1.3 Global error quantification

This section corresponds to box labelled "Error quantification" in Figure 1.4.

After running the model, the simplest analysis of its performance consists of measuring the aggregated error of predictions against ground true values over the whole test set. Different metrics and criteria can be adopted for this task. Common error measures are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). These metrics account for the distance between points representing ground true values ($\mathbf{Y}$) and model's predictions ($\tilde{\mathbf{Y}}$), taken as the $L1$ norm (MAE) or the square of the $L2$ norm (RMSE) of the distances. In many industrial applications of machine learning, difference between ground true and predicted values can be more or less important depending on whether that difference is due to *overestimating* or *underestimating*. Take, for instance, the case of MS-S18 model, whose predictions are a measure of the probability of failure of aeronautical structural components. Clearly, underestimating the risk is much more dangerous than overestimating it. For cases such as this one, a useful measure of the error is the **residue**. The residual error of a given point i is measured as

$$\mathbf{e}(i) = \mathbf{Y}(i) - \tilde{\mathbf{Y}}(i) \tag{1.2}$$

The drawback of the residual error is that, when using it as an aggregated indicator for the whole test set, residues can cancel out. A null MAE or RMSE account for a perfectly fitted model (ground true values and predictions are equal). That is not the case for the residual error.

An interesting scalar metric for the global performance of the model is the coefficient of determination R^2 [16] of the scatter distribution of $\mathbf{Y}$ vs $\tilde{\mathbf{Y}}$. R^2 is a measure of the goodness of fit of predicted to ground true values. Illustration of this is provided in Figure 1.10. The most important conclusion of Figure 1.10 is that the error shows to be heteroscedastic[2]. This property makes it necessary to study the error distribution, as well as the error distribution conditioned on input and output space ($P(e)$, $P(e|\mathbf{X})$ and $P(e|\mathbf{Y})$, resp.) as is discussed in the following sections.

[2]This means the variance of the error is not constant along some variable range (in this case, that variable is the output variable named "RF Net Tension"). In Figure 1.10 we can clearly see that dispersion grows with the output value. Cfr.[22, p. 374].

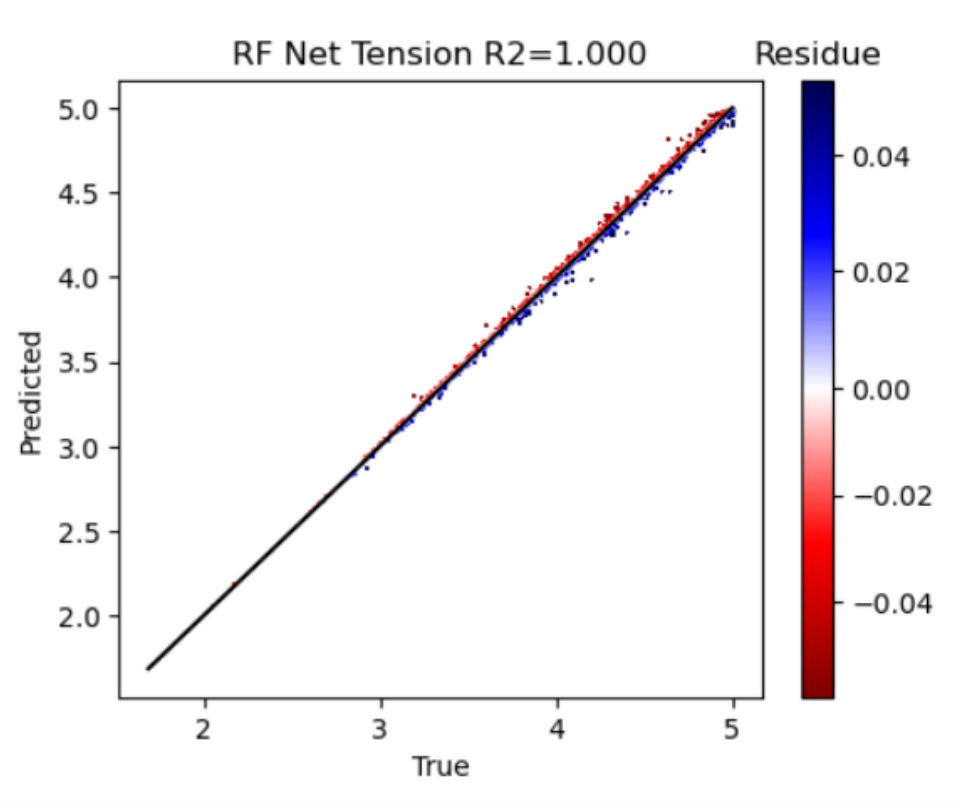

Figure 1.10: Scatter plot of ground true (x axis) against predicted values (y axis) and R^2 coefficient. In this case, $R^2 = 1,000$ indicates a perfect fit of predicted to ground true values. Mismatches due to underestimating failure risk are labelled in red, while those due to overestimating failure risk are labelled in blue. Similar graphs can be computed for every pair $\{y_j, \hat{y}_j\}$ of features in the output variables $\mathbf{Y} = \{y_1, y_2, \ldots, y_m\}$.

1.4 Distributed error quantification.

This section corresponds to box labelled "$P(E)$" in Figure 1.11. From now on, the terms "error" and "residue" (as defined in the previous section) will be used as synonyms.

Beyond global error statistics, the PDF of the residue, $P(e)$, is of great importance for

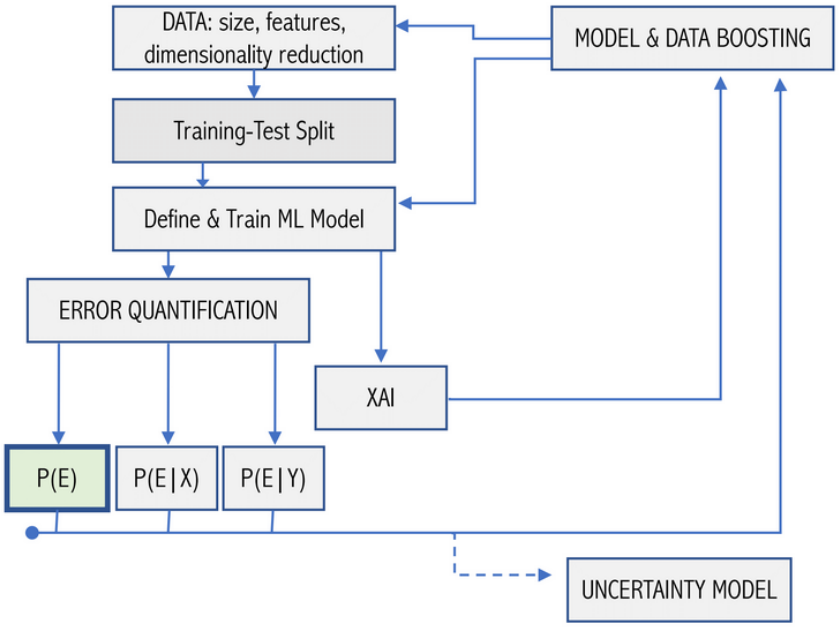

Figure 1.11: Location of section 1.4 in the validation pipeline.

validation purposes. The main goal of this section is finding the analytic expression of $P(e)$ and computing important statistics of it, and, when the first is not possible, computing the important statistics from the empirical distribution of the residue (sampled from test set points) with the method of bootstrapping[3]. There are three main reasons for studying $P(e)$:

A. Non gaussian error distribution

It is common in industry-applied problems to find situations in which the error between ground-true values and predictions coming from a regression model (let it be a surrogate ANN, some parametric regressor, etc.) which is supposed to fit any function describing a complex system follows absolutely non-Gaussian distributions (see *e.g.* [18–21]). Amongst the reasons for

[3] Vid. following paragraphs

this, the most frequent are, on the one hand, non-homogeneous data sampling in the training set (leading to uncovered regions and isolated points) which can cause poor model performance due to non-interpolation-regime operation, and on the other hand, the inherent difficulty encountered at predicting outputs for specific input configurations due to strongly non-linear physics or governing equations (mathematically this manifests in the form of high gradients). When the error is non-Gaussian, concentration statistics such as MAE or RMSE stop being informative. In such case, a comprehensive analysis on $P(e)$ is more adequate.

B. Outlier detection

Outliers are strange events in a population sampled from a known PDF, in the sense that it is not expected to find them, or that their position is far away from expected. Outlier detection helps identifying strange phenomena which the engineer in charge could decide to investigate. Imagine, for instance, that certain residue e_x was systematically sampled from the test set with an unusual frequency, compared with similar values. In this case, it would be necessary to assess whether this high frequency is statistically expectable from the residue's PDF or not. If e_x was found to be an outlier, determining the underlying reason triggering such high frequency would help with model boosting.

Outlier detection relies on knowing $P(e)$. The probability of finding a residue larger than a given magnitude x is measured as $p_{>x} = \int_x^\infty P(e)\, de$. If we find some x for which $p_{>x} << 1$, all samples of the residue $e > x$ would be classified as outliers. This simple idea lies behind standard outlier-detection methods such as the z-score and the gESD (which are later discussed).

C. Uncertainty measuring

The marginalised distribution of $P(e)$ is the first step in the journey towards building an **uncertainty model**. This is the ultimate milestone of the whole validation pipeline, since it provides precise information about *how much* and *when* the model's predictions are trustworthy. This is the whole point of section 1.7. The uncertainty model relies on the marginalised distribution of the residue for building confidence intervals which embed the model's error with a given statistical confidence level.

A simple plot can help us have a first intuition for the cause of subsequent results presented in this section. In Figure 1.12 ground true values and model's predictions are represented in a double histogram for MS-18.

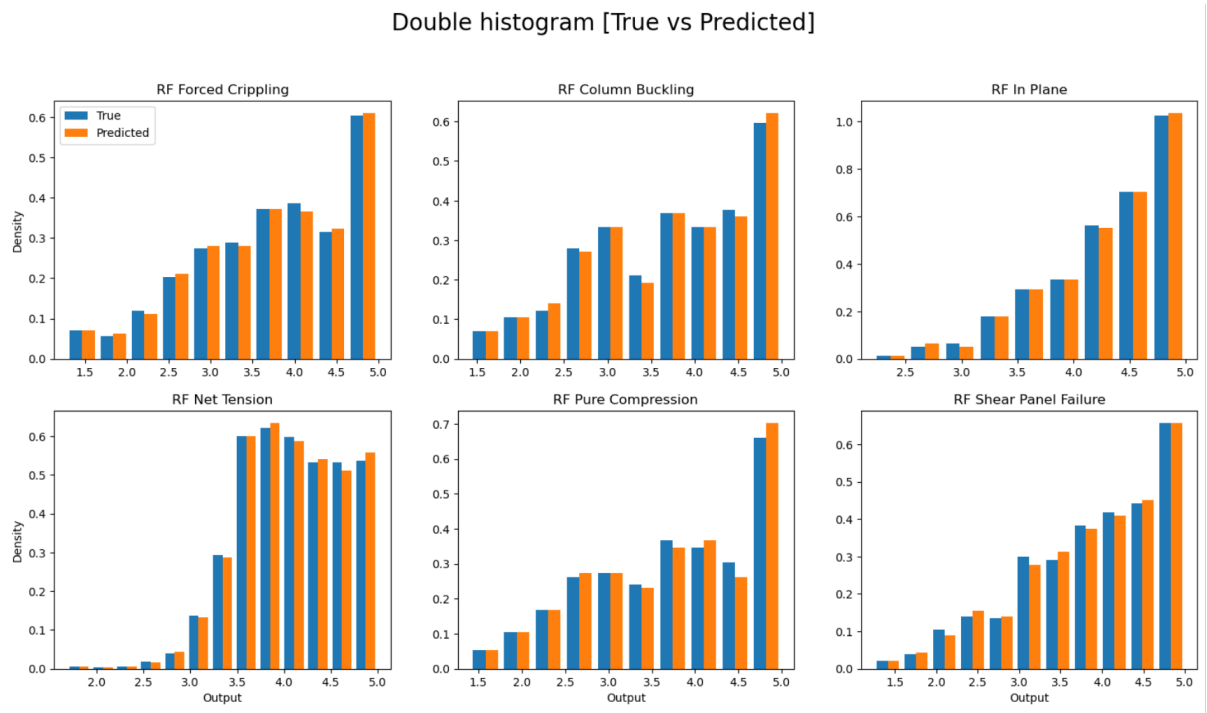

Figure 1.12: Double histogram depicting ground true values and model's predictions for the six output variables of MS-18.

If the model accurately predicts results along the whole output variables range (or equivalently it is not heteroscedastic), we would expect both the true and the predicted results to come from the same (unknown) distribution (this hypothesis is taken as $H0$). We can test this with a (2 sample) goodness-of-fit test like the Kolmogorov-Smirnov. Once again, the null hypothesis is rejected only at a 95% confidence level (*i.e.* if the p-value derived from the K-S test is lower than 0.05). The corresponding p-values of the K-S test for the six output variables of MS-18 are depicted in Table 1.7.

In Table 1.7 we can see how the K-S test rejects the null hypothesis in some cases despite distributions in Figure 1.12 looking very similar. This is due to the K-S test sensitivity to the size of datasets.

As it has been previously mentioned, outlier detection and uncertainty models both rely on the PDF of the residue, $P(e)$. The main goal of this section is finding the analytical definition of $P(e)$[4], and measuring important statistics of it. This is addressed with a focus similar to that followed in Figure 1.12 and Table 1.7, but instead of assessing the fitness of the predictions ($\hat{y}$) distribution to the ground true values (y) distribution, we try to assess the goodness of fit of the

[4]This might not always be possible. When it is not, non-parametrical bootstrapping is given as an alternate solution (vid. next paragraphs).

Table 1.7: p-value results for the 2-sample Kolmogorov-Smirnov test performed on distributions showed in Figure 1.12. Hypothesis $H0$ is that ground true and predicted values both come from the same (unknown) distribution.

Hypothesis Tests Results (p-value)	KS
RF Forced Crippling	0.03763
RF Column Buckling	0.11906
RF In Plane	0.03428
RF Net Tension	0.60624
RF Pure Compression	0.08367
RF Shear Panel Failure	0.03510

empirical residue distribution to some well-known distributions. For reference, the (empirical) error distribution of the six output variables is depicted in Figure 1.13, as well as the corresponding cumulative distributions, given in Figure 1.14.

Under the assumption that $P(e)$ can be described by some well-known parametric distribution, we use the 1-sample K-S test (coupled with a minimal-squares based optimizer to find the optimal set of parameters for each distribution) to compare distributions of Figure 1.13 to the following distributions:

- Normal
- Laplace
- Cauchy
- JohnsonSU

P-values from the K-S test are given in Table 1.8. As we can see, the normal distribution does not fit any of the output variables' error. While Laplace and Cauchy distributions' p-values from the K-S test are above the 0.05 threshold in five of the six output variables, they are well below the obtained p-values for the JohnsonSU[23] distribution (vid. Figure 1.15). In fact, when augmenting the dataset size from the illustrative-sized employed here (10,000 items) to a more realistic $800,000$ items, neither of Laplace and Cauchy distributions pass the test (their p-values drop to near-zero orders of magnitude). This happens due to the K-S sensibility to the size of

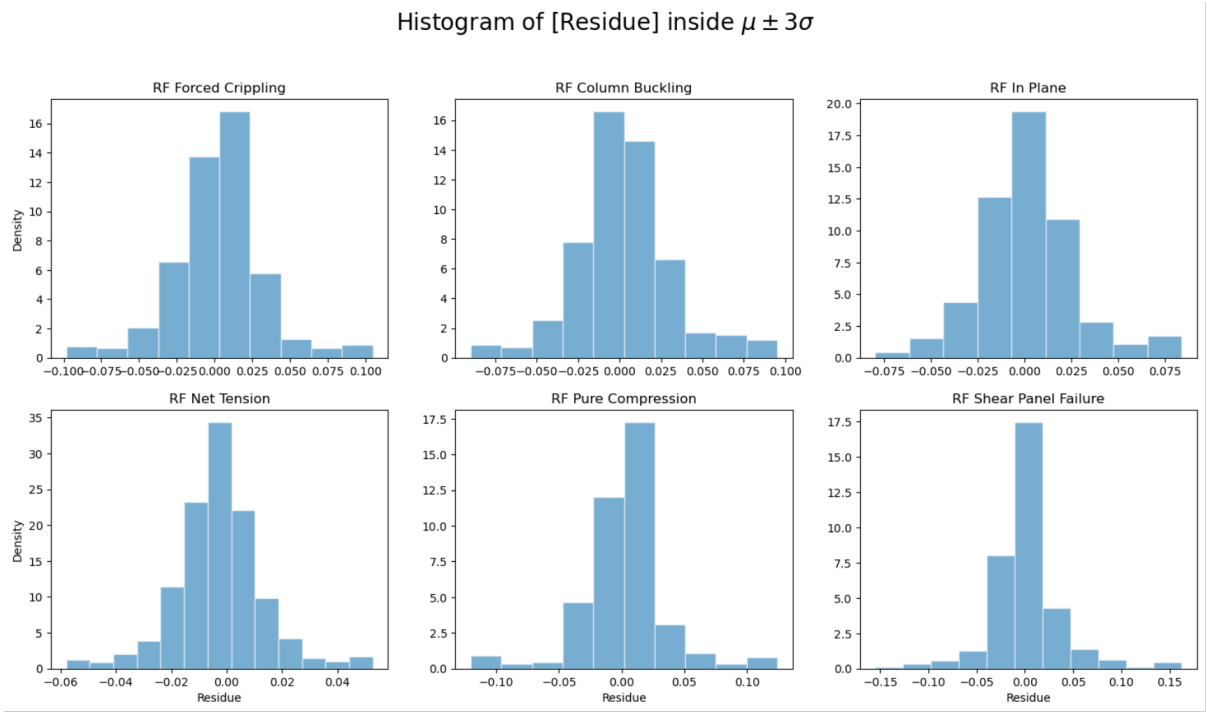

Figure 1.13: Empirical residue distribution sampled from the test set, for each of the six output variables of MS-S18. x-axis limits have been truncated to $\mu \pm 3\sigma$, where the most part of the error lies. Histograms have been appropriately binned for a correct visualization.

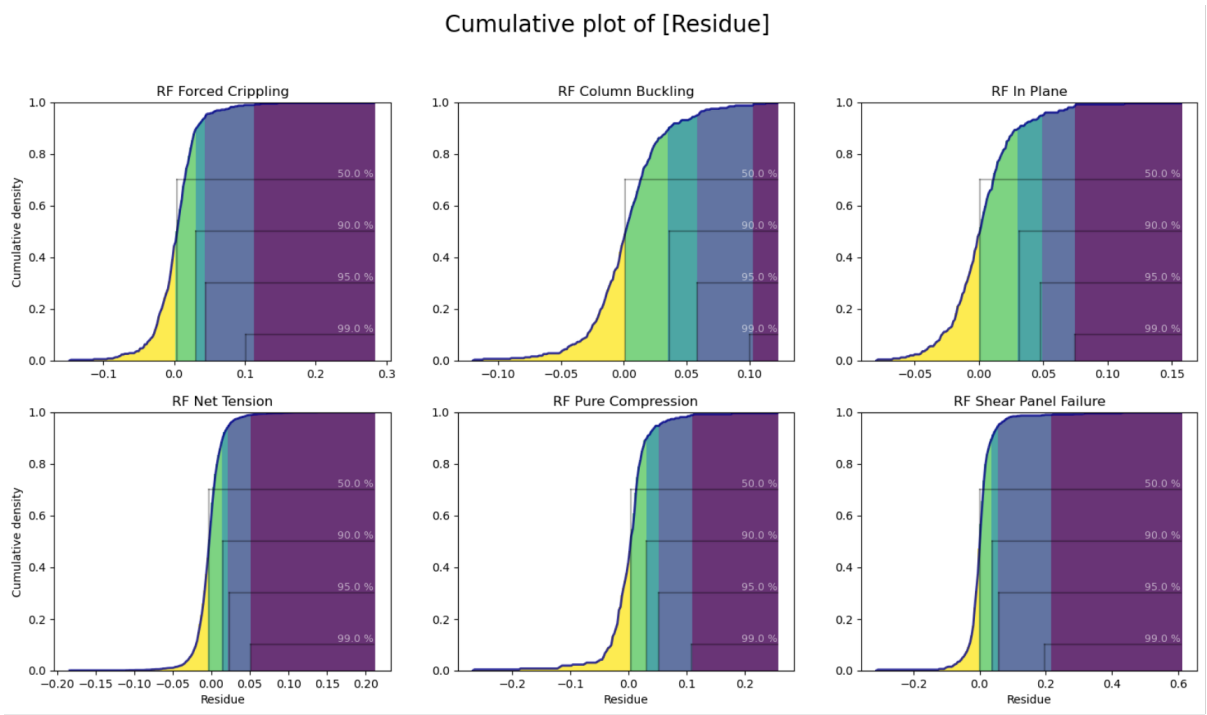

Figure 1.14: Cumulative error distributions corresponding to the PDFs showed (as binned histograms) in Figure 1.13.

Table 1.8: P-values of the K-S test comparing the empirical sample of $P(e)$ to the theoretical distributions indicated in each column. The null hypothesis $H0$ (the empirical distribution has been sampled from the one figuring in a given column) is rejected when $p - \text{value} < 0.05$.

	Fit to each output-metric			
	Norm	**Laplace**	**Cauchy**	**Johnsonsu**
RF Forced Crippling	0.004	0.138	0.138	0.613
RF Column Buckling	0.043	0.328	0.435	0.954
RF In Plane	0.005	0.691	0.691	0.616
RF Net Tension	0.000	0.060	0.001	0.843
RF Pure Compression	0.001	0.053	0.438	0.791
RF Shear Panel Failure	0.000	0.002	0.453	0.716

data, which makes the test more strict when the datasets are large (as would be expected). We conclude that the Laplace and Cauchy distributions just pass the K-S anecdotally for the unrealistically small dataset size which has been used for illustrative purposes, and we also conclude that the only theoretical distribution (of the list which has been checked) that fits the MS-18's error distribution is the JohnsonSU.

Provided that there exists a simple transformation of the JohnsonSU distribution's random variable (vid. Figure 1.16) that converges to a normal distribution, one can, with the information obtained from the K-S test (that is, assuming the error data comes from a JohnsonSU distribution) apply standard outlier detection methods to a transformed variable $z \sim \mathcal{N}(0, 1)$, like the z-score (every point located outside $\mu \pm 3\sigma$ is considered to be an outlier) and the generalised Extreme Studentized Deviate[24] (gESD), thus fulfilling the aims described at the beginning of this section concerning outlier detection (vid. Table 1.9).

Results shown in Table 1.7 rise some concerns about the method followed until now. The immediate concern that arises is what would happen if we were unable to find a theoretical distribution that fits some variable's error distribution with a statistically significant confidence level. In fact, the most common statistical distributions (normal, Laplace, Cauchy) do not properly fit the MS-18 error distribution (when using an industrial-sized dataset, not the one employed for illustration here), and we've had to rely on the (rather exotic) JohnsonSU distribution. Without a parametrized distribution that properly fits the error, an uncertainty model cannot be computed. Recall building an accurate uncertainty model is the main motivation for this section. Computing empirical statistics of the empirical distribution of the residue

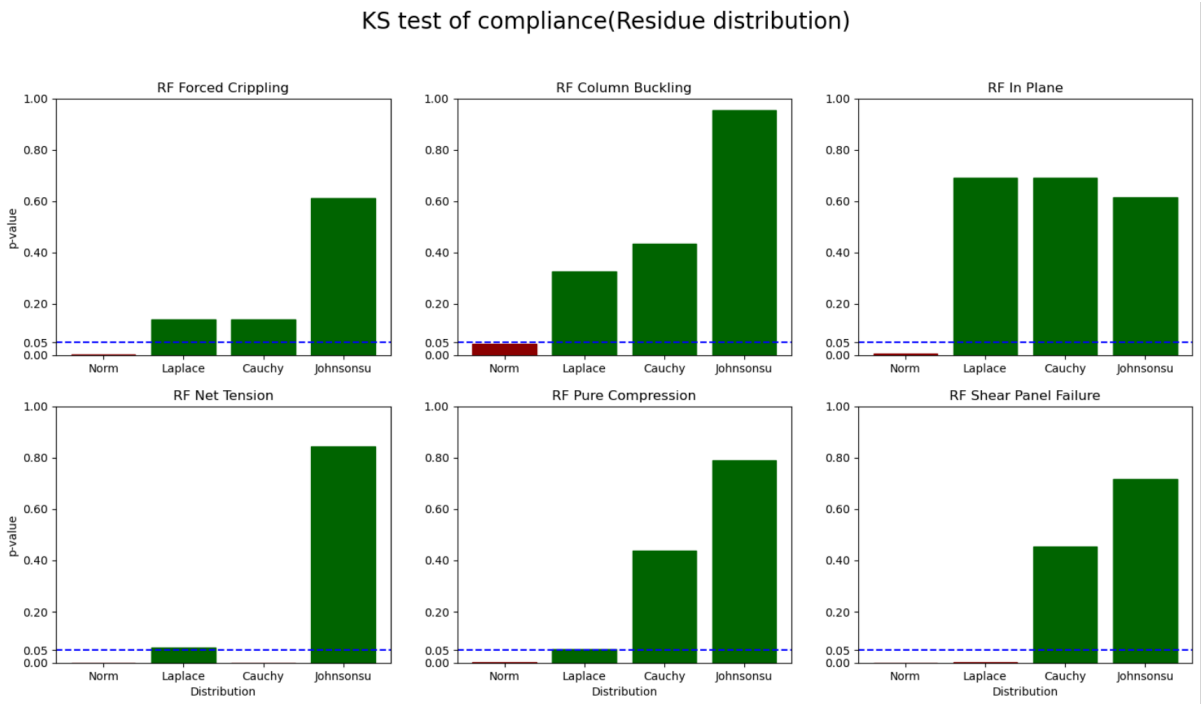

Figure 1.15: Graphical comparison of the p-values resulting from the K-S test for the MS-18 data.

(obtained from test set points) is possible, but assuming the empirical statistics are the same than the true, theoretical ones is not possible. To solve this obstacle, an alternative method for calculating informative statistics of the error's distribution that do not rely on knowing the parametrized analytic expression of it is therefore needed. The method provided here is known as non-parametric bootstrapping[17] and the main concept behind it is showed in algorithm 2. This algorithm:

1. Samples N points with replacement from the original population S. N.B. replacement makes the new and the original populations (possibly) different.
2. Statistic x is calculated in the new population.
3. Steps 1 and 2 are repeated $M >> 1$ times, giving a collection of x's (called X).
4. If M is sufficiently large, X converges to a Gaussian population. The bootstrapped CI for the statistic x with a 95% confidence is bounded by the percentiles 2.5% and 97.5% of X.

Some informative statistics of the error distribution are given in Table 1.10. The statistics are computed twice, once in the empirical distribution of $P(e)$ sampled from the outputs of $\mathcal{S}^{\text{test}}$,

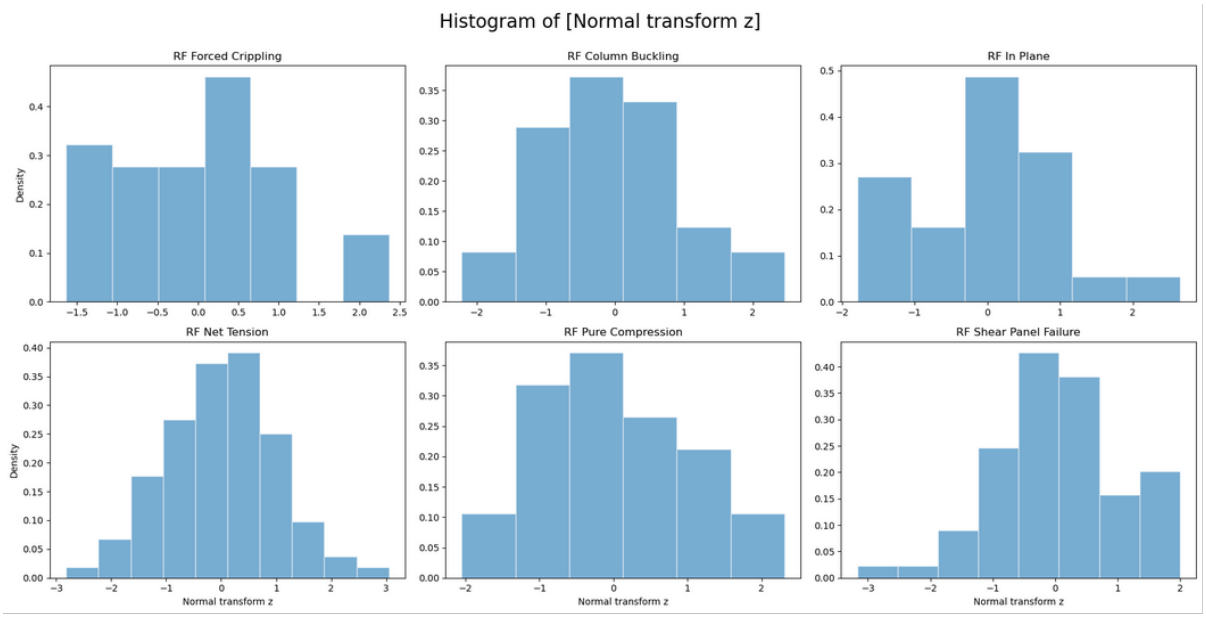

Figure 1.16: (Binned) histograms depicting the distribution of the variable $z = \gamma + \sinh \frac{e-\xi}{\lambda}$, where $e \sim JohnsonSU(\gamma, \delta, \xi, \lambda)$ is the sampled residue. It can be shown that z converges to a normal distribution[23].

Table 1.9: Outlier detection taking $e \sim JohnsonSU$ as $H0$. N.B. for this table the usual requirements for classifying a point as an outlier ($z \in \sigma \pm 3\mu$ for the z-score and significance $1 - a \geq 95\%$ for the gESD) have been softened here for illustration purposes to $z \in \sigma \pm 1.2\mu$ and $a = 0.45$ for both tests, respectively.

Outliers assuming JohnsonSU distribution

	RF Forced Crippling	RF Column Buckling	RF In Plane	RF Net Tension	RF Pure Compression	RF Shear Panel Failure
st.normaltest pvalue	0.64	0.86	0.44	0.76	0.95	0.44
Total number of cases	38	31	25	279	26	69
Outliers outside 1.5 stds	5 / 13.16%	4 / 12.90%	3 / 12.00%	36 / 12.90%	4 / 15.38%	8 / 11.59%
Lower bound z=-1.5	-0.04	-0.04	-0.01	-0.03	-0.07	-0.04
Upper bound z=1.5	0.04	0.04	0.03	0.02	0.06	0.04
GESD outliers (a=0.45)	0-25 / 65.79%	0-1 / 3.23%	0-1 / 4.00%	0-270 / 96.77%	0-18 / 69.23%	0-1 / 1.45%
GESD lower bound	0.00	-0.09	-0.02	0.00	0.01	-0.10
GESD upper bound	0.01	0.04	0.04	0.00	0.02	0.07

Algorithm 2: Non-parametric bootstrapping

Data: Population $S = \{S_1, S_2, \ldots, S_N\}$ with unknown PDF.
Result: Statistic x's CI
Initial ize list: $CI = [0]_{1\times 2}$;
Initialize list: $X = [0]_{1\times M}$;
for $i = 1, \ldots, M >> 1$ **do**
    $S_i \leftarrow C_S(N, N)$;
    $X(i) \leftarrow x_{S_i}$;
end
$CI(1) \leftarrow P_X^{2.5\%}$;
$CI(2) \leftarrow P_X^{97.5\%}$;

and the other one in the form of bootstrapped CIs.

To better understand the utility of $P(e)$ for building an uncertainty model, the simple idea behind these models is presented here, although it is further discussed in section 1.7.

In Table 1.11, some quantiles of the error distribution are presented. For reference, they are computed as empirical statistics of the empirical error distribution, and using bootstrapping (in this case their confidence intervals are given instead). The simplest uncertainty model which can be built with this information assumes that, for future samples of the residue, the quantiles of Table 1.11 will still hold true (*i.e.* , the empirical and the theoretical quantiles coincide). For instance, we would assume that, according to Table 1.11, for future samples the error of variable "RF Forced Crippling" will belong to the interval $[-0.0432, 0.0448]$ (defined by percentiles 5^{th} and 95^{th}) with a frequency equal to 90%. If we wanted to soften the assumption that the empirical and the theoretical quantiles are the same, we could use the bootstrapped quantiles instead. That way, with a 95% confidence we could claim that the error of "RF Forced Crippling" variable will belong to $[-0.0548., 0.0586]$ with a frequency of 90% as a minimum, given that we now from bootstrapping that, with a 95% confidence, the true 5^{th} quantile belongs to the range $[-0.0548, -0.0353]$ and the true 95^{th} quantile belongs to $[0.0379, 0.0586]$.

Of course, the uncertainty model described in the last paragraph can be fine-tuned using additional information about the error distribution. If we found $P(e)$ to be heteroscedastic, we could benefit from conditioning our uncertainty model to certain regions of the input (or the output) space. This is the main motivation for section 1.5, section 1.6 and section 1.7.

Table 1.10: Summary of bootstrapped error statistics. For the median, the Wilson-score[25] is used for computing the confidence interval.

Confidence level at 95%. BS=Confidence Interval (percentile bootstrap), WS=CI (wilson-score)

	count	min	mean	(BS) mean	median	(WS) median	std	(BS) std	IQR	(BS) IQR	kurtosis	(BS) kurtosis	skewness	(BS) skewness	max
RF Forced Crippling	383	-0.15	0.0032	0.0062 -0.00071	0.0041	0.0068 0.000079	0.034	0.039 0.028	0.030	0.035 0.025	14.	24. 2.7	1.4	3.1 -0.74	0.28
RF Column Buckling	319	-0.12	0.0027	0.0061 -0.00045	0.0014	0.0047 -0.0014	0.031	0.034 0.027	0.030	0.035 0.027	2.7	3.8 1.6	0.22	0.81 -0.33	0.12
RF In Plane	252	-0.08	0.0020	0.0059 -0.0013	0.00065	0.0028 -0.0022	0.027	0.031 0.023	0.027	0.030 0.023	5.0	8.3 0.84	1.1	1.9 0.13	0.16
RF Net Tension	2794	-0.18	-0.0026	-0.0019 -0.0034	-0.0028	-0.0022 -0.0033	0.018	0.02 0.017	0.016	0.016 0.015	17.	25. 6.2	0.72	2.0 -0.75	0.21
RF Pure Compression	265	-0.27	0.0019	0.0073 -0.0028	0.0040	0.0077 0.0012	0.041	0.052 0.032	0.029	0.034 0.025	15.	21. 5.1	-0.31	2.4 -2.9	0.26
RF Shear Panel Failure	695	-0.31	0.0038	0.0081 -0.000057	0.00073	0.0026 -0.0011	0.053	0.067 0.038	0.028	0.031 0.026	44.	63. 20.	4.0	5.9 0.037	0.61

Table 1.11: Bootstrapped percentiles (1^{st}, 5^{th}, 10^{th}, 90^{th}, 95^{th} and 99^{th}) of the residue distribution, calculated with a 95% confidence interval using Wilson-score.

Confidence level at 95%. BS=Confidence Interval (percentile bootstrap), WS=CI (wilson-score)

	1%	(WS) 1%	5%	(WS) 5%	10%	(WS) 10%	median	(WS) median	90%	(WS) 90%	95%	(WS) 95%	99%	(WS) 99%
RF Forced Crippling	-0.0834	-0.0712 -0.115	-0.0432	-0.0353 -0.0548	-0.0286	-0.0253 -0.0361	0.00405	0.00676 0.0000785	0.0309	0.0389 0.027	0.0448	0.0586 0.0379	0.101	0.147 0.0768
RF Column Buckling	-0.0782	-0.0605 -0.111	-0.0441	-0.0337 -0.0509	-0.0283	-0.0246 -0.0373	0.00139	0.00472 -0.00138	0.0362	0.0453 0.0297	0.0582	0.0656 0.0442	0.100	0.113 0.0686
RF In Plane	-0.0586	-0.0449 -0.0798	-0.0386	-0.0306 -0.0449	-0.0284	-0.0203 -0.0337	0.000646	0.00281 -0.00220	0.0310	0.0418 0.0233	0.0481	0.0664 0.0355	0.0747	0.158 0.0664
RF Net Tension	-0.0506	-0.0444 -0.0633	-0.0277	-0.0261 -0.0298	-0.0198	-0.0187 -0.0213	-0.00280	-0.00224 -0.00329	0.0148	0.0158 0.0135	0.0231	0.0250 0.0211	0.0514	0.0596 0.0437
RF Pure Compression	-0.118	-0.0549 -0.267	-0.0428	-0.0352 -0.074	-0.0289	-0.0244 -0.0389	0.00400	0.00772 0.00124	0.0314	0.0407 0.0258	0.0521	0.0715 0.0369	0.109	0.259 0.0683
RF Shear Panel Failure	-0.0994	-0.0775 -0.123	-0.0498	-0.0392 -0.0578	-0.0307	-0.0258 -0.0361	0.000728	0.00259 -0.00108	0.0376	0.0445 0.0299	0.0572	0.0696 0.0489	0.197	0.316 0.0902

1.5 Distributed error quantification: conditioning the error distribution on the input and the output space

Bibliography

[1] P. Bijlaard. "On the Buckling of Stringer Panels Including Forced Crippling". In: *Journal of the Aeronautical Sciences* 22.7 (1955), pp. 491–501 (cit. on p. 1).

[2] F. P. Preparata and M. I. Shamos. "Convex Hulls: Basic Algorithms". In: *Computational Geometry: An Introduction*. New York, NY: Springer New York, 1985, pp. 95–149. ISBN: 978-1-4612-1098-6. DOI: 10.1007/978-1-4612-1098-6_3. URL: https://doi.org/10.1007/978-1-4612-1098-6_3 (cit. on p. 2).

[3] D. Barrett et al. "Measuring abstract reasoning in neural networks". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 511–520. URL: https://proceedings.mlr.press/v80/barrett18a.html (cit. on p. 3).

[4] B. M. Lake and M. Baroni. "Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks". In: *CoRR* abs/1711.00350 (2017). arXiv: 1711.00350. URL: http://arxiv.org/abs/1711.00350 (cit. on p. 3).

[5] D. Saxton et al. "Analysing Mathematical Reasoning Abilities of Neural Models". In: *CoRR* abs/1904.01557 (2019). arXiv: 1904.01557. URL: http://arxiv.org/abs/1904.01557 (cit. on p. 3).

[6] T. Ebert, J. Belz, and O. Nelles. "Interpolation and extrapolation: Comparison of definitions and survey of algorithms for convex and concave hulls". In: *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE. 2014, pp. 310–314 (cit. on p. 3).

[7] W.-Y. Loh, C.-W. Chen, and W. Zheng. "Extrapolation errors in linear model trees". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.2 (2007), 6–es (cit. on p. 3).

[8] P. Klesk. "Construction of a Neurofuzzy Network Capable of Extrapolating (and Interpolating) With Respect to the Convex Hull of a Set of Input Samples in R". In: *IEEE Transactions on Fuzzy Systems* 16.5 (2008), pp. 1161–1179. DOI: 10.1109/TFUZZ.2008.924337 (cit. on p. 3).

[9] R. Balestriero, J. Pesenti, and Y. LeCun. "Learning in high dimension always amounts to extrapolation". In: *arXiv preprint arXiv:2110.09485* (2021) (cit. on pp. 3, 4, 10).

[10] S. Marsland. *Machine Learning: An Algorithmic Perspective*. 2nd ed. Boca Raton, USA: Chapman & Hall/CRC, 2015 (cit. on pp. 3, 7).

[11] I. Bárány and Z. Füredi. "On the shape of the convex hull of random points". In: *Probability theory and related fields* 77 (1988), pp. 231–240 (cit. on p. 3).

[12] L. Bonnasse-Gahot. "Interpolation, extrapolation, and local generalization in common neural networks". In: *arXiv preprint arXiv:2207.08648* (2022) (cit. on pp. 4, 10, 13).

[13] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417 (cit. on p. 9).

[14] B. Rosner and D. Grove. "Use of the Mann–Whitney U-test for clustered data". In: *Statistics in medicine* 18.11 (1999), pp. 1387–1400 (cit. on p. 13).

[15] R. Velez Ibarrola and A. Garcia Perez. *Calculo de probabilidades y Estadistica Matematica*. 1st ed. Madrid, Spain: Universidad Nacional de Educacion a Distancia, 1994 (cit. on p. 15).

[16] D. Zhang. "A coefficient of determination for generalized linear models". In: *The American Statistician* 71.4 (2017), pp. 310–316 (cit. on p. 20).

[17] B. Efron. "Bootstrap methods: another look at the jackknife". In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 569–593 (cit. on p. 28).

[18] G. Chen, J. R. Gott, and B. Ratra. "Non-Gaussian Error Distribution of Hubble Constant Measurements". In: *Publications of the Astronomical Society of the Pacific* 115.813 (2003), p. 1269 (cit. on p. 22).

[19] P. Pernot, B. Huang, and A. Savin. "Impact of non-normal error distributions on the benchmarking and ranking of Quantum Machine Learning models". In: *Machine Learning: Science and Technology* 1.3 (2020), p. 035011 (cit. on p. 22).

[20] D. Smyl et al. "Learning and correcting non-Gaussian model errors". In: *Journal of Computational Physics* 432 (2021), p. 110152 (cit. on p. 22).

[21] L. Chai et al. "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12 (2019), pp. 1919–1931 (cit. on p. 22).

[22] J. D. Jobson. *Applied multivariate data analysis: regression and experimental design*. Springer Science & Business Media, 2012 (cit. on p. 20).

[23] M. C. Jones and A. Pewsey. "Sinh-arcsinh distributions". In: *Biometrika* 96.4 (2009), pp. 761–780 (cit. on pp. 25, 29).

[24] B. Rosner. "Percentage points for a generalized ESD many-outlier procedure". In: *Technometrics* 25.2 (1983), pp. 165–172 (cit. on p. 27).

[25] E. B. Wilson. "Probable inference, the law of succession, and statistical inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212 (cit. on p. 31).

[26] O. Montenbruck and E. Gill. *Satellite Orbits: Models, Methods and Applications*. 1st ed. Wessling, Germany: Springer, 2001 (cit. on p. 38).

[27] K. Yamanaka and F. Ankersen. "New State Transition Matrix for Relative Motion on an Arbitrary Elliptical Orbit". In: *Journal of Guidance, Control & Dynamics* 25.1 (2002), pp. 60–66. DOI: 10.2514/2.4875 (cit. on p. 39).

[28] D.-W. Gim and K. Alfriend. "State Transition Matrix of Relative Motion for the Perturbed Noncircular Reference Orbit". In: *Journal of Guidance Control and Dynamics* 26 (Nov. 2003), pp. 956–971. DOI: 10.2514/2.6924 (cit. on p. 39).

[29] S.-S. Exchange. *Practical Uses of an STM*. 2019. URL: https://space.stackexchange.com/questions/32916 (cit. on p. 40).

[30] ESA. *Precise Orbit Determination*. 2011. URL: https://gssc.esa.int/navipedia/index.php/Precise_Orbit_Determination (cit. on p. 40).

[31] C. Chao and F. Hoots. *Applied Orbit Perturbation and Maintenance*. Aerospace Press, 2018. ISBN: 9781523123346 (cit. on p. 45).

[32] W. E. Wiesel. *Modern Astrodynamics*. 2nd ed. Beavercreek, Ohio: Aphelion Press, 2010 (cit. on p. 49).

[33] M. Eckstein, C. Rajasingh, and P. Blumer. "Colocation Strategy and Collision Avoidance for the Geostationary Satellites at 19 Degrees West". In: *CNES International Symposium on Space Dynamics*. Vol. 25. Oberpfaffenhofen, Germany: DLR GSOC, Nov. 1989, pp. 60–66 (cit. on p. 50).

[34] S. D'Amico and O. Montenbruck. "Proximity Operations of Formation-Flying Spacecraft Using an Eccentricity/Inclination Vector Separation". In: *Journal of Guidance, Control & Dynamics* 29.3 (2006), pp. 554–563. DOI: 10.2514/1.15114 (cit. on pp. 50, 56, 57).

[35] D.-W. Gim and K. T. Alfriend. "Satellite Relative Motion Using Differential Equinoctial Elements". In: *Celestial Mechanics and Dynamical Astronomy* 92.4 (2005), pp. 295–336. DOI: 10.1007/s10569-004-1799-0 (cit. on pp. 51, 52).

[36] S. D'Amico. *Relative Orbital Elements as Integration Constants of Hill's Equations*. TN 05-08. Oberpfaffenhofen, Germany: Deutsches Zentrum für Luft- und Raumfahrt (DLR), 2005 (cit. on pp. 51, 56).

[37] H. Schaub. "Relative Orbit Geometry Through Classical Orbit Element Differences". In: *Journal of Guidance, Control & Dynamics* 27.5 (2004), pp. 839–848. DOI: 10.2514/1.12595 (cit. on pp. 51, 56).

[38] G. Gaias, C. Colombo, and M. Lara. "Accurate Osculating/Mean Orbital Elements Conversions for Spaceborne Formation Flying". In: (Feb. 2018). https://www.researchgate.net/publication/340378956_Accurate_OsculatingMean_Orbital_Elements_Conversions_for_Spaceborne_Formation_Flying (cit. on p. 52).

[39] T. Vincent Peters and R. Noomen. "Linear Cotangential Transfers and Safe Orbits for Elliptic Orbit Rendezvous". In: *Journal of Guidance, Control & Dynamics* 44.4 (2021), pp. 732–748. DOI: 10.2514/1.G005152 (cit. on pp. 58, 59).

[40] R. H. Lopes, I. Reid, and P. R. Hobson. "The two-dimensional Kolmogorov-Smirnov test". In: (2007).

[41] G. Schoups and J. A. Vrugt. "A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors". In: *Water Resources Research* 46.10 (2010).

[42] B. D. Tapley, B. E. Schutz, and G. H. Born. *Statistical Orbit Determination*. 1st ed. Amsterdam, Netherlands: Elsevier, 2004.

[43] K. T. Alfriend and Srinivas. *Spacecraft Formation Flying*. 1st ed. Oxford, United Kingdom: Elsevier, 2010.

[44] R. H. Battin. *An Introduction to the Mathematics and Methods of Astrodynamics, Revised Edition*. 1st ed. Reston, Virginia: American Institute of Aeronautics and Astronautics, 1999.

[45] D. Brouwer and G. M-Clemence. *Methods of Celestial Mechanics*. Brackley Square House, London: Academic Press, 1961.

[46] H. Schaub and J. L. Junkins. *Analytical Mechanics of Space Systems*. Reston, VA: AIAA Education Series, Oct. 2003. DOI: 10.2514/4.861550.

[47] W. Fehse. *Automated Rendezvous and Docking of Spacecraft*. Cambridge Aerospace Series. Cambridge: Cambridge University Press, 2003. DOI: 10.1017/CBO9780511543388.

[48] P. K. S. Dennis D. McCarthy. *Time: From Earth Rotation to Atomic Physics*. Wiley-VCH, 2009. ISBN: 3527407804; 9783527407804.

[49] K. Wakker. *Fundamentals of Astrodynamics*. Jan. 2015. ISBN: 978-94-6186-419-2.

[50] A. H. Nayfeh. *Perturbation Methods*. Weinheim, Germany: Wiley-VCH, 2004.

[51] W. M. Kaula. *Theory of Satellite Geodesy*. Mineola, New York: Dover Publications, 2013.

[52] F. Tisserand. *Traité de Mécanique Céleste*. Vol. 1. Paris, 1889.

[53] J. Sullivan, S. Grimberg, and S. D'Amico. "Comprehensive Survey and Assessment of Spacecraft Relative Motion Dynamics Models". In: *Journal of Guidance, Control & Dynamics* 40.8 (2017), pp. 1837–1859. DOI: 10.2514/1.G002309.

[54] H. Schaub and K. T. Alfriedn. "Hybrid Cartesian and Orbit Element Feedback Law for Formation Flying Spacecraft". In: *Journal of Guidance, Control & Dynamics* 25.2 (2002), pp. 387–393. DOI: 10.2514/2.4893.

[55] N. Capitaine. "The Celestial Pole Coordinates". In: *Celestial Mechanics and Dynamical Astronomy* 48 (1990), pp. 127–143.

[56] D. D. McCarthy. *IERS Conventions (1992)*. 21. Paris, France: Central Bureau of IERS - Observatoire de Paris, 1996.

[57] J. Williams. *LVLH Transformations*. https://degenerateconic.com/uploads/2015/03/lvlh.pdf. 2014.

[58] H. Fiedler. *Analysis of TerraSAR-L Cartwheel Constellations*. https://elib.dlr.de/22345/. Oberpfaffenhofen, Germany: Deutsches Zentrum für Luft- und Raumfahrt (DLR), Nov. 2003.

[59] G. W. Hill. "Researches in the Lunar Theory". In: *American Journal of Mathematics* 1.2 (1878), pp. 129–147. DOI: 10.2307/2369304.

[60] W. H. Clohessy and R. S. Wiltshire. "Terminal Guidance System for Satellite Rendezvous". In: *Journal of the Aerospace Sciences* 27.9 (1960), pp. 653–658. DOI: 10.2514/8.8704.

[61] R. Broucke. "On the Matrizant of the Two-Body Problem". In: *Astronomy and Astrophysics* 6 (June 1970), p. 173.

[62] J. Tschauner and P. Hempel. "Optimale Beschleunigungsprogramme fÃ¼r das Rendezvous-Manöver". In: *Astronautica Acta* 10 (1964), pp. 296–307.

[63] T. Carter. "State Transition Matrices for Terminal Rendezvous Studies: Brief Survey and New Example". In: *Journal of Guidance Control and Dynamics* 21 (Jan. 1998), pp. 148–155. DOI: 10.2514/2.4211.

[64] O. Montenbruck, M. Kirschner, and S. D'Amico. "E/I-vector separation for safe switching of the GRACE formation". In: *Aerospace Science and Technology* 10 (2006), pp. 628–635. DOI: 10.1016/j.ast.2006.04.001.

[65] G. Gaias, J.-S. Ardaens, and C. Colombo. "Precise line-of-sight modelling for angles-only relative navigation". In: *Advances in Space Research* 67.11 (2021). Satellite Constellations and Formation Flying, pp. 3515–3526. DOI: 10.1016/j.asr.2020.05.048.

[66] G. Gaias, C. Colombo, and M. Lara. "Analytical Framework for Precise Relative Motion in Low Earth Orbits". In: *Journal of Guidance, Control, and Dynamics* 43.5 (2020), pp. 915–927. DOI: 10.2514/1.G004716.

[67] D. Brouwer. "Solution of the problem of artificial satellite theory without drag". In: *Astronomical Journal* 64.5 (Nov. 1959), p. 378. DOI: 10.1086/107958.

[68] R. H. Lyddane. "Small eccentricities or inclinations in the Brouwer theory of the artificial satellite". In: *Astronomical Journal* 68 (Oct. 1963), p. 555. DOI: 10.1086/109179.

[69] A. Deprit. "Canonical transformations depending on a small parameter". In: *Celestial Mechanics* 1.1 (Mar. 1969), pp. 12–30. DOI: 10.1007/BF01230629.

[70] A. Deprit. "Delaunay Normalisations". In: *Celestial Mechanics* 26.1 (Jan. 1982), pp. 9–21. DOI: 10.1007/BF01233178.

[71] G. Hori. "Theory of General Perturbation with Unspecified Canonical Variable". In: *Publications of the Astronomical Society of Japan* 18 (Jan. 1966), pp. 287–296.

[72] Y. Chihabi and S. Ulrich. "Spacecraft Formation Guidance Law using a State Transition Matrix With Gravitational, Drag and Third-Body Perturbations". In: Jan. 2020. DOI: 10.2514/6.2020-1460.

[73] G. Gaias, J. Ardaens, and O. Montenbruck. "Model of J2 Perturbed Satellite Relative Motion with Time-Varying Differential Drag". In: *Celestial Mechanics and Dynamical Astronomy* (2015).

[74] A. Koenig, T. Guffanti, and S. D'Amico. "New State Transition Matrices for Relative Motion of Spacecraft Formations in Perturbed Orbits". In: Sept. 2016. DOI: 10.2514/6.2016-5635.

[75] A. Biria and R. Russell. "A Satellite Relative Motion Model Including J2 and J3 via Vinti's Intermediary". In: Feb. 2016.

[76] K. Alfriend and H. Yan. "An Orbital Elements Based Approach to the Nonlinear Formation Flying Problem". In: Toulouse, France, Feb. 2016.

[77] Mathworks. *Convert complex diagonal form into real diagonal form*. 2022. URL: https://www.mathworks.com/help/matlab/ref/cdf2rdf.html.

[78] M. R. Delgado. *Lecture notes: Basics of Orbital Mechanics I*. Apr. 2008.

[79] F. G. Nievinski. *subtightplot*. https://www.mathworks.com/matlabcentral/fileexchange/39664-subtightplot. 2013.

[80] J. C. Lansey. *linspecer*. https://www.mathworks.com/matlabcentral/fileexchange/42673-beautiful-and-distinguishable-line-colors-colormap. 2015.

[81] Jan. *WindowAPI*. https://www.mathworks.com/matlabcentral/fileexchange/31437-windowapi. 2013.

[82] T. Davis. *Arrow3*. https://www.mathworks.com/matlabcentral/fileexchange/14056-arrow3. 2022.

[83] E. Duenisch. *latexTable*. https://www.mathworks.com/matlabcentral/fileexchange/44274-latextable. 2016.