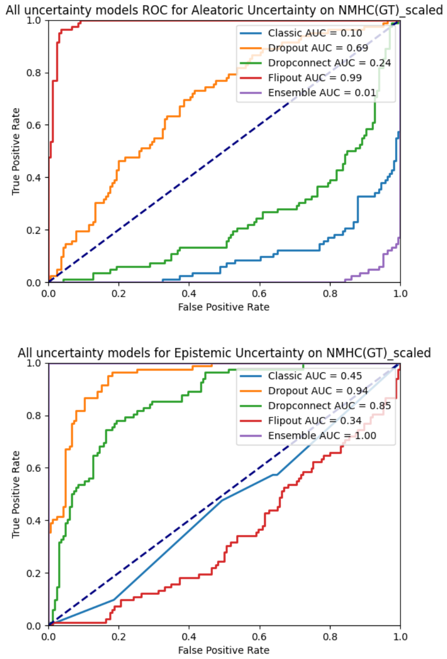


ROC curve NOx(GT)


All uncertainty models ROC for Aleatoric Uncertainty on PT08.S1(CO)_scalec

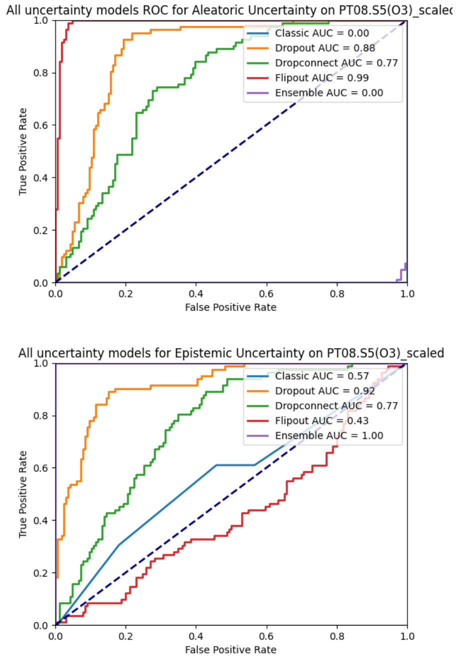
All uncertainty models for Epistemic Uncertainty on PT08.S1(CO)_scaled

ROC curve of PT08.S1(CO)

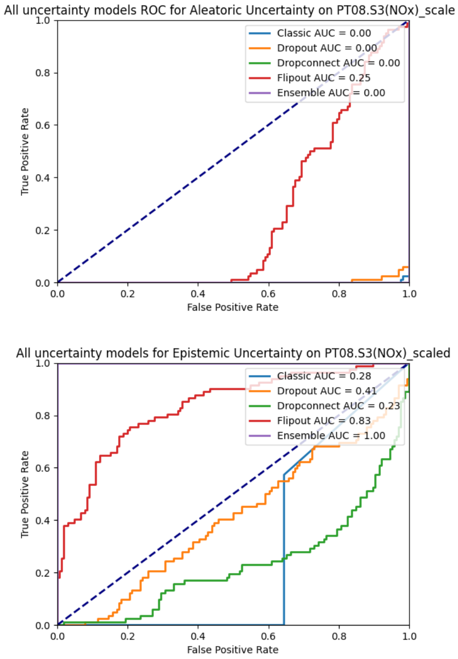

All uncertainty models ROC for Aleatoric Uncertainty on NMHC(GT)_scaled

All uncertainty models for Epistemic Uncertainty on NMHC(GT)_scaled

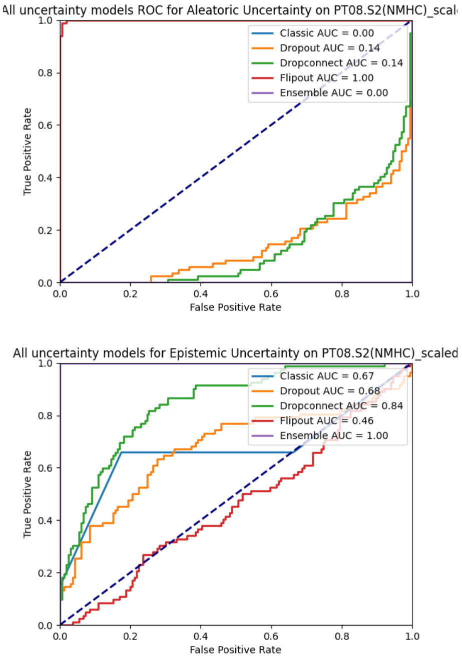
ROC curve of NMHC(GT)



ROC curve of PT08.S5(O3)



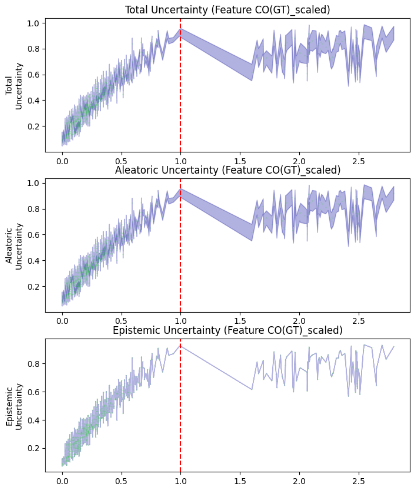
ROC curve of PT08.S3(NOx)



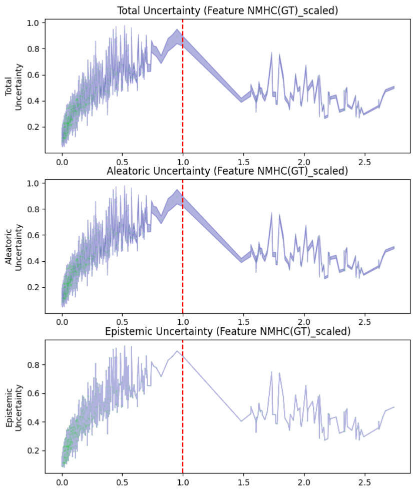
ROC curve of PT08.S2(NMHC)

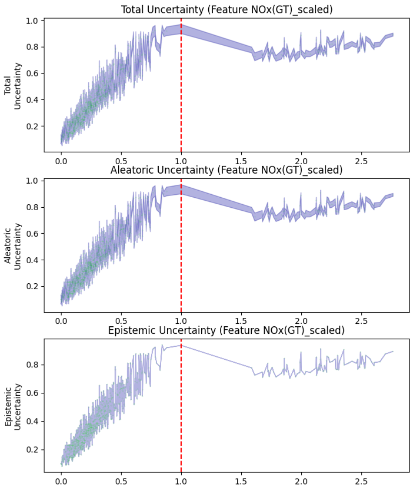
C Visualizatoin of Aleatoric and Epistemic Uncertainty

C.1 Visualization of uncertainties with mean and standard deviation, NLL loss function

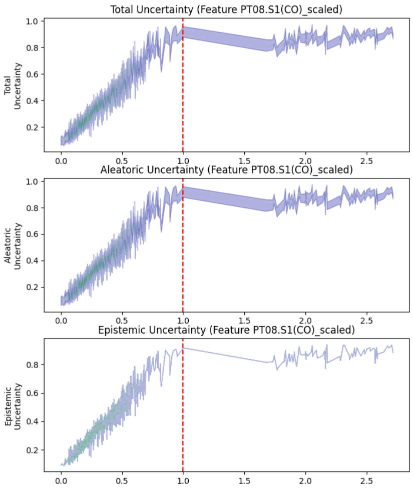


Classic CO(GT)

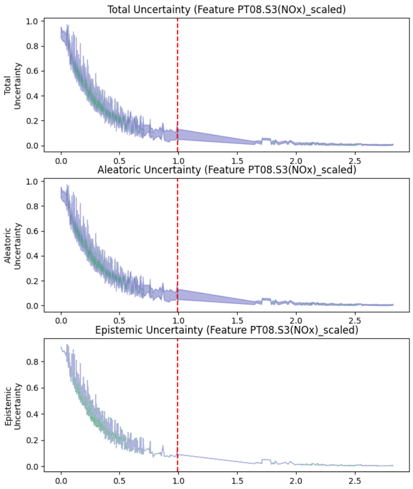




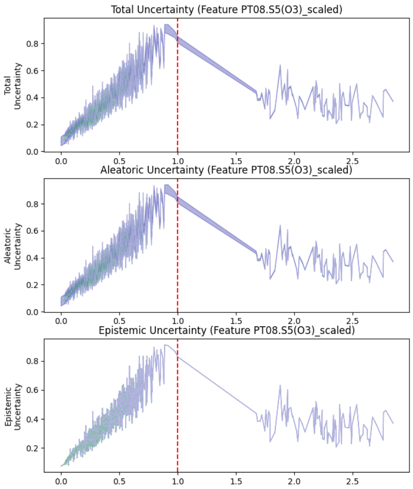
Classic NOx(GT)



Classic PT08.S1(CO)

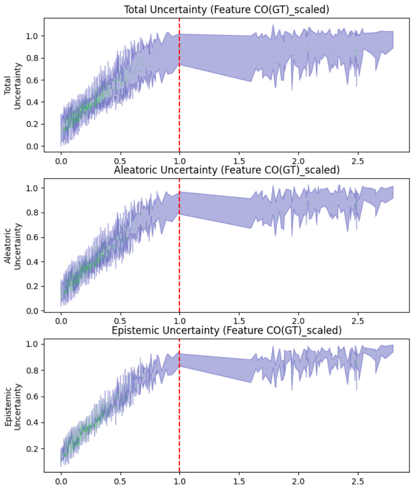


Classic PT08.S3(NOx)

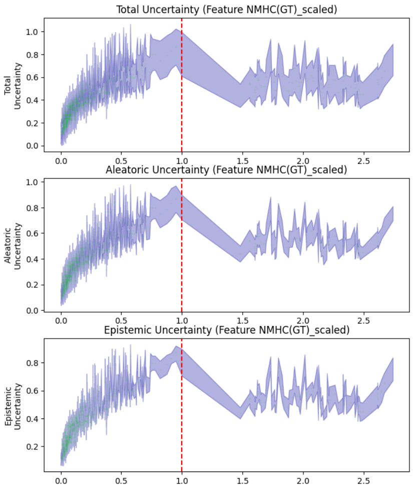


Classic PT08.S5(O3)

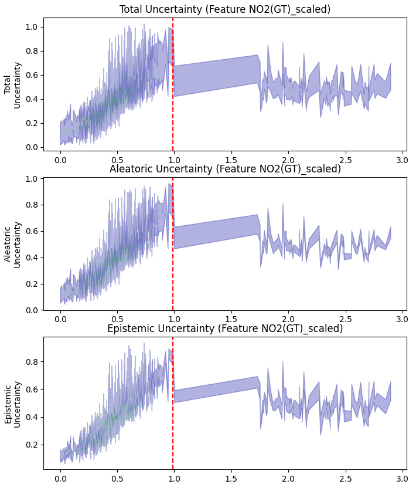
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



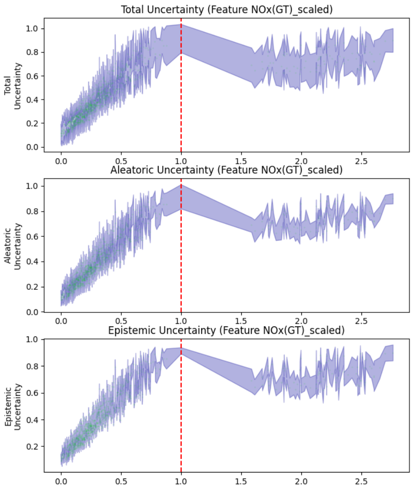
Dropout CO(GT)



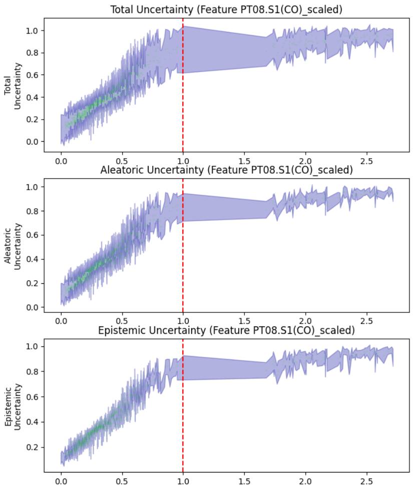
Dropout NMHC(GT)



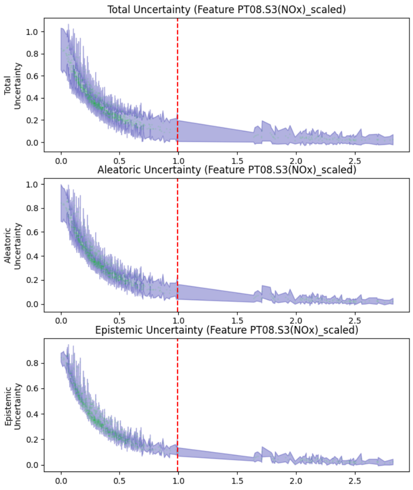
Dropout NO2(GT)



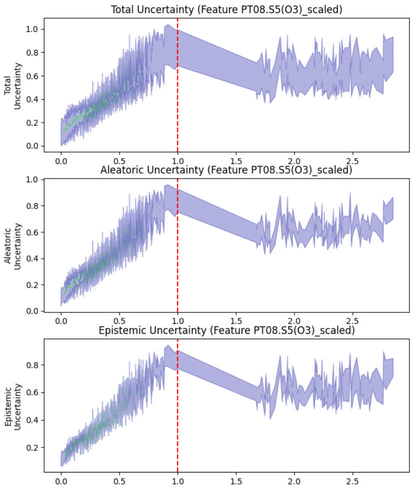
Dropout NOx(GT)



Dropout PT08.S1(CO)

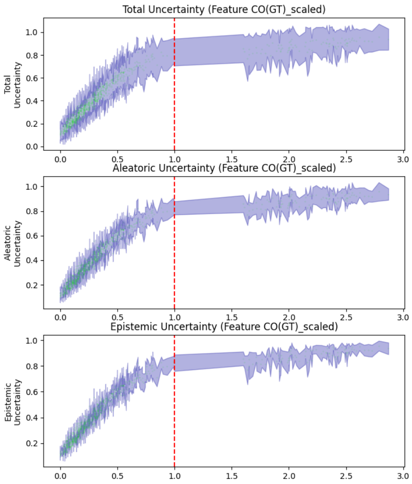


Dropout PT08.S3(NOx)

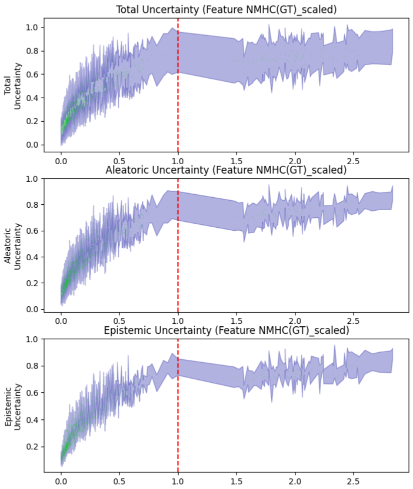


Dropout PT08.S5(O3)

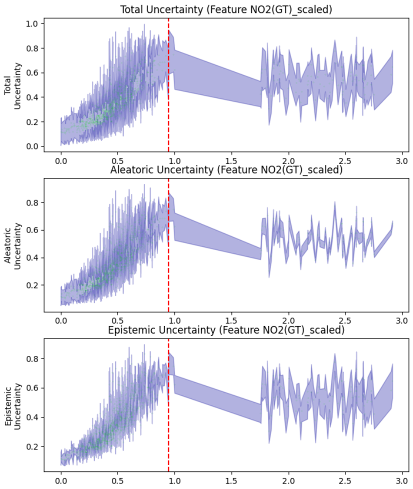
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



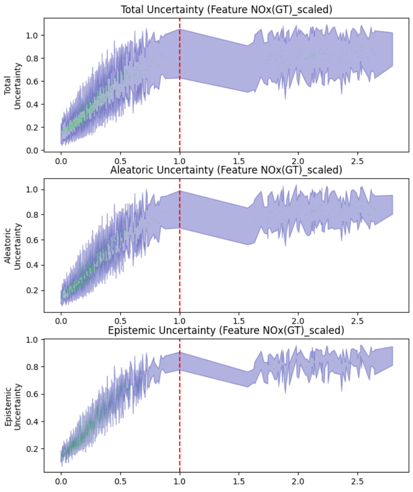
Dropconnect CO(GT)



Dropconnect NMHC(GT)



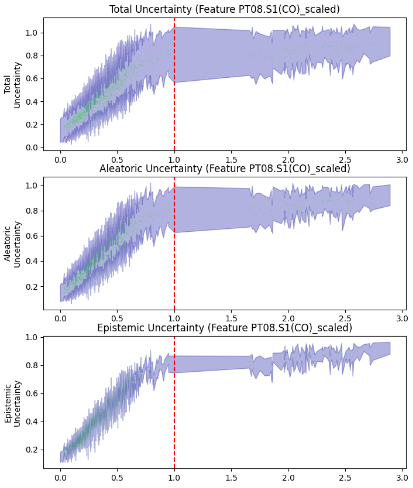
Dropconnect NO2(GT)


Total Uncertainty (Feature NOx(GT)_scaled)

Aleatoric Uncertainty (Feature NOx(GT)_scaled)

Epistemic Uncertainty (Feature NOx(GT)_scaled)

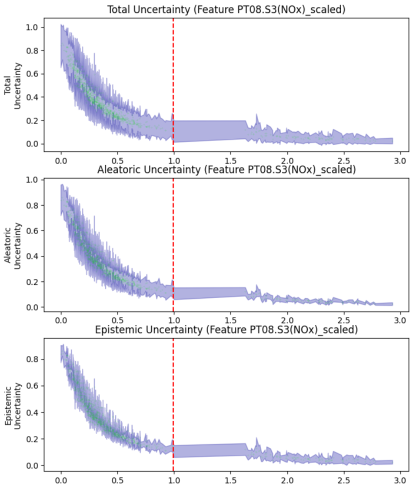
Dropconnect NOx(GT)


Total Uncertainty (Feature PT08.S1(CO)_scaled)

Aleatoric Uncertainty (Feature PT08.S1(CO)_scaled)

Epistemic Uncertainty (Feature PT08.S1(CO)_scaled)

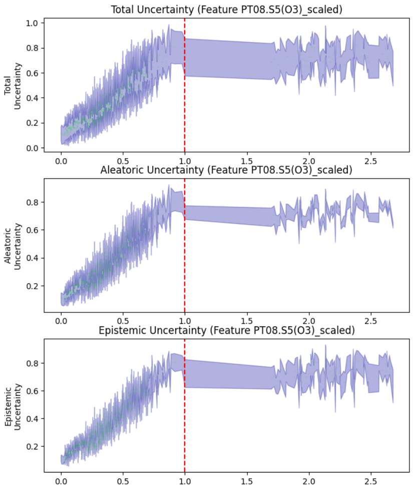
Dropconnect PT08.S1(CO)


Total Uncertainty (Feature PT08.S3(NOx)_scaled)

Aleatoric Uncertainty (Feature PT08.S3(NOx)_scaled)

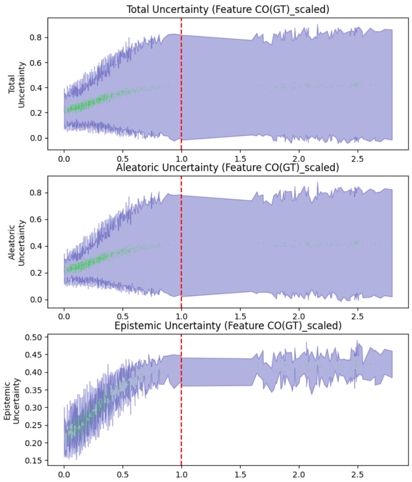
Epistemic Uncertainty (Feature PT08.S3(NOx)_scaled)

Dropconnect PT08.S3(NOx)

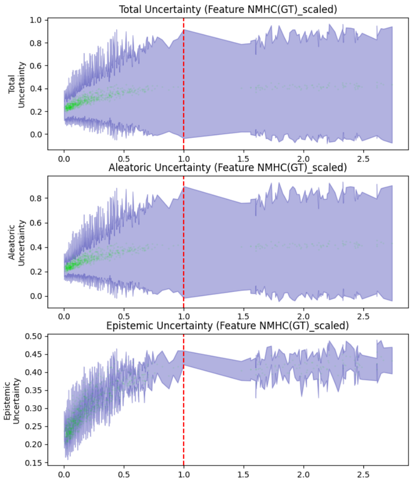


Dropconnect PT08.S5(O3)

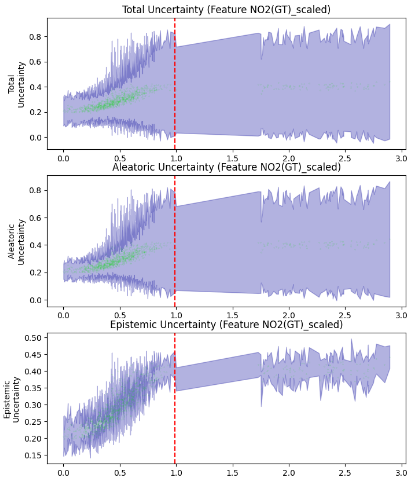
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



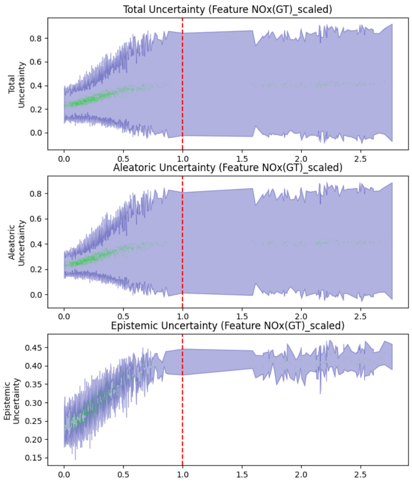
Flipout CO(GT)



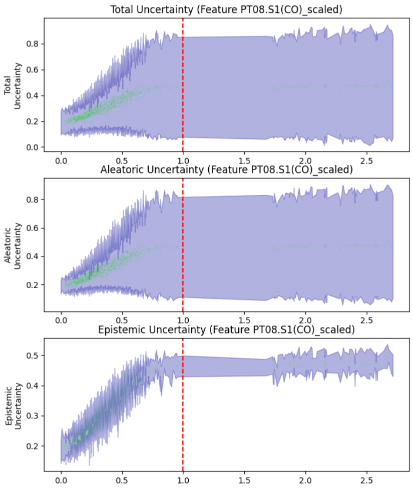
Flipout NMHC(GT)



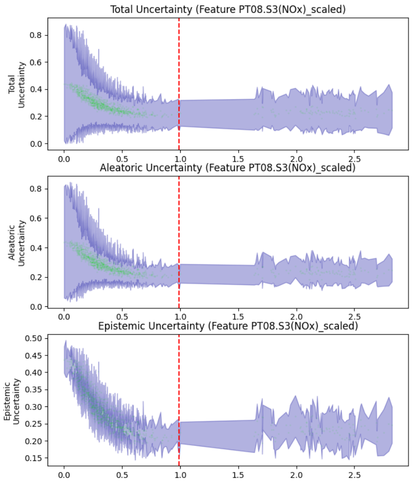
Flipout NO2(GT)



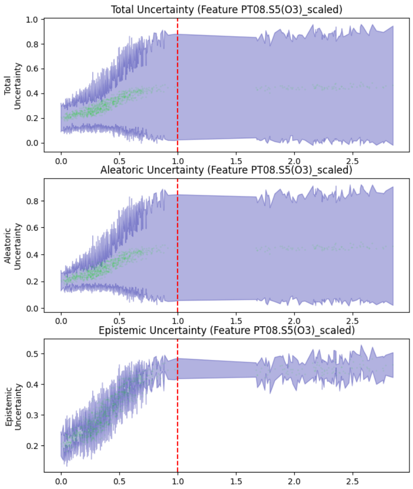
Flipout NOx(GT)



Flipout PT08.S1(CO)

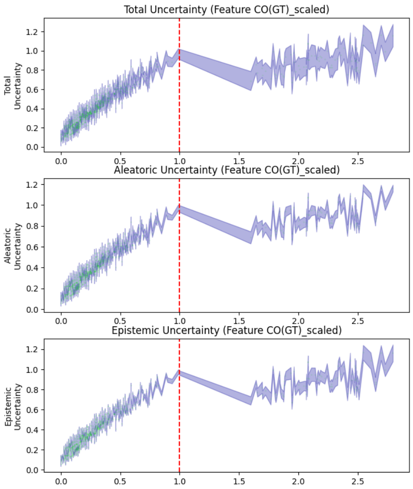


Flipout PT08.S3(NOx)

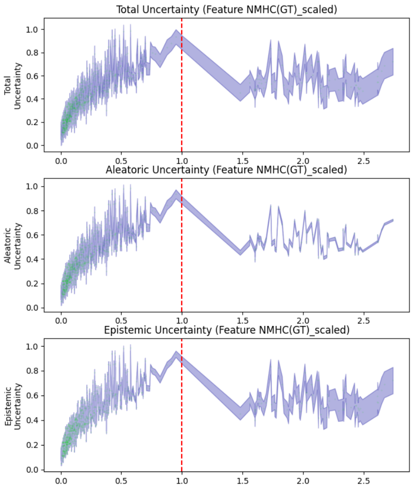


Flipout PT08.S5(O3)

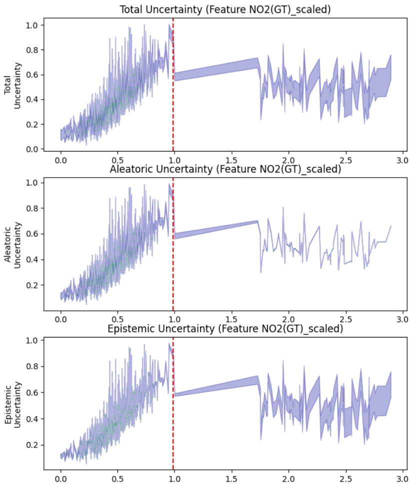
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



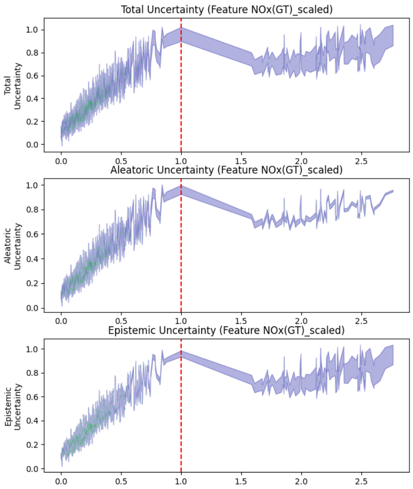
Ensemble CO(GT)



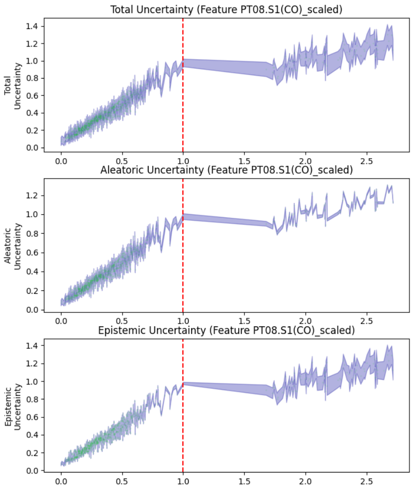
Ensemble NMHC(GT)



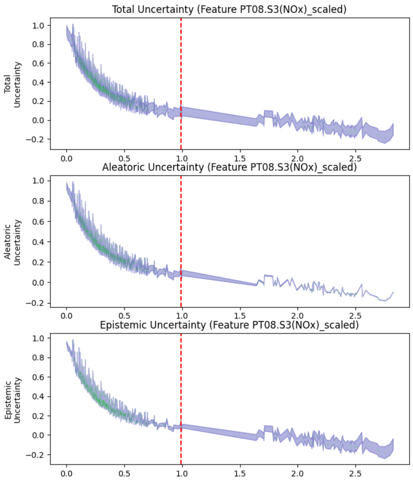
Ensemble NO2(GT)



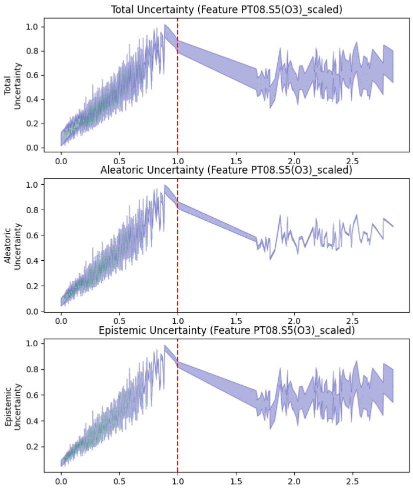
Ensemble NOx(GT)



Ensemble PT08.S1(CO)



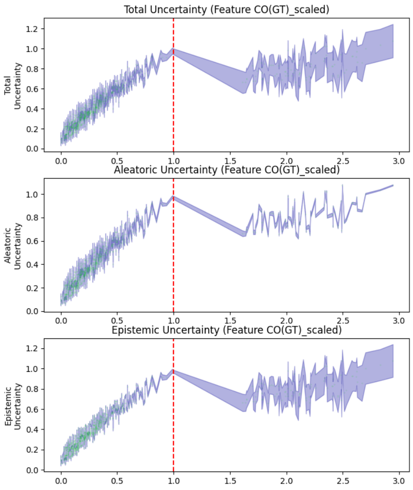
Ensemble PT08.S3(NOx)



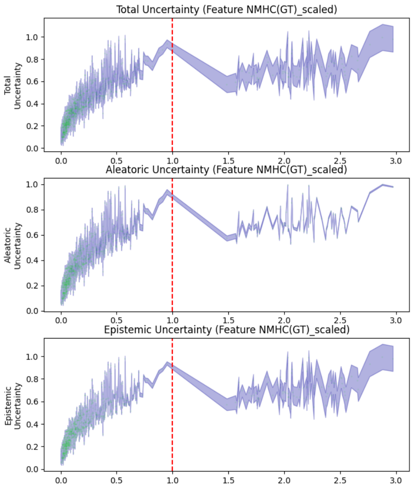
Ensemble PT08.S5(O3)

Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.

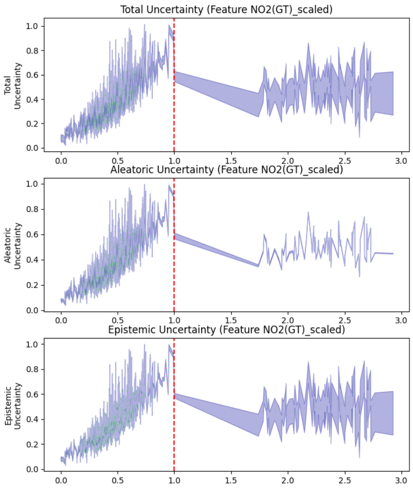
C.2 Visualization of uncertainties with mean and standard deviation, β-NLL loss function



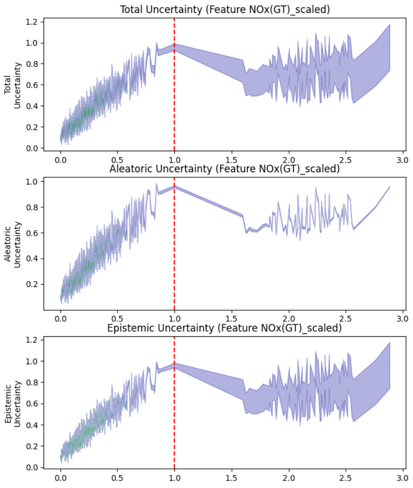
Ensemble CO(GT)



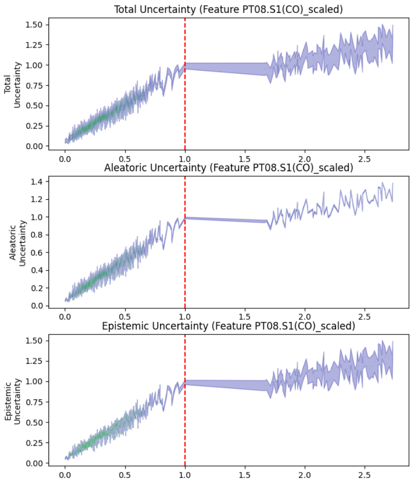
Ensemble NMHC(GT)



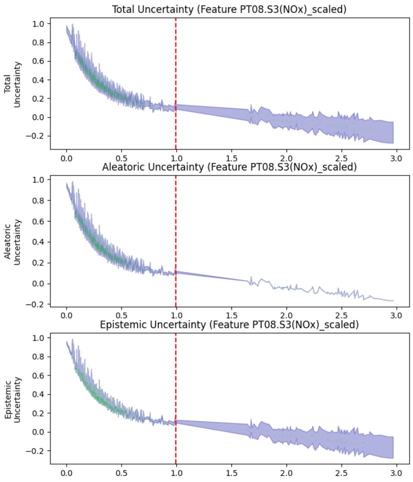
Ensemble NO2(GT)



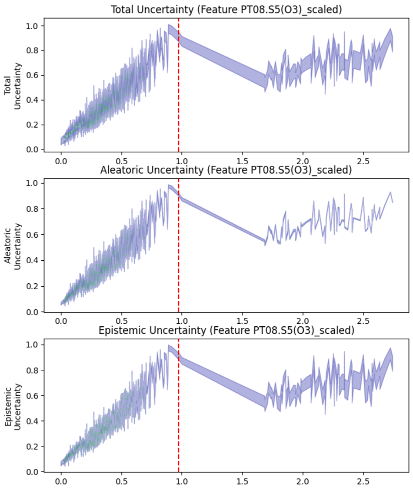
Ensemble NOx(GT)



Ensemble PT08.S1(CO)

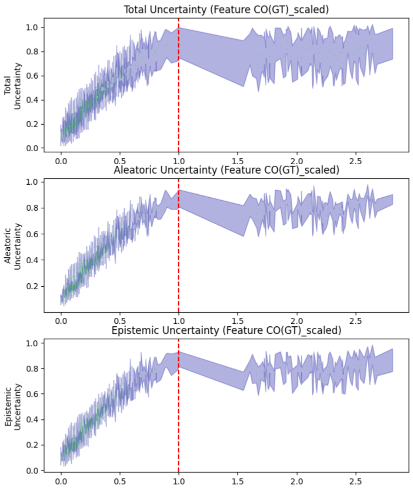


Ensemble PT08.S3(NOx)

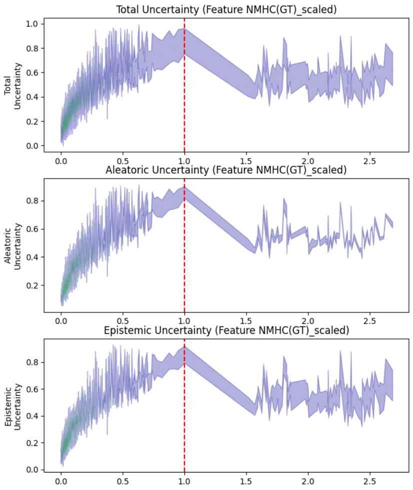


Ensemble PT08.S5(O3)

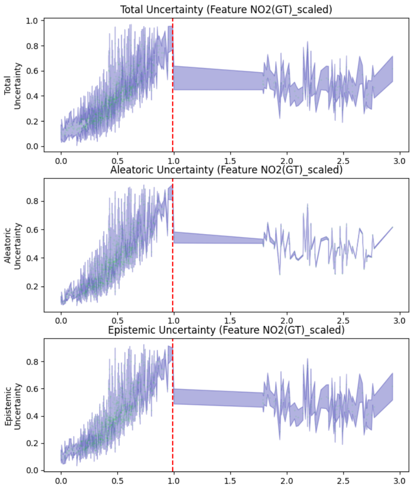
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using β-NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



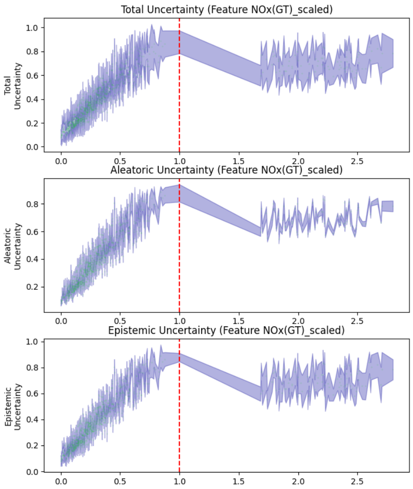
Dropout CO(GT)



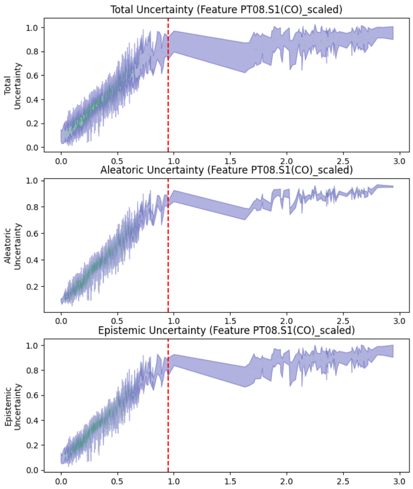
Dropout NMHC(GT)



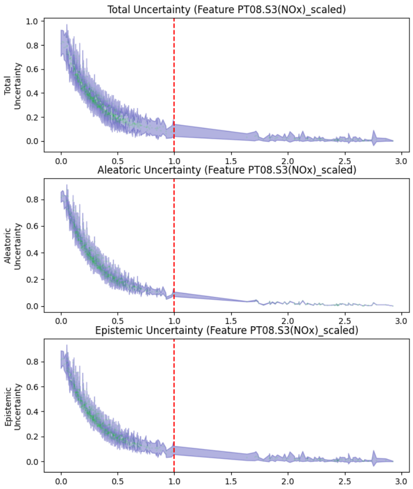
Dropout NO2(GT)



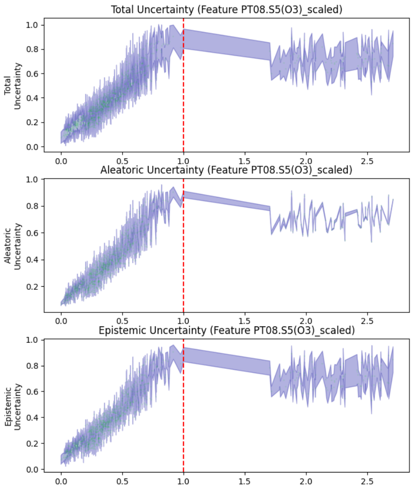
Dropout NOx(GT)



Dropout PT08.S1(CO)

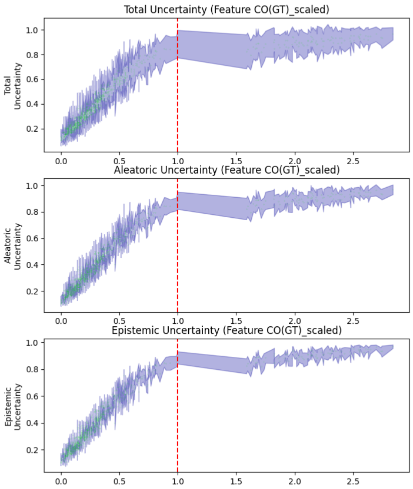


Dropout PT08.S3(NOx)

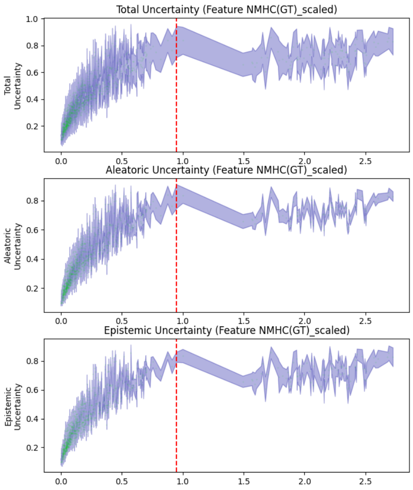


Dropout PT08.S5(O3)

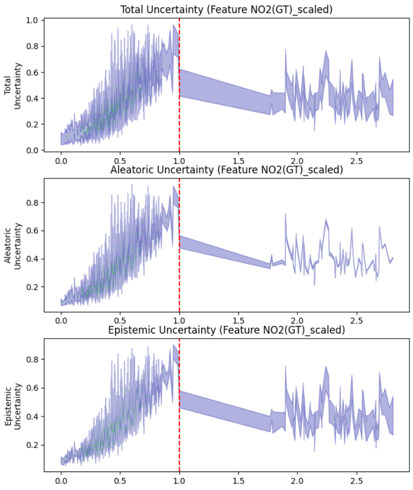
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using β-NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



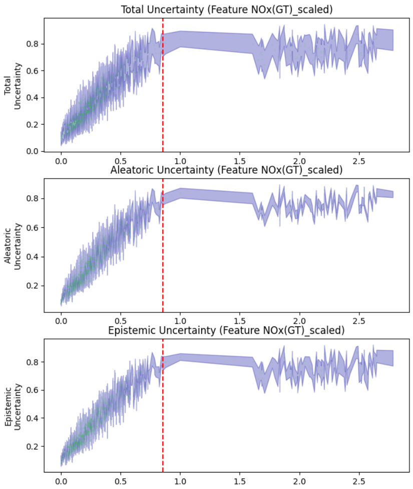
Dropconnect CO(GT)



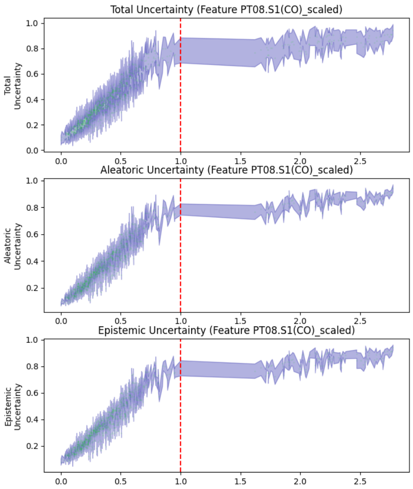
Dropconnect NMHC(GT)



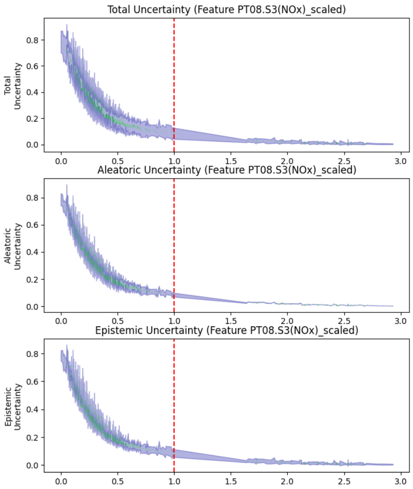
Dropconnect NO2(GT)



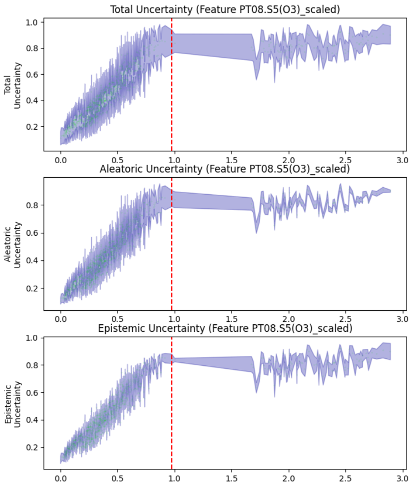
Dropconnect NOx(GT)



Dropconnect PT08.S1(CO)

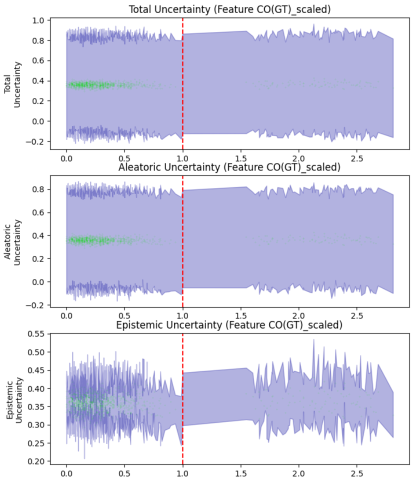


Dropconnect PT08.S3(NOx)

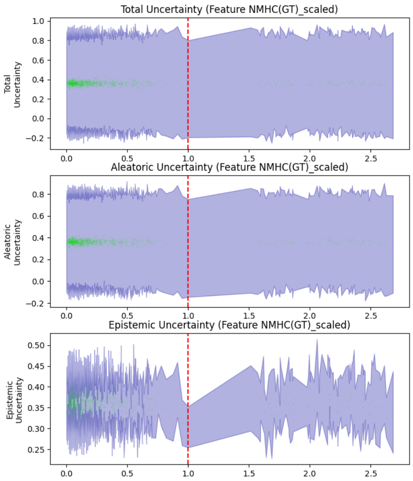


Dropconnect PT08.S5(O3)

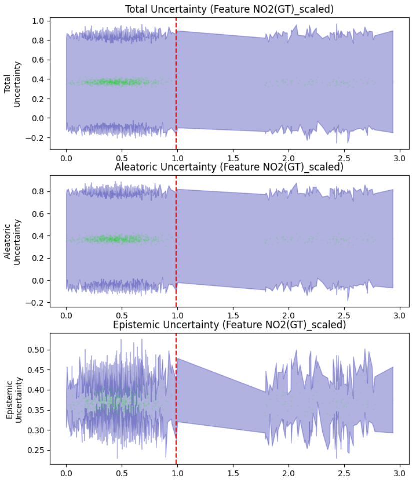
Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using β-NLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.



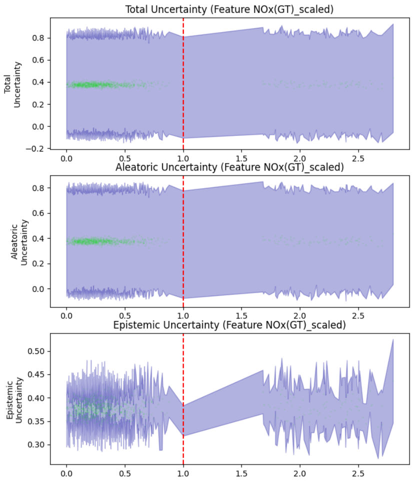
Flipout CO(GT)



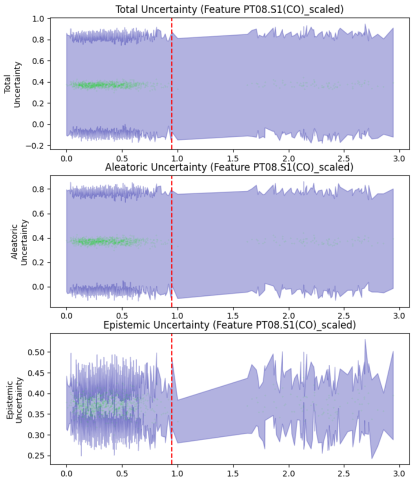
Flipout NMHC(GT)



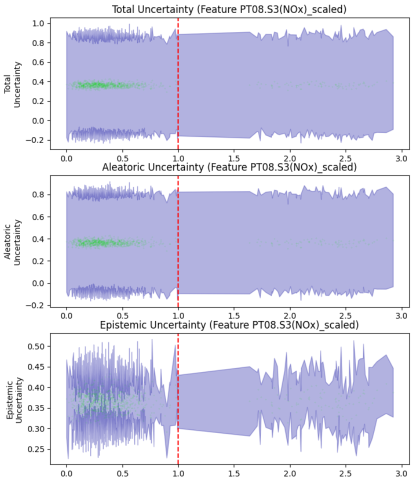
Flipout NO2(GT)



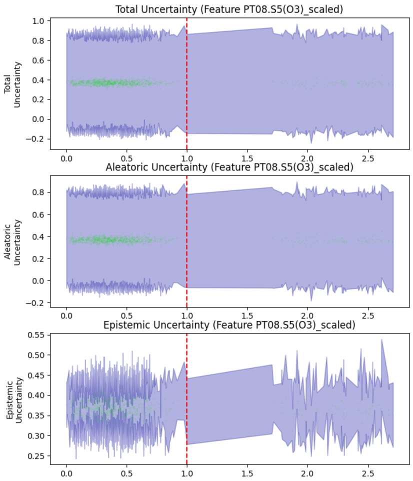
Flipout NOx(GT)



Flipout PT08.S1(CO)



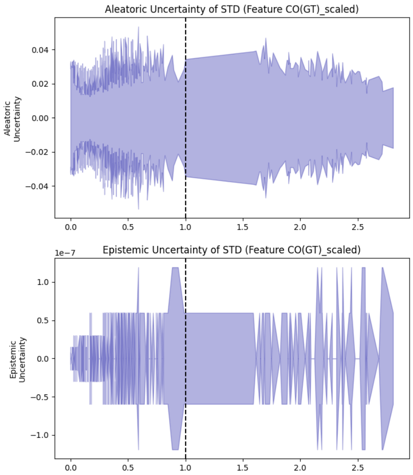
Flipout PT08.S3(NOx)



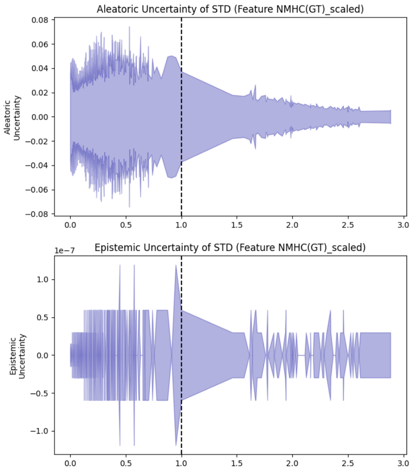
Flipout PT08.S5(O3)

Comparison of total, aleatoric and epistemic uncertainty on all uncertainty methods using BNLL loss function for an individual feature (PT08.S2(NMHC)), all values to the left of the vertical red line are ID and the rest are OOD.

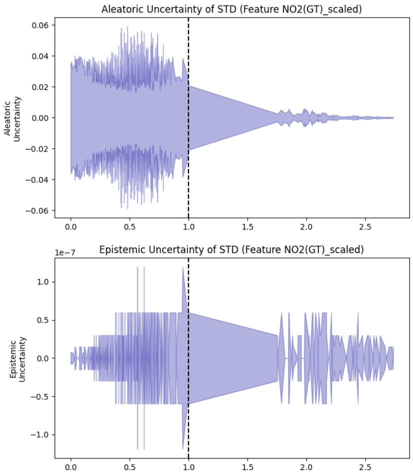
C.3 Visualization of uncertainties with only standard Deviation, NLL loss function



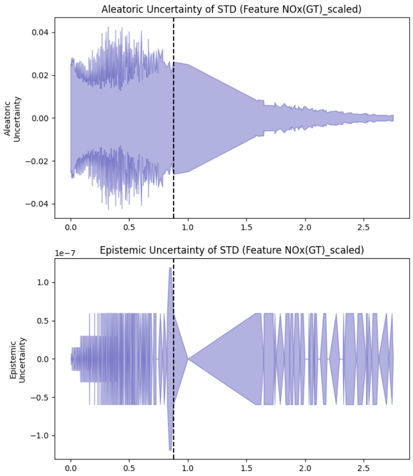
Classic CO(GT)



Classic NMHC(GT)

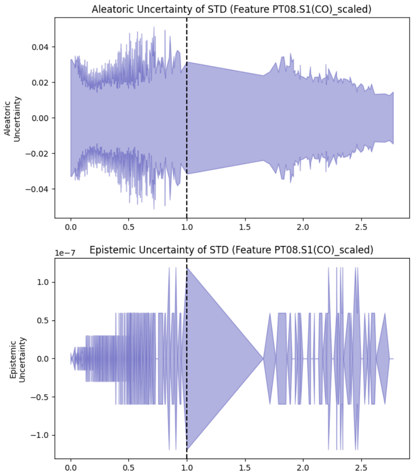


Classic NO2(GT))


Aleatoric Uncertainty of STD (Feature NOx(GT)_scaled)

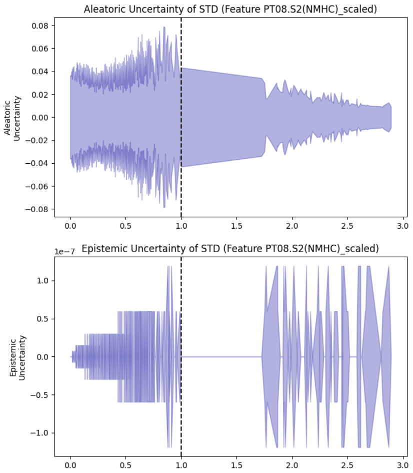
Epistemic Uncertainty of STD (Feature NOx(GT)_scaled)

Classic NOx(GT)


Aleatoric Uncertainty of STD (Feature PT08.S1(CO)_scaled)

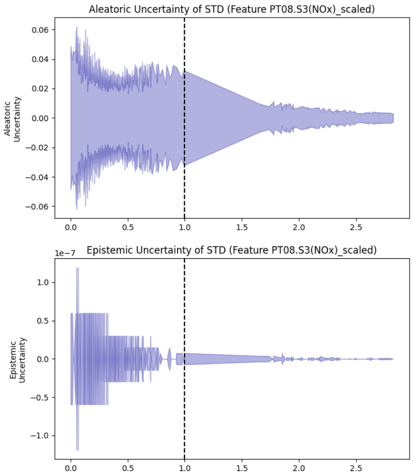
Epistemic Uncertainty of STD (Feature PT08.S1(CO)_scaled)

Classic PT08.S1(CO)

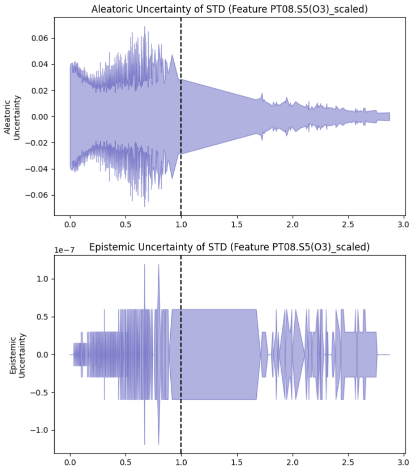

Aleatoric Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

Epistemic Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

Classic PT08.S2(NMHC)

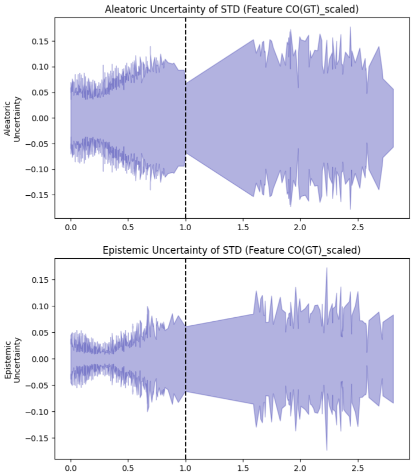


Classic PT08.S3(NOx

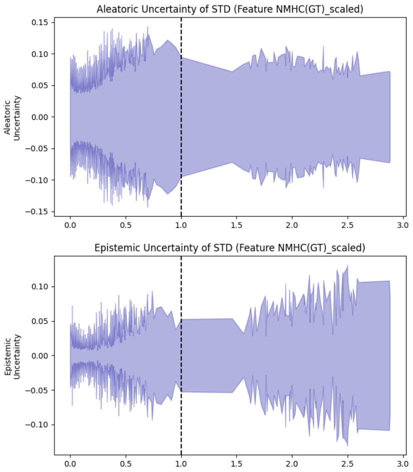


Classic PT08.S5(O3)

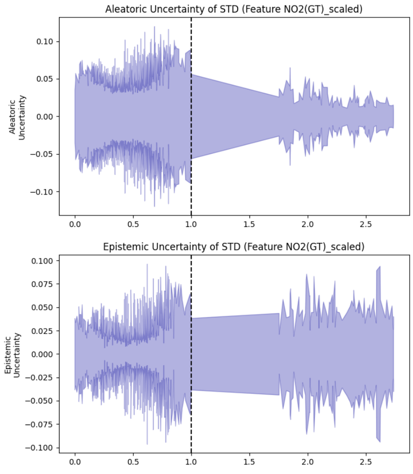
Comparison of aleatoric and epistemic uncertainty on Classic without the mean predictions using NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



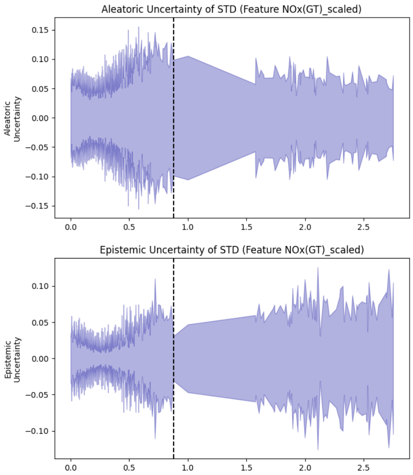
Dropout CO(GT)



Dropout NMHC(GT)

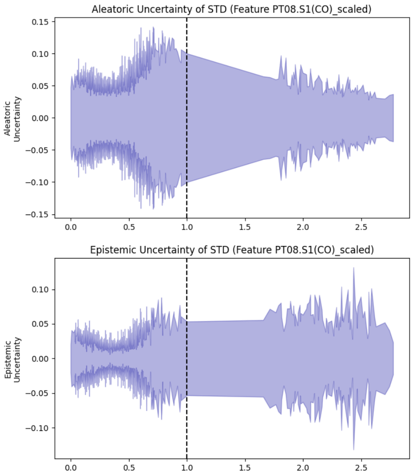


Dropout NO2(GT))


Aleatoric Uncertainty of STD (Feature NOx(GT)_scaled)

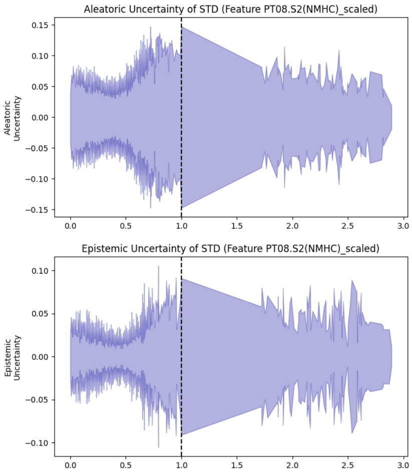
Epistemic Uncertainty of STD (Feature NOx(GT)_scaled)

Dropout NOx(GT)


Aleatoric Uncertainty of STD (Feature PT08.S1(CO)_scaled)

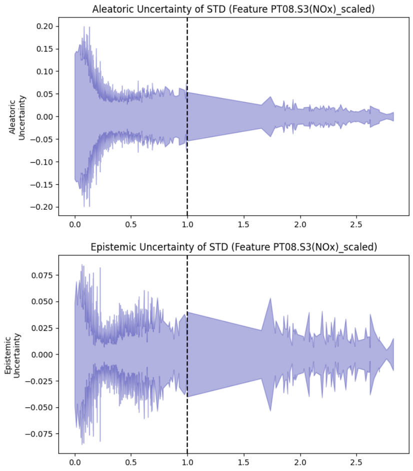
Epistemic Uncertainty of STD (Feature PT08.S1(CO)_scaled)

Dropout PT08.S1(CO)

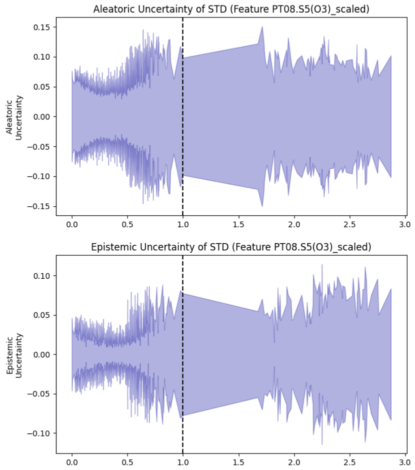

Aleatoric Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

Epistemic Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

Dropout PT08.S2(NMHC)

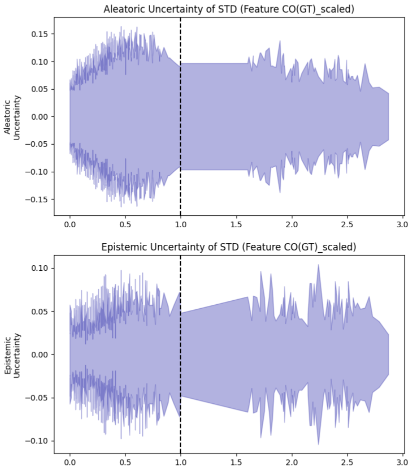


Dropout PT08.S3(NOx



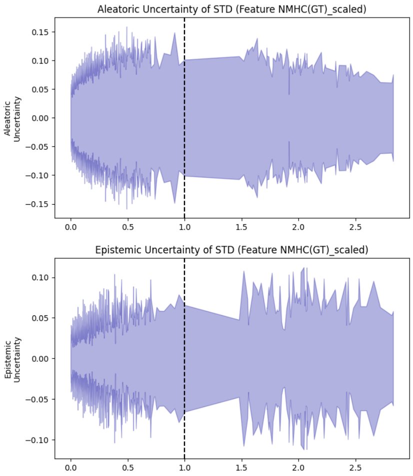
Dropout PT08.S5(O3)

Comparison of aleatoric and epistemic uncertainty on Dropout without the mean predictions using NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD


Aleatoric Uncertainty of STD (Feature CO(GT)_scaled)

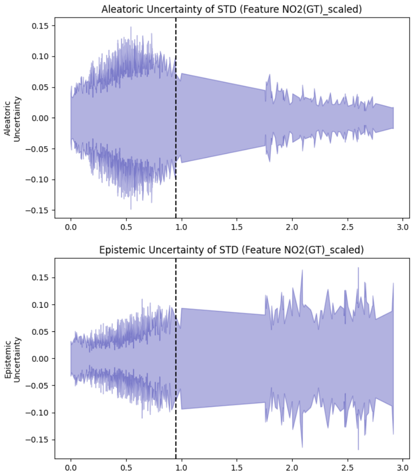
Epistemic Uncertainty of STD (Feature CO(GT)_scaled)

Dropconnect CO(GT)


Aleatoric Uncertainty of STD (Feature NMHC(GT)_scaled)

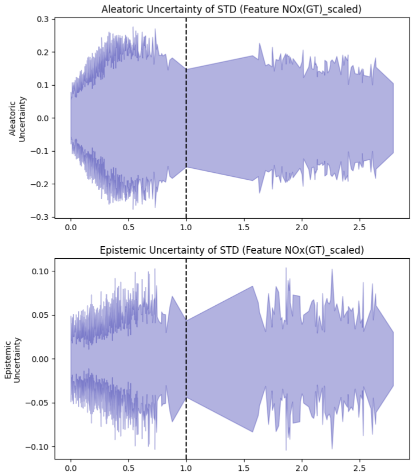
Epistemic Uncertainty of STD (Feature NMHC(GT)_scaled)

Dropconnect NMHC(GT)


Aleatoric Uncertainty of STD (Feature NO2(GT)_scaled)

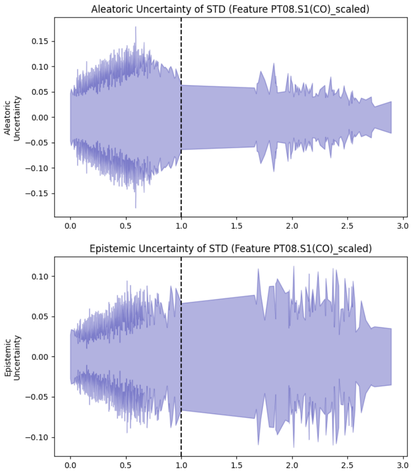
Epistemic Uncertainty of STD (Feature NO2(GT)_scaled)

Dropconnect NO2(GT))


Aleatoric Uncertainty of STD (Feature NOx(GT)_scaled)

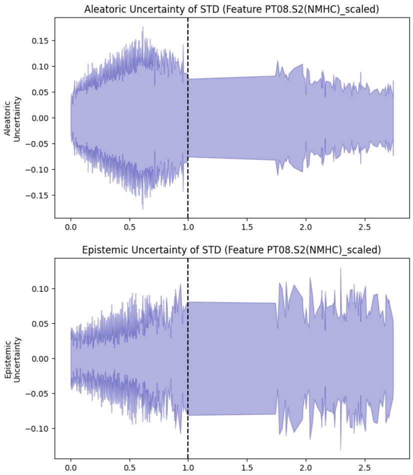
Epistemic Uncertainty of STD (Feature NOx(GT)_scaled)

Dropconnect NOx(GT)


Aleatoric Uncertainty of STD (Feature PT08.S1(CO)_scaled)

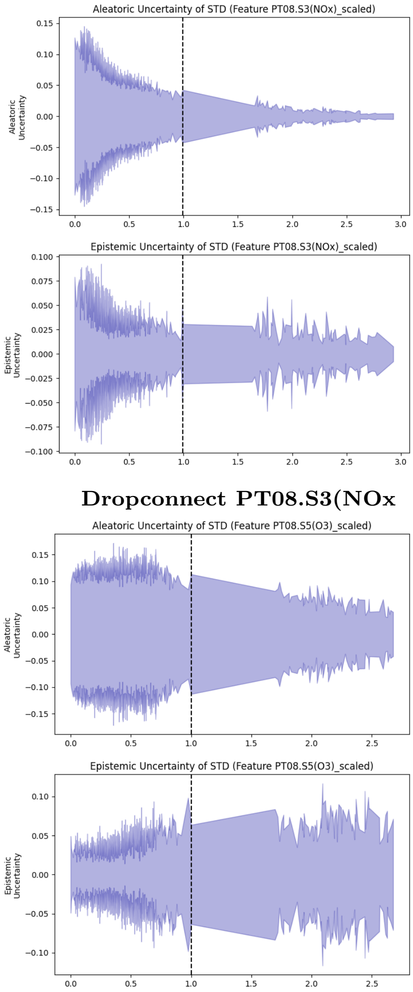
Epistemic Uncertainty of STD (Feature PT08.S1(CO)_scaled)

Dropconnect PT08.S1(CO)


Aleatoric Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

Epistemic Uncertainty of STD (Feature PT08.S2(NMHC)_scaled)

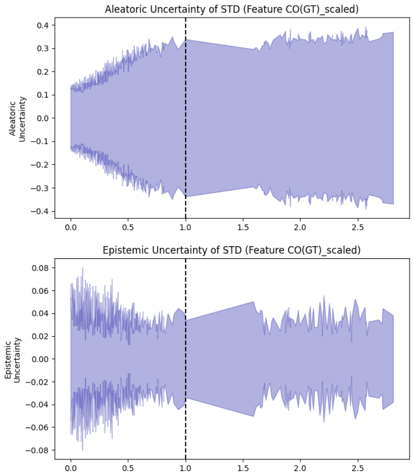
Dropconnect PT08.S2(NMHC)



Dropconnect PT08.S3(NOx

Dropconnect PT08.S5(O3)

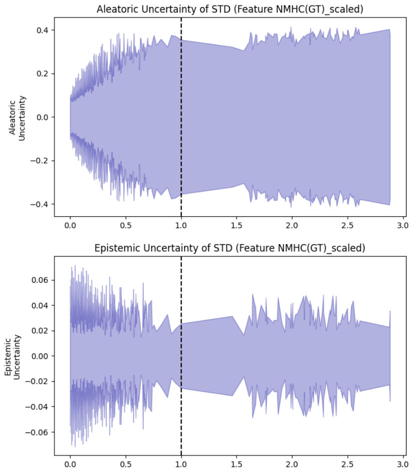
Comparison of aleatoric and epistemic uncertainty on Dropconnect without the mean predictions using NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



Aleatoric Uncertainty of STD (Feature CO(GT)_scaled)

Epistemic Uncertainty of STD (Feature CO(GT)_scaled)

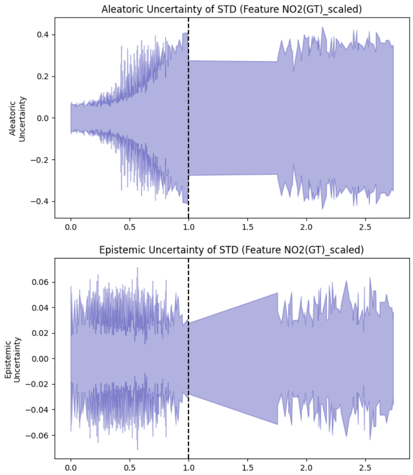
Flipout CO(GT)



Aleatoric Uncertainty of STD (Feature NMHC(GT)_scaled)

Epistemic Uncertainty of STD (Feature NMHC(GT)_scaled)

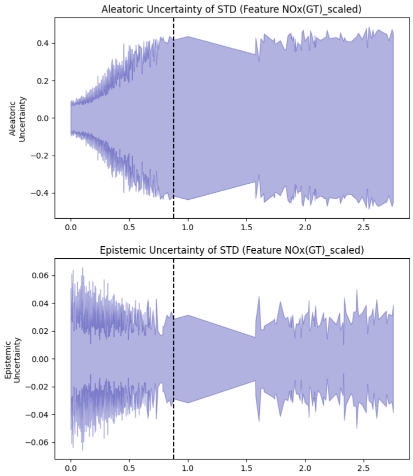
Flipout NMHC(GT)



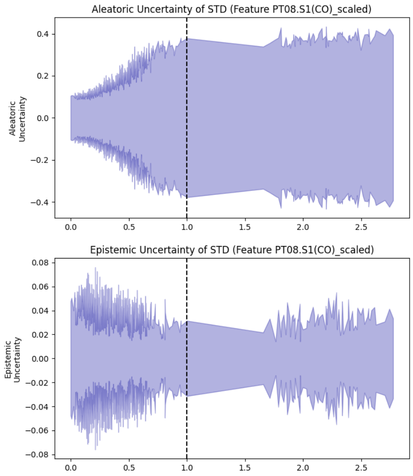
Aleatoric Uncertainty of STD (Feature NO2(GT)_scaled)

Epistemic Uncertainty of STD (Feature NO2(GT)_scaled)

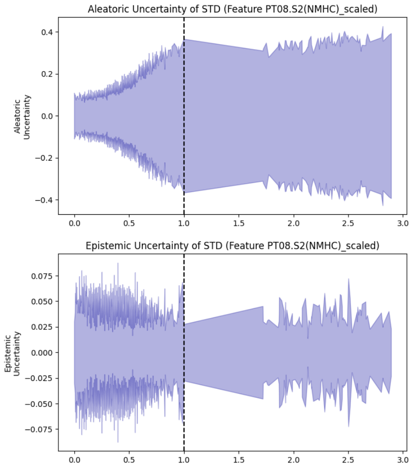
Flipout NO2(GT))



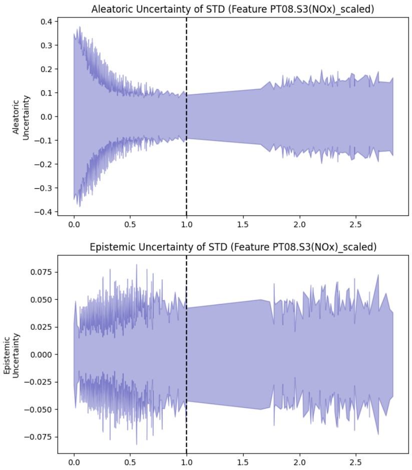
Flipout NOx(GT)



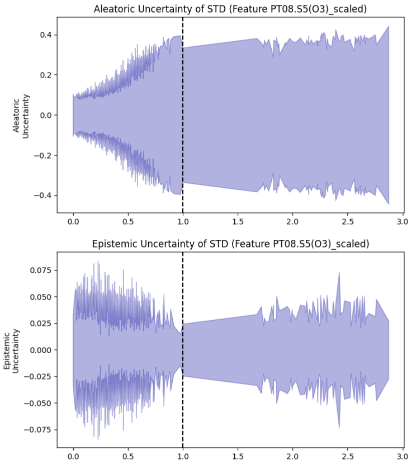
Flipout PT08.S1(CO)



Flipout PT08.S2(NMHC)

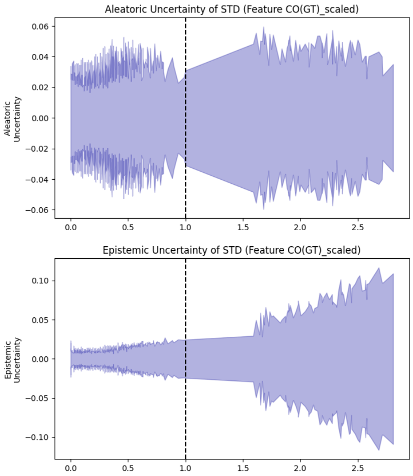


Flipout PT08.S3(NOx

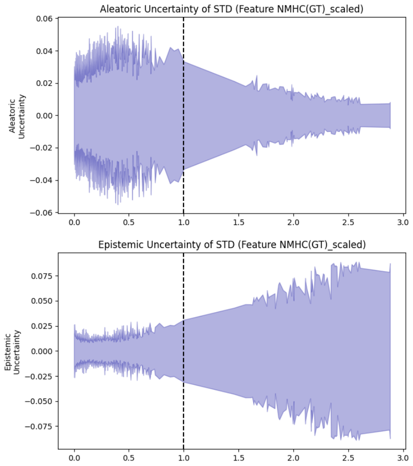


Flipout PT08.S5(O3)

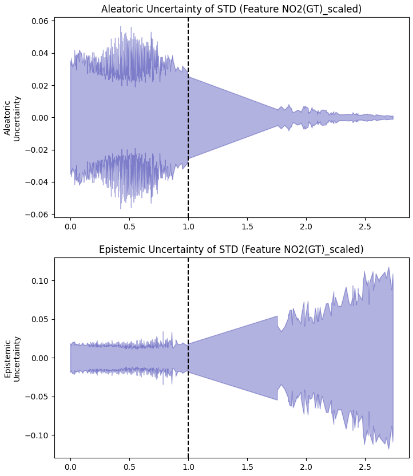
Comparison of aleatoric and epistemic uncertainty on Flipout without the mean predictions using NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



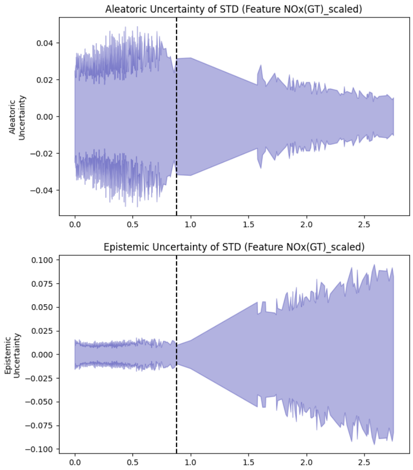
Ensemble CO(GT)



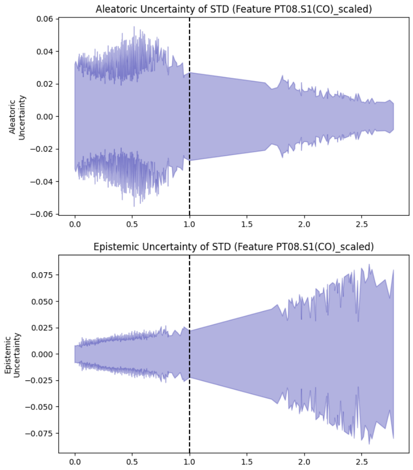
Ensemble NMHC(GT)



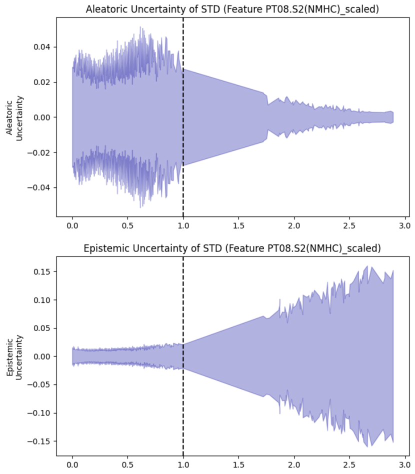
Ensemble NO2(GT))



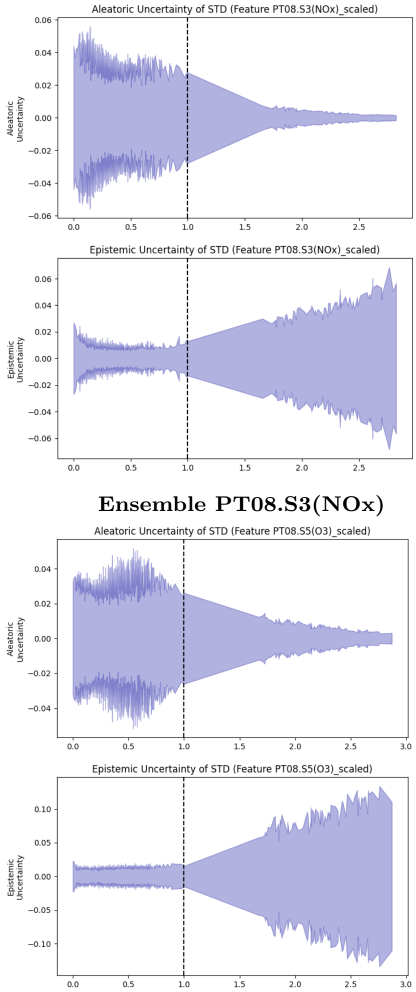
Ensemble NOx(GT)



Ensemble PT08.S1(CO)



Ensemble PT08.S2(NMHC)

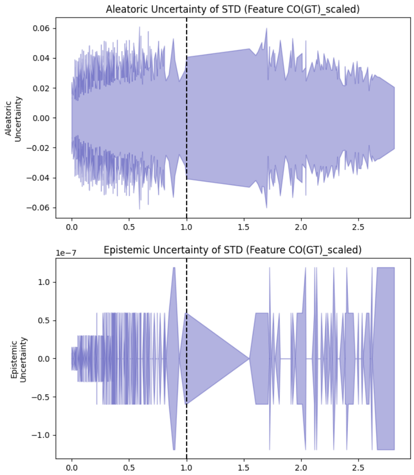


Ensemble PT08.S3(NOx)

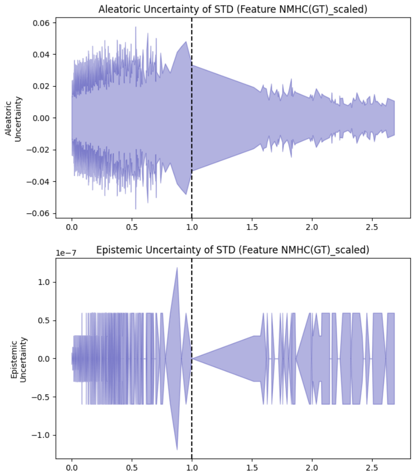
Ensemble PT08.S5(O3)

Comparison of aleatoric and epistemic uncertainty on Ensemble without the mean predictions using NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD

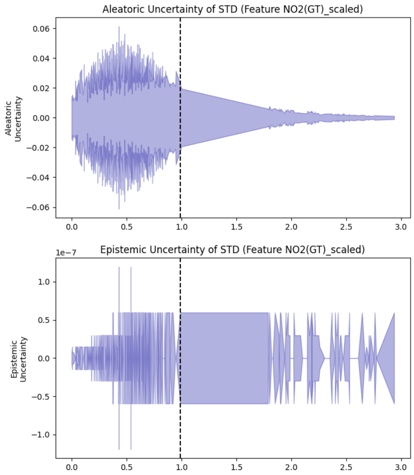
C.4 Visualization of uncertainties with only Standard Deviation, β-NLL loss function



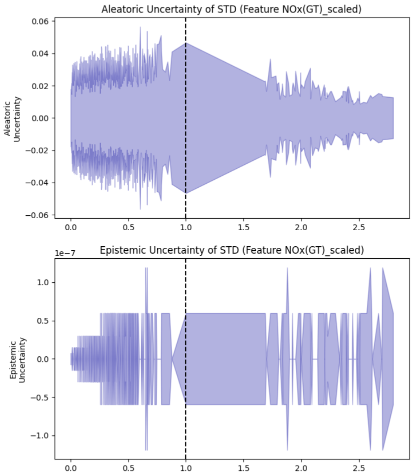
Classic CO(GT)



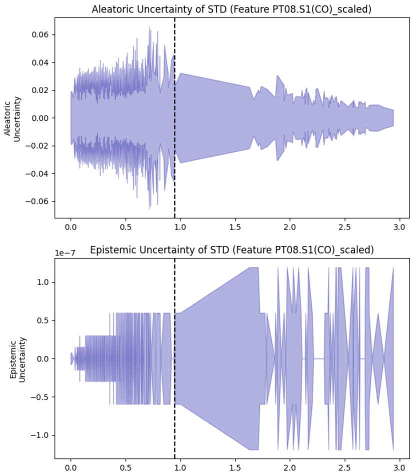
Classic NMHC(GT)



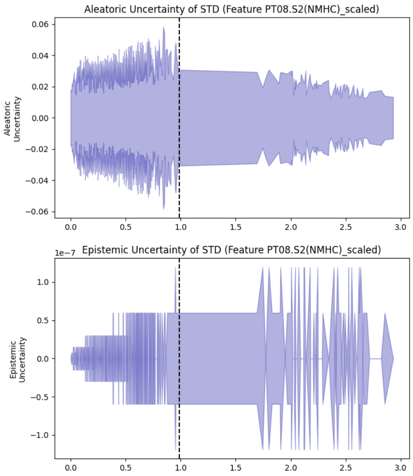
Classic NO2(GT))



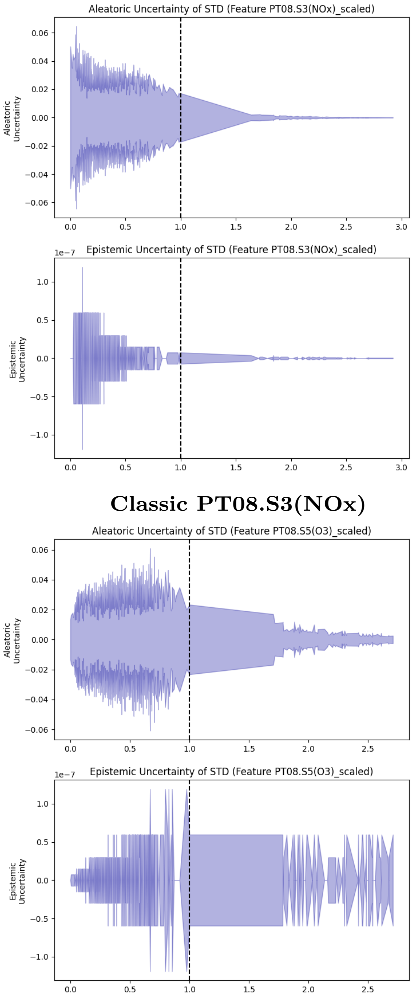
Classic NOx(GT)



Classic PT08.S1(CO)



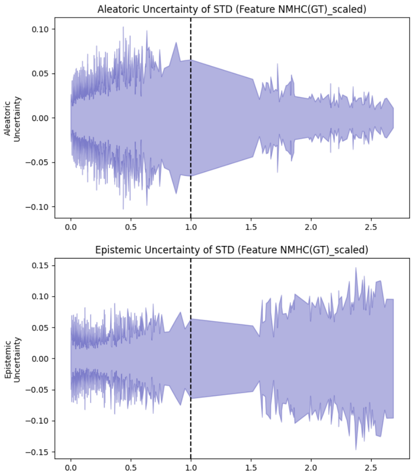
Classic PT08.S2(NMHC)



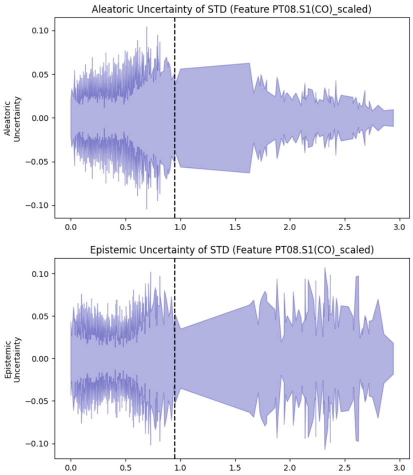
Classic PT08.S3(NOx)

Classic PT08.S5(O3)

Comparison of aleatoric and epistemic uncertainty on Classic without the mean predictions using β-NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD

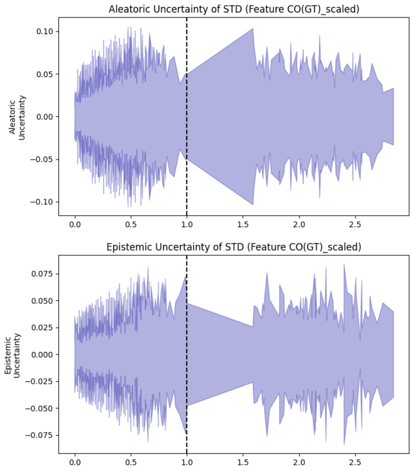


Dropout NMHC(GT)

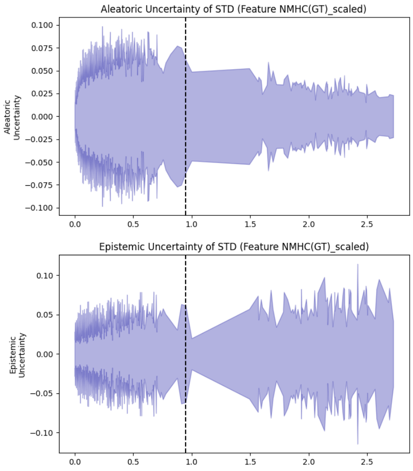


Dropout PT08.S1(CO)

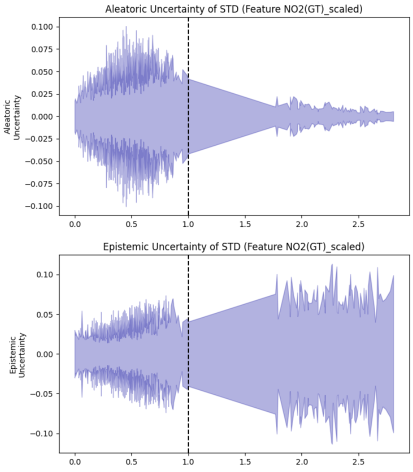
Comparison of aleatoric and epistemic uncertainty on Droput without the mean predictions using β-NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



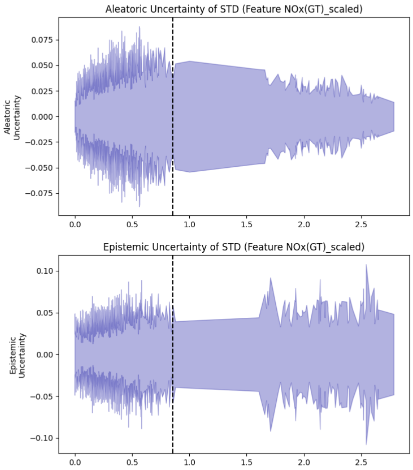
Dropconnect CO(GT)



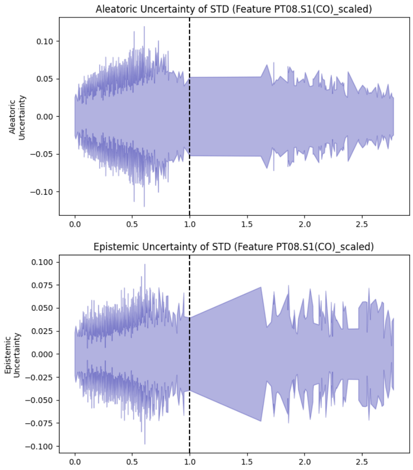
Dropconnect NMHC(GT)



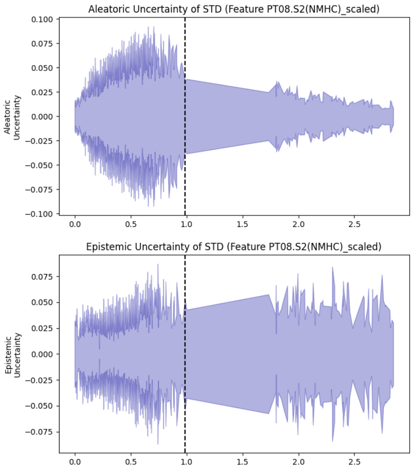
Dropconnect NO2(GT)



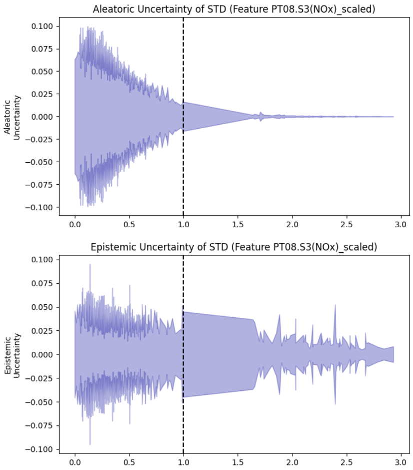
Dropconnect NOx(GT)



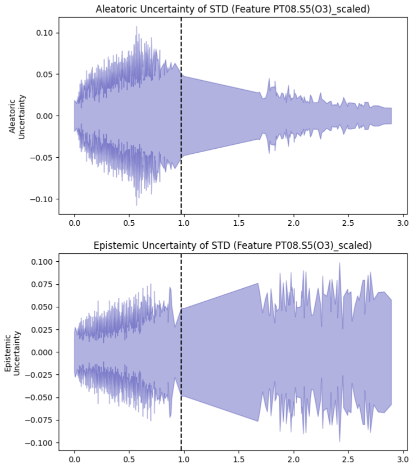
Dropconnect PT08.S1(CO)



Dropconnect PT08.S2(NMHC)

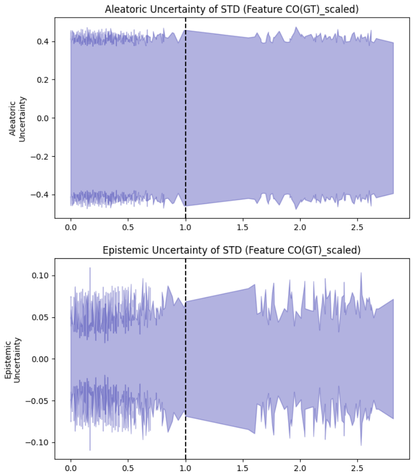


Dropconnect PT08.S3(NOx)

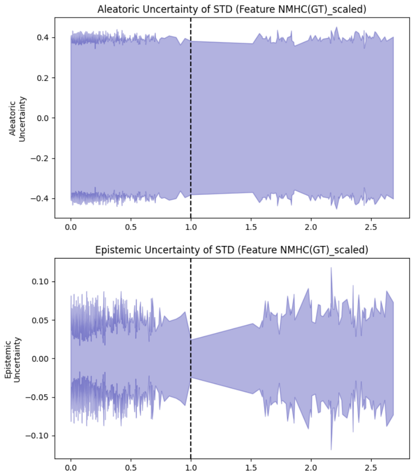


Dropconnect PT08.S5(O3)

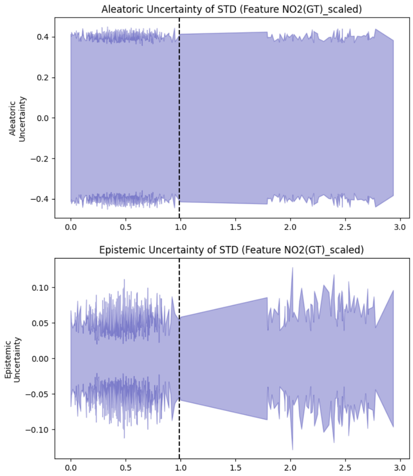
Comparison of aleatoric and epistemic uncertainty on Dropconnect without the mean predictions using β-NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



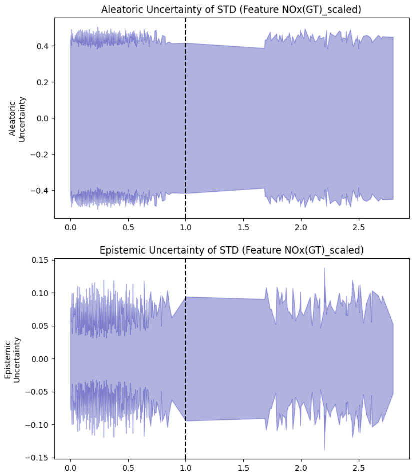
Flipout CO(GT)



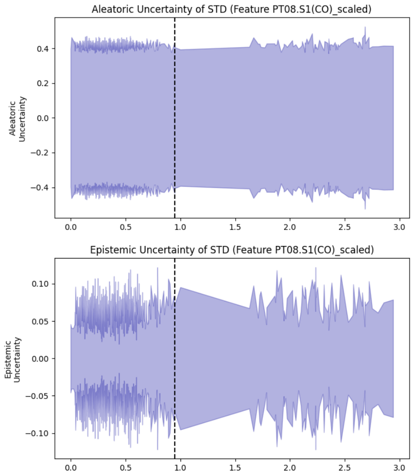
Flipout NMHC(GT)



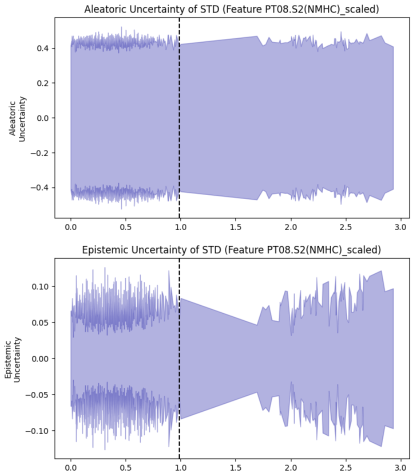
Flipout NO2(GT))



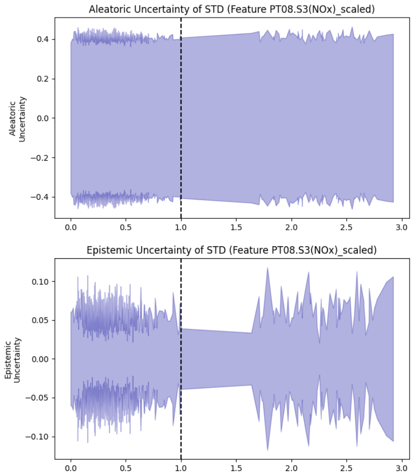
Flipout NOx(GT)



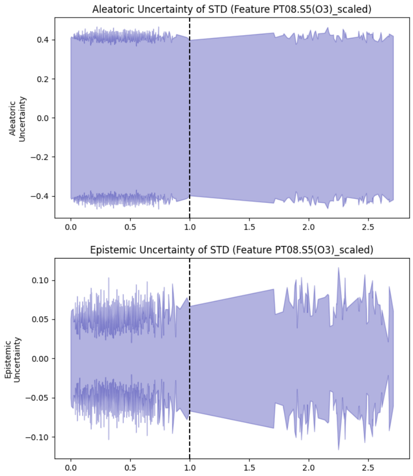
Flipout PT08.S1(CO)



Flipout PT08.S2(NMHC)

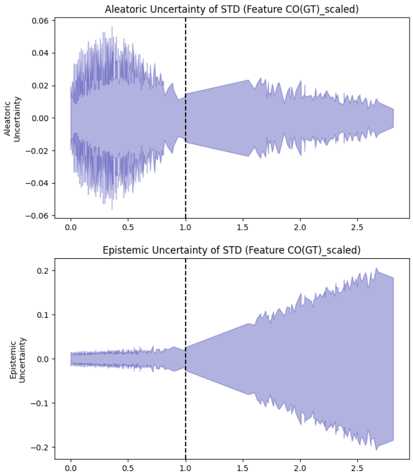


Flipout PT08.S3(NOx)



Flipout PT08.S5(O3)

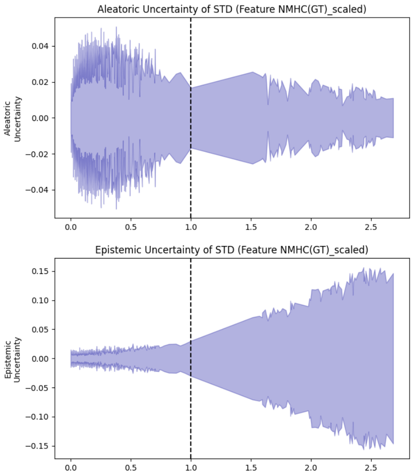
Comparison of aleatoric and epistemic uncertainty on Flipout without the mean predictions using β-NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD



Aleatoric Uncertainty of STD (Feature CO(GT)_scaled)

Epistemic Uncertainty of STD (Feature CO(GT)_scaled)

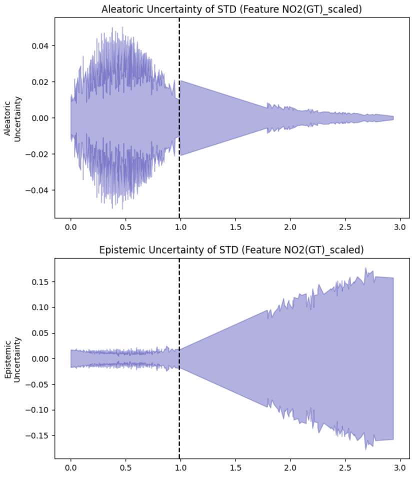
Ensemble CO(GT)



Aleatoric Uncertainty of STD (Feature NMHC(GT)_scaled)

Epistemic Uncertainty of STD (Feature NMHC(GT)_scaled)

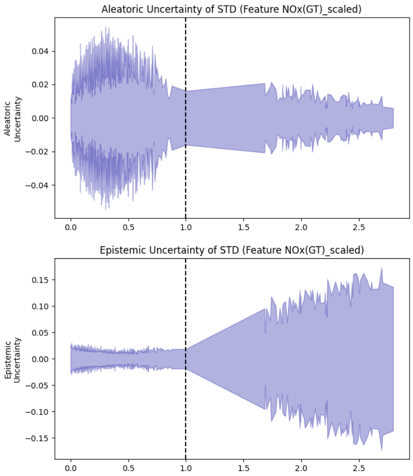
Ensemble NMHC(GT)



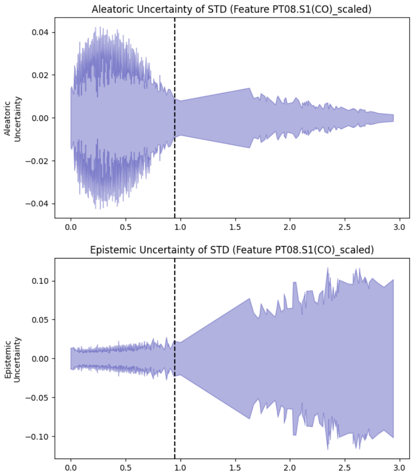
Aleatoric Uncertainty of STD (Feature NO2(GT)_scaled)

Epistemic Uncertainty of STD (Feature NO2(GT)_scaled)

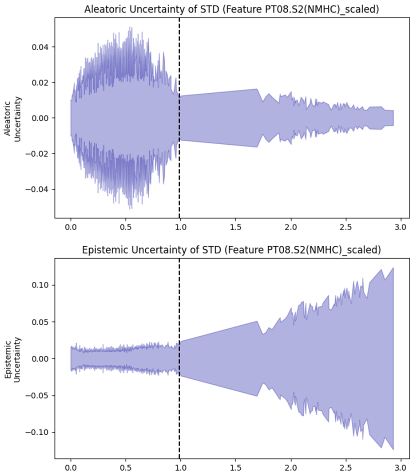
Ensemble NO2(GT))



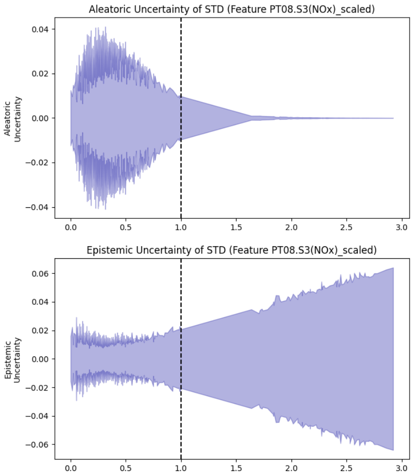
Ensemble NOx(GT)



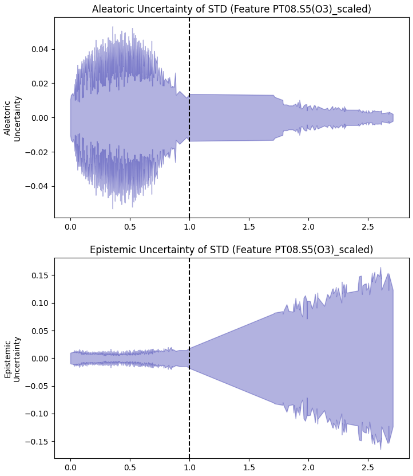
Ensemble PT08.S1(CO)



Ensemble PT08.S2(NMHC)



Ensemble PT08.S3(NOx)



Ensemble PT08.S5(O3)

Comparison of aleatoric and epistemic uncertainty on Ensemble without the mean predictions using β-NLL loss function for remaining features, all values to the left of the vertical black line are ID and the rest are OOD