Sensitivity of the SIMulation-EXtrapolation (SIMEX) Methodology to Mis-specification of the Statistical Properties of the Measurement Errors

GABRIELE VILLARINI[1], DARIO TREPPIEDI[2], LEONARDO V. NOTO[2]

[1]IIHR—Hydroscience and Engineering, University of Iowa, Iowa City, USA

[2]Dipartimento di Ingegneria, Università degli Studi di Palermo, Palermo, Italy

Manuscript submitted to

Theoretical and Applied Climatology

2 September 2022

Revised February 2023

Corresponding author address:

Gabriele Villarini, IIHR-Hydroscience & Engineering, The University of Iowa, 323B C. Maxwell Stanley Hydraulics Laboratory, Iowa City, 52242, Iowa, USA. E-mail: gabriele-villarini@uiowa.edu. Tel.: (319) 384-0596

ABSTRACT

In hydrometeorological and environmental studies, it is common to seek relations between two variables (predictand and predictor), one of which (predictor) is affected by uncertainties. These errors unavoidably affect the results of the analyses by providing erroneous estimates of the parameters of the predictor-predictand model. A possible solution is represented by the SIMulation-EXtrapolation (SIMEX) methodology. This approach follows two steps: 1) perturbation of the predictor with increasing level of uncertainties (multiples of the known error variance); and 2) finding a relation between the model's parameters and level of uncertainty, which allows their extrapolation to the error-free case.

The application of the SIMEX methodology requires the a priori knowledge of the mean, variance, and distribution of the measurement errors. However, in hydrologic and climatologic studies, this is not the case and the impact of an erroneous specification of these statistical properties on the results of the analyses has received little attention. The aim of this study is to investigate the sensitivity of the SIMEX methodology to mis-specification of the error characteristics. By using a simulation-based approach, we investigate the impact of an imperfect knowledge of the characteristics of the errors associated with the predictor (mean, variance, and probability distribution). Our results suggest that SIMEX is robust against mis-specification of the moments and distribution of the measurement errors, that it performs better than standard linear regression, even when these statistical properties are erroneously specified and that, for these reasons, it could find an useful application to seasonal forecasting of hydroclimatic variables.

Keywords: Simulation-extrapolation; SIMEX; Uncertainties; Measurement Errors

1. Introduction

In several disciplines, we are interested in finding the relation between two random variables, denoted by X and Y, where the former is the predictor and the latter the predictand. In general, while we usually assume that Y is error-free, this may not be the case for X: instead of measuring X (latent variable), we actually measure another variable, W, which represents the combination of X and the measurement errors associated with it. In hydrometeorological and environmental studies, this is a very common problem. For instance, consider the case in which we want to relate radar rainfall estimates to the average of different rain gage measurements within the pixel of interest (considered to be the "ground truth"). When we relate these two quantities, we generally neglect to account for the uncertainties associated with computing pixel-averaged rainfall based on a limited number of rain gages (e.g., Villarini et al. 2008). As mentioned in Hasan et al. (2014), by neglecting these uncertainties, we introduce a bias when converting radar reflectivity to rainfall. Another example is related to the impacts of uncertainties in input variables in the estimation of the parameters of hydrologic models (e.g., Chowdhury and Sharma 2008). Moreover, consider for instance a time series of precipitation from a global climate model (GCM) that we want to fit with a gamma distribution. It is well-known that there are large uncertainties associated with GCM outputs, especially for a variable like precipitation. If we neglect to account for these uncertainties and apply standard approaches, our estimation of the parameters of the gamma distribution will be biased (Woldemeskel et al. 2014).

More specifically, let us focus on the case in which X and Y are linearly related:

$$\mathbf{Y}=\beta_0+\beta_1\mathbf{X}+\varepsilon \quad (1)$$

where ε represents the uncertainties with respect to the regression line and is normally distributed with mean equal to zero and standard deviation equal to σ_ε.

Assuming a classic error model (Carroll et al. 2007), instead of measuring X we measure W, which accounts for the measurement errors U in X in an additive form:

$$W=X+U \quad (2)$$

where U is independent of both X and Y and is described by a Gaussian distribution with mean μ_U equal to zero and standard deviation σ_U.

Considering Y as the predictand and W as the predictor, ordinary least squares method (OLS) provides a consistent estimation of $\hat{\beta}_{1,\text{naïve}} = \lambda\beta_1$ rather than β_1 (e.g., Carroll et al. 2007) where:

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_U^2} < 1 \quad (3)$$

and σ_x^2 represents the variance of the variable X, and λ is called the reliability ratio (Fuller 1987). The estimate of the slope by means of OLS is therefore biased when $\sigma_U^2 > 0$. In general, the presence of measurement errors not only introduces an attenuation of the slope β_1, but makes the data noisier as well (e.g., Carroll et al. 2007).

Since it is clear how measurement errors may significantly affect the results of our analyses, they should be accounted for. In the field of applied statistics, this is a well-known problem and it has been the object of a vast literature (e.g., Fuller 1987; Gleser 1990; Brown and Mariano 1993; Carroll et al. 2007). Among all of the proposed approaches, a very effective method to reduce the bias associated with the presence of measurement errors is the SIMulation-EXtrapolation (SIMEX) method, first introduced by Cook and Stefanski (1994). SIMEX is based on the idea of adding measurement errors to the data as multiples of the known error variance and finding a relation between the targeted parameter

and the level of added noise, allowing the extrapolation of the results to the error-free case. We will discuss this approach in more details in the next section. Among other fields, this simulation-based method has been widely used in ecology (e.g., Kangas 1998 , Stoklosa et al. 2015; Ponzi et al. 2019; Kinane et al. 2021), and biostatistics and medicine (e.g., Lauzon et al. 2013; Guolo 2014; Alexeeff et al. 2016; Oh et al. 2018). However, in hydrometeorological and environmental studies, it has found limited application. The first application of this methodology was discussed by Chowdhury and Sharma (2007) to infill values of the Southern Oscillation Index based on sea surface temperature anomalies. Chowdhury and Sharma (2008) used the SIMEX methodology to quantify the bias in key storage parameters in the Sacramento Model. Then Woldemeskel et al (2012) showed the improvement in the estimation of the parameters of future droughts when SIMEX was used to account for uncertainties in the GCM outputs. Finally, Hasan et al. (2014) used this method when relating radar reflectivity to the ground truth, accounting for the uncertainties associated with a "ground truth" obtained from the average of multiple point measurements.

One of the possible obstacles towards a more widespread use of this methodology is the required knowledge of the statistical characteristics (i.e., mean, variance, and probability distribution) of the measurement error U. For the case of unknown variance, Devanarayan and Stefanski (2002) proposed a modification of the SIMEX approach, which however requires independent replicate measurements. However, as described in Chowdhury and Sharma (2008), an outstanding "issue to be investigated in greater detail is the specification of the error distribution for various hydro-climatological variables." Therefore, what would happen if we mis-specify the error variance, or the mean, or even

the entire probability distribution still remains open questions. Carroll et al. (2007) wrote that "minor violations to the assumption of normality of the measurement errors is not critical in practice." To the best of our knowledge, this is one of the very few indications about the robustness of SIMEX to mis-specification of the statistical characteristics of U. Therefore, a study in which this issue is addressed is still lacking and it would be important to show that even an imperfect knowledge of the measurement error characteristics could lead to more accurate results compared to the case in which we neglect them. In this study we tackle this question in a simulation framework, where we know the underlying relation between X and Y (Section 2). In Section 3 we will show that SIMEX tends to perform better than OLS even under mis-specification of the error characteristics, while Section 4 discusses the main points made and closes the article.

2. SIMulation-EXtrapolation: SIMEX

There are several papers describing the SIMEX methodology in details for linear and non-linear models, homoschedastic or heteroschedastic errors, both in additive or multiplicative forms. The reader is pointed to Carroll et al. (2007) and references therein for more information. Here we consider the linear model in equation (1), and the additive error model in equation (2) to describe the measurement errors in X. We also assume that we know that U is normally distributed with mean equal to zero and variance equal to σ_U^2.

The SIMEX methodology consists of a simulation and an extrapolation step. In the simulation step, we generate m-1 additional datasets which are obtained by increasing the level of measurement error by $(1+\zeta)\sigma_U^2$, with $\zeta\epsilon[0;\ \zeta_m]$ and known. By using OLS, we would consistently estimate $\beta_1\sigma_x^2/[\sigma_x^2+(1+\zeta)\sigma_U^2]$. We can now focus our attention to a

nonlinear regression problem, in which the independent variable is ζ and the dependent one is $\hat{\beta}_1$. Asymptotically, we have:

$$\mathrm{E}\left(\hat{\beta}_1|\zeta\right) = \mathrm{G}(\zeta) = \frac{\beta_1\sigma_x^2}{\sigma_x^2+(1+\zeta)\sigma_U^2} \quad (4)$$

where $G(0) = \hat{\beta}_{1,naïve}$ (i.e., the slope estimated by OLS) while $\mathrm{G}(-1) = \hat{\beta}_{1,\mathrm{SIMEX}}$. Here $\hat{\beta}$ represents the estimated value of the β coefficient.

We can summarize this methodology in the following steps:

1- We consider m to be equal to 7, with ζ that set to assume the values [0.5; 1.0; 1.5; 2.0; 2.5; 3.0]. Simulate independent errors with variance equal to $\zeta\sigma_U^2$ and add it to the measured variable W. For each ζ, we repeat this step 200 times (we have explored the sensitivity of the results to the number of replicates, without finding them to be sensitive).

2- Estimate the slope β_1 for each of the new simulated datasets. For each value of ζ, we have 200 estimates of the slope; for each one of them, take the average value. Therefore, we have six points (plus the naïve estimate).

3- Plot β_1 as a function of ζ and fit a quadratic function (see Section 5.3.2 in Carroll et al. (2007) for considerations about the extrapolant function). Once parameterized G(ζ), we want to obtain the value of the function for ζ = -1, so that G(-1) is the targeted $\beta_{1,\mathrm{SIMEX}}$.

It is clear that we need to have information about the statistical properties of the measurement errors. As mentioned before, we analyze the impacts of mis-specification of the statistical properties of the measurement errors on the SIMEX methodology in a simulation framework. Similar to Chowdhury and Sharma (2007), we start by simulating

X from a standard uniform distribution. To generated Y, we use the linear model in equation (1), and consider four possible sets of values for β_0 and β_1:

1. $\beta_0 = 0$ and $\beta_1 = 0.3$;
2. $\beta_0 = 0$ and $\beta_1 = 0.8$;
3. $\beta_0 = 0$ and $\beta_1 = 2.0$;
4. $\beta_0 = 0.8$ and $\beta_1 = 0.8$.

These values could be representative of changes in hydroclimatic variables of annual to decadal to centennial time scales, including precipitation, temperature, and discharge. Moreover, we consider ε to have a normal distribution with mean equal to zero and standard deviation equal to 0.1. As far as the measurement error U is concerned, we assume that it is normally distributed with mean equal to zero and $\sigma_u = 0.3$. We perform our analysis using the SIMEX package (Lederer and Seibold 2019) in R (R Core Team 2022).

For case 2 ($\beta_0 = 0$ and $\beta_1 = 0.8$), we show how SIMEX can correct for the bias in the regression coefficient in Figure 1. By neglecting the measurement uncertainties, OLS would return a slope value $\hat{\beta}_{1,\text{naïve}}$ equal to 0.64. On the other hand, using the SIMEX approach we have that $\hat{\beta}_{1,\text{SIMEX}}$ is equal to 0.78. Therefore, given an additive error model and knowing the measurement error mean, variance, and distribution, SIMEX can remove the bias from the estimation of the parameters. Similar improvements were observed for the other three scenarios as well: the SIMEX methodology was able to provide an estimate of the slope and intercept that was close to the target values, much more so than by using OLS.

We apply SIMEX to seasonal forecasting to explore its potential applicability to hydroclimatological variables. Figure 2 shows an example related to the prediction of the

Southern Oscillation Index (SOI) in February using the sea surface temperature (SST) averaged over the Niño3.4 region and forecasted from the beginning of February (i.e., 0.5-lead time forecast) to the beginning of November of the previous year (i.e., 3.5-lead time forecast). Results are based on the average of 12 members of the GFDL-CM2p5-FLOR-B01 by the Geophysical Fluid Dynamics Laboratory and part of the North American Multi-Model Ensemble (NMME; Kirtman et al. 2014). The study period ranges from 1981 to 2020, and we use 1981-2204 for calibration, and 2005-2020 for validation. As a first step, we compute σ_u from the 12 members of the GFDL model, and we then apply the SIMEX methodology to estimate the slope and intercept. For short lead times, the regression lines between SOI and Niño3.4 SST based on SIMEX and OLS are almost identical. As the lead time increases, there is more separation between the two lines, with a slightly better performance by SIMEX in terms of root mean squared error. These results point to the potential suitability of SIMEX for seasonal forecasting, especially for long lead times.

3. Results

Let us start by investigating the impact of mis-specification of the mean and variance of the measurement errors under the above mentioned four scenarios. To accomplish this task, we apply the SIMEX methodology assuming that U is normally distributed with mean equal to zero and standard deviation equal to 0.3, and investigate what happens if we increase the value of μ_U from 0 to 1, and at the same time we have σ_U ranging from 0.3 to 0.6.

In Figures 3-6, we have plotted the results for our analysis. On the x-axis we have increasing values of μ_U from 0 to 1, while on the y-axis we have plotted the uncertainties

associated with σ_U in percentage of its true value. As far as the intercept is concerned (left panels), we show $\Delta_{\beta_0,i} = (\beta_0 - \hat{\beta}_{0,i})/\beta_1$, where $\hat{\beta}_{0,i}$ is the estimated intercept using OLS or SIMEX (indexed by i), β_0 is its true value, and β_1 is value of the true slope. As far as the slope is concerned (right panels), we show $\Delta_{\beta_1,i} = 100(\beta_1 - \hat{\beta}_{1,i})/\beta_1$, where $\hat{\beta}_{1,i}$ is the value of the slope estimated from the data using OLS or SIMEX (indexed by i). Here we normalize by the differences in the intercept with respect to the true slope rather than the true intercept because the latter is set to zero in some of our simulations, affecting the computation of the ratio. More generally, the normalization of the results with respect to β_1 provides indications on the generalization of these results.

As expected, the slope is sensitive only to mis-specifications of the standard deviation of the measurement errors (Figures 3-6). This statement is valid for both OLS and SIMEX methodologies. We have increasing errors for increasing uncertainties in σ_U, since we only partially account for the measurement errors: based on equation (3), the reliability ratio λ is still smaller than 1, resulting in an attenuated slope. Overall, with OLS we make an error between 20% and 50% of the true slope value for uncertainties in σ_U from 0% to 100%. On the other hand, using SIMEX we have errors in the estimation of the slope ranging from 5% to 30%. Based on our results, the improvements from the use of the SIMEX methodology with respect to OLS are larger as the uncertainties in σ_U are smaller. On the other hand, as the uncertainties in the measurement error standard deviation increase, the advantage of using SIMEX with respect to OLS decreases. This is made clearer from the bottom-right panels in Figures 3-6, in which the ratio between the OLS and SIMEX errors $(\Delta_{\beta_1,\text{naïve}}/\Delta_{\beta_1,\text{SIMEX}})$ decreases from a value greater than 10 for small errors in σ_U to values closer to 1 (i.e., they perform in a similar way) for large uncertainties in the

measurement error. Comparing the results for scenarios 1 to 3 (Figures 3-5), we notice a slightly reduced sensitivity of both OLS and SIMEX to mis-specification of σ_U. Comparing Figures 4 and 6, when estimating the slope the presence of a larger intercept tends to attenuate the sensitivity of both of the approaches to mis-specifications.

While the slope is sensitive exclusively to erroneous specification of the measurement error standard deviation, the intercept is mostly sensitive to mis-specification of its mean μ_U. This feature was expected, because errors with a mean different from zero would introduce an offset that would be compensated by creating or shifting an intercept. For small values of μ_U, the intercept is not sensitive to the error standard deviation. On the other hand, as μ_U increases, the sensitivity of the intercept to increasing uncertainties in σ_U increases: the attenuation in the estimation of the slope tends to reduce the impact of the increasing values of μ_U. Based on Figures 3-6 (bottom-left panel) where we show $\Delta_{\beta_0,\text{naïve}}/\Delta_{\beta_0,\text{SIMEX}}$, the intercept estimated with SIMEX tends to be very close to what estimated from OLS.

Finally, we have looked at the impact of mis-specification of the measurement error distribution and the role of sample size (Figure 7). Throughout this study, we have assumed to know the error distribution, and, in particular, that it can be described by a Gaussian distribution. However, in many hydrometeorological applications we do not have this type of information. Therefore, we want to consider the case in which we erroneously assume that the errors are normally distributed, even though they are actually described by another distribution. In particular, we consider U to follow Laplace, logistic, lognormal, and gamma distributions (e.g., Johnson et al. 1994). The Laplace and logistic distributions have a mean and standard deviation equal to 0 and 0.3, respectively. As far as the lognormal

distribution is concerned, these values are also the values of the mean and standard deviation of the corresponding Gaussian distribution. Finally, the gamma distribution has a mean of 0.5 and a standard deviation of 0.3. To focus on the impact of higher moment orders (e.g., skewness, kurtosis), we have linearly transformed the gamma and lognormally distributed errors to match the values 0.0 and 0.3 for mean and standard deviation, preserving the shape of their distributions. We consider six sample sizes (i.e., n = 25, 50, 100, 200, 400, and 800), and present our results in Figure 7, where the true values of intercept and slope are $\beta_0 = 0$ and $\beta_1 = 0.8$.

Regardless of the distribution and the sample size, the SIMEX method outperforms OLS in estimating β_1. If we use the median of the $\hat{\beta}_1$ by SIMEX and OLS as reference, the gap between the two approaches tends to remain constant, with OLS estimating a slope value around 0.6, while SIMEX is closer to the target value of 0.8. As we increase the sample size, the variability associated with the estimation of the slope decreases, especially as we go from 25 to 100. For sample sizes larger than 100, the marginal improvement in terms of performance decreases. Neither of the approaches displays a strong dependence on the distribution of U. Not surprisingly, when the measurement error follows a Gaussian distribution, SIMEX performs well. However, even when U follows a Laplace, gamma or logistic distribution, the sensitivity of the results is very limited and indeed almost indistinguishable from what observed for the Gaussian case. The largest departure is for the lognormal distribution, whose mis-specification has a detectable signal for both SIMEX and OLS. Therefore, based on these results and in agreement with Carroll et al. (2007), the SIMEX methodology is robust to violations to the assumption of Gaussian distribution of the measurement errors.

4. Discussion and Conclusions

In studies investigating the existence of a relation between two variables, the presence of errors in a predictor could significantly affect the results of the analyses. A possible approach to account for measurement errors is represented by the SIMEX methodology. It is a simple albeit powerful methodology in the case in which we have information about the statistical properties of the measurement errors. However, its sensitivity to an erroneous specification of these errors has received little attention and was the topic of this study. Our findings can be summarized as follows:

1- The slope is sensitive to mis-specifications of the measurement error standard deviation σ_U and insensitive to mis-specification of the measurement error mean μ_U.

2- For low values of the mean μ_U (i.e., measurements that are unbiased or have small biases), the intercept is sensitive only to mis-specifications of the measurement error mean. As the values of μ_U increases, it tends to depend on σ_U as well.

3- Departure of the statistical distribution of the measurement error from a Gaussian distribution was found to be not very sensitive to the case of Laplace, logistic and gamma distribution, while a comparatively worse performance was exhibited by for the case of the lognormal distribution. Therefore, mis-specifications of the distribution of the measurement errors do not affect the results of the SIMEX methodology.

In this study we have focused on linear regression. However, SIMEX has been successfully applied to other and more complex models (e.g., Carroll et al. 2007), and the

sensitivity of SIMEX to error mis-specification in these setups should be evaluated in future studies.

Even under mis-specifications of the statistical properties of the measurement errors, the SIMEX method performs better than the standard OLS. Therefore, our findings suggest that SIMEX could be a very valuable approach in those applications where the measurements of a predictor are affected by errors for which limited information about their statistical characteristics are available, including for sub-seasonal to seasonal forecasting of hydroclimatic variables (e.g., precipitation, temperature, large-scale climate indices) or to account for the role of representativeness errors in rain gauges in ground validation of remote sensing estimates (e.g., Villarini and Krajewski 2008).

Author Declarations

Funding: Gabriele Villarini acknowledges support from the USACE Institute for Water Resources. Leonardo V. Noto acknowledges financial support for Dario Treppiedi provided by Consorzio Interuniversitario per l'Idrologia and by Autorità di bacino del Distretto idrografico della Sicilia.

Conflict of interest/Competing interests: None

Ethics approval/declarations: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material/ Data availability: Not applicable

Code availability: The codes are available on Github (https://github.com/dtreppiedi/simex-mis-specification).

Authors' contributions: GV. conceptualized the study, and D.T. performed the simulations and prepared the figures. All authors interpreted the results and wrote the manuscript.

References

Alexeeff, S. E., R. J. Carroll and B. Coull, Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures, *Biostatistics*, 17(2), 377-389, 2016.

Brown, B. W. and R. S. Mariano. Stochastic simulations for inference in nonlinear errors-in-variables models, *Handbook of Statistics*. New York, North Holland, 11, 611-627, 1993..

Carroll, R. J., D. Ruppert, L. A. Stefanski and C. M. Crainiceanu. *Measurement Error in Nonlinear Models - A Modern Perspective*, CRC Press, 2007..

Chowdhury, S. and A. Sharma, Mitigating parameter bias in hydrological modelling due to uncertainty in covariates, *Journal of Hydrology*, 340(3-4), 197-204, 2007.

Chowdhury, S. and A. Sharma, A simulation based approach for representation of rainfall uncertainty in conceptual rainfall runoff models, *Hydrological Research Letters*, 25-8, 2008.

Cook, J. R. and L. A. Stefanski, Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association*, 89(428), 1314-1328, 1994.

Devanarayan, V. and L. A. Stefanski, Empirical simulation extrapolation for measurement error models with replicate measurements, *Statistics & Probability Letters*, 59(3), 219-225, 2002.

Fuller, W. A. *Measurement Error Models*, New York, John Wiley & Sons, 1987..

Gleser, L. J. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models, *Statistical Analysis of Error Measurement Models and Application*. P. J. Brown and W. A. Fuller, Providence, Rhode Island, American Methematics Society, 1990..

Guolo, A., The SIMEX approach to measurement error correction in meta-analysis with baseline risk as covariate, *Statistics in Medicine*, 33(12), 2062-2076, 2014.

Hasan, M. M., A. Sharma, F. Johnson, G. Mariethoz and A. Seed, Correcting bias in radar Z–R relationships due to uncertainty in point rain gauge networks, *Journal of Hydrology*, 5191668-1676, 2014.

Johnson, N. L., S. Kotz and N. Balakrishnan. *Continuous Univariate Distributions*, New York, John Wiley & Sons, 1994..

Kangas, A. S., Effect of errors-in-variables on coefficients of a growth model and on prediction of growth, *Forest Ecology and Management*, 102(2-3), 203-212, 1998.

Kinane, S. M., C. R. Montes, T. J. Albaugh and D. R. Mishra, A Model to Estimate Leaf Area Index in Loblolly Pine Plantations Using Landsat 5 and 7 Images, *Remote Sensing*, 13(6), 1140, 2021.

Kirtman, B. P., D. Min, J. M. Infanti, J. L. Kinter, D. A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M. P. Mendez, E. Becker, P. T. Peng, P. Tripp, J. Huang, D. G. DeWitt, M. K. Tippett, A. G. Barnston, S. H. Li, A. Rosati, S. D. Schubert, M. Rienecker, M. Suarez, Z. E. Li, J. Marshak, Y. K. Lim, J. Tribbia, K. Pegion, W. J. Merryfield, B. Denis and E. F. Wood, The North American Multimodel Ensemble Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction, *Bulletin of the American Meteorological Society*, 95(4), 585-601, 2014.

Lauzon, C. B., C. Crainiceanu, B. C. Caffo and B. A. Landman, Assessment of bias in experimentally measured diffusion tensor imaging parameters using SIMEX, *Magnetic Resonance in Medicine*, 69(3), 891-902, 2013.

Lederer, W. and H. Seibold, *simex: SIMEX- And MCSIMEX-Algorithm for Measurement Error Models*, 2019..

Oh, E. J., B. E. Shepherd, T. Lumley and P. A. Shaw, Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with SIMEX, *Statistics in Medicine*, 37(8), 1276-1289, 2018.

Ponzi, E., L. F. Keller and S. Muff, The simulation extrapolation technique meets ecology and evolution: A general and intuitive method to account for measurement error, *Methods in Ecology and Evolution*, 10(10), 1734-1748, 2019.

R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2022..

Stoklosa, J., C. Daly, S. D. Foster, M. B. Ashcroft and D. I. Warton, A climate of uncertainty: accounting for error in climate variables for species distribution models, *Methods in Ecology and Evolution*, 6(4), 412-423, 2015.

Villarini, G. and W. F. Krajewski, Empirically-based modeling of spatial sampling uncertainties associated with rainfall measurements by rain gauges, *Advances in Water Resources*, 31(7), 1015-1023, 2008.

Villarini, G., P. V. Mandapaka, W. F. Krajewski and R. J. Moore, Rainfall and sampling uncertainties: A rain gauge perspective, *Journal of Geophysical Research-Atmospheres*, 113(D11), 2008.

Woldemeskel, F. M., A. Sharma, B. Sivakumar and R. Mehrotra, A framework to quantify GCM uncertainties for use in impact assessment studies, *Journal of Hydrology*, 5191453-1465, 2014.

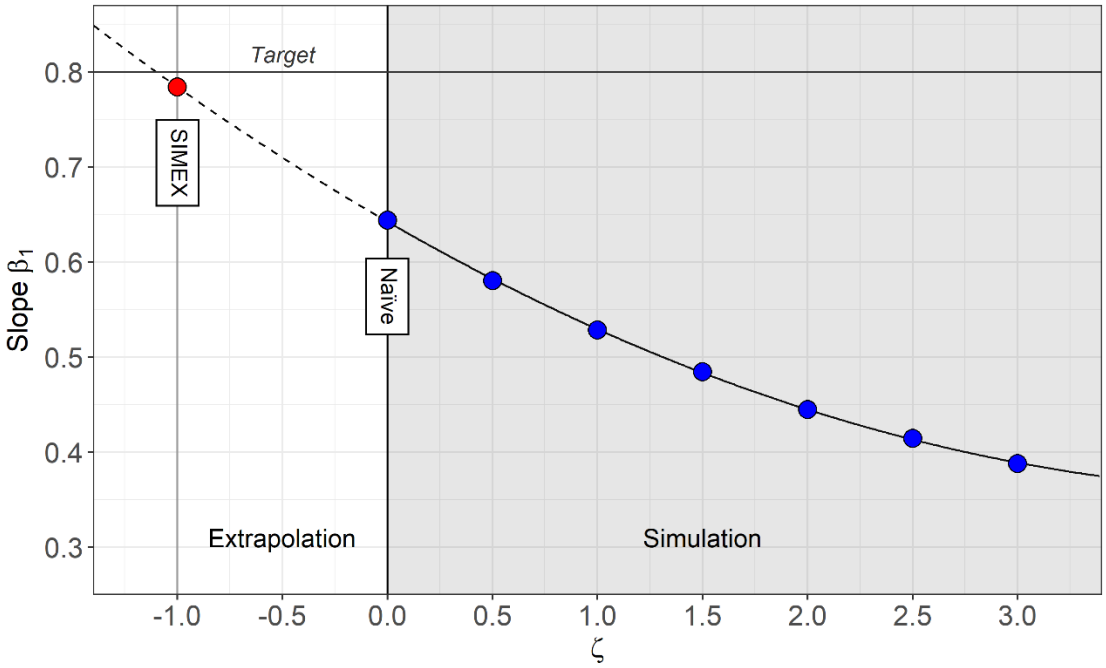

Figure 1. Plot of the slope β_1 as a function of ζ for $\beta_0 = 0$ and $\beta_1 = 0.8$. A value of ζ equal to 0 corresponds to the naïve estimator ($\hat{\beta}_{1,\text{naïve}} = 0.64$), while a value of -1 represents the SIMEX estimator ($\hat{\beta}_{1,\text{SIMEX}} = 0.78$).

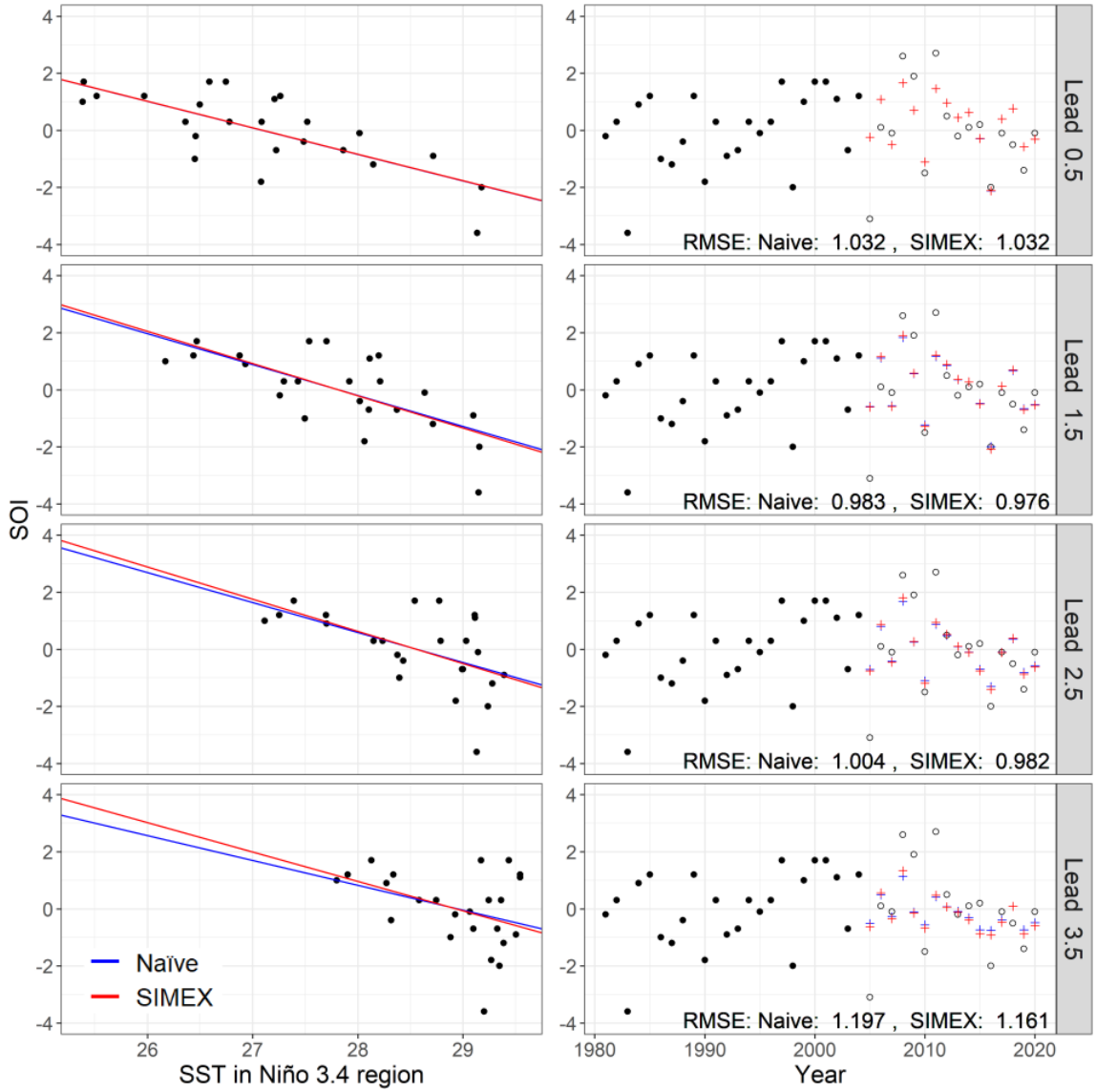

Figure 2. Example of the applicability of SIMEX in seasonal forecasting of the Southern Oscillation Index (SOI) in February as a function of sea surface temperature (SST) forecasts in the Niño3.4 region. Each row has a different lead time (from shortest to longest moving from the top to the bottom). The left column shows the scatterplot between the two variables, together with the regression lines based on OLS (blue) and SIMEX (red). The right panels show the time series of SOI used for the training of the model (black circles), the SOI values used for validation (white circles) and the forecasts based on the regression lines in the left panels, together with the corresponding root mean squared error (RMSE) values.

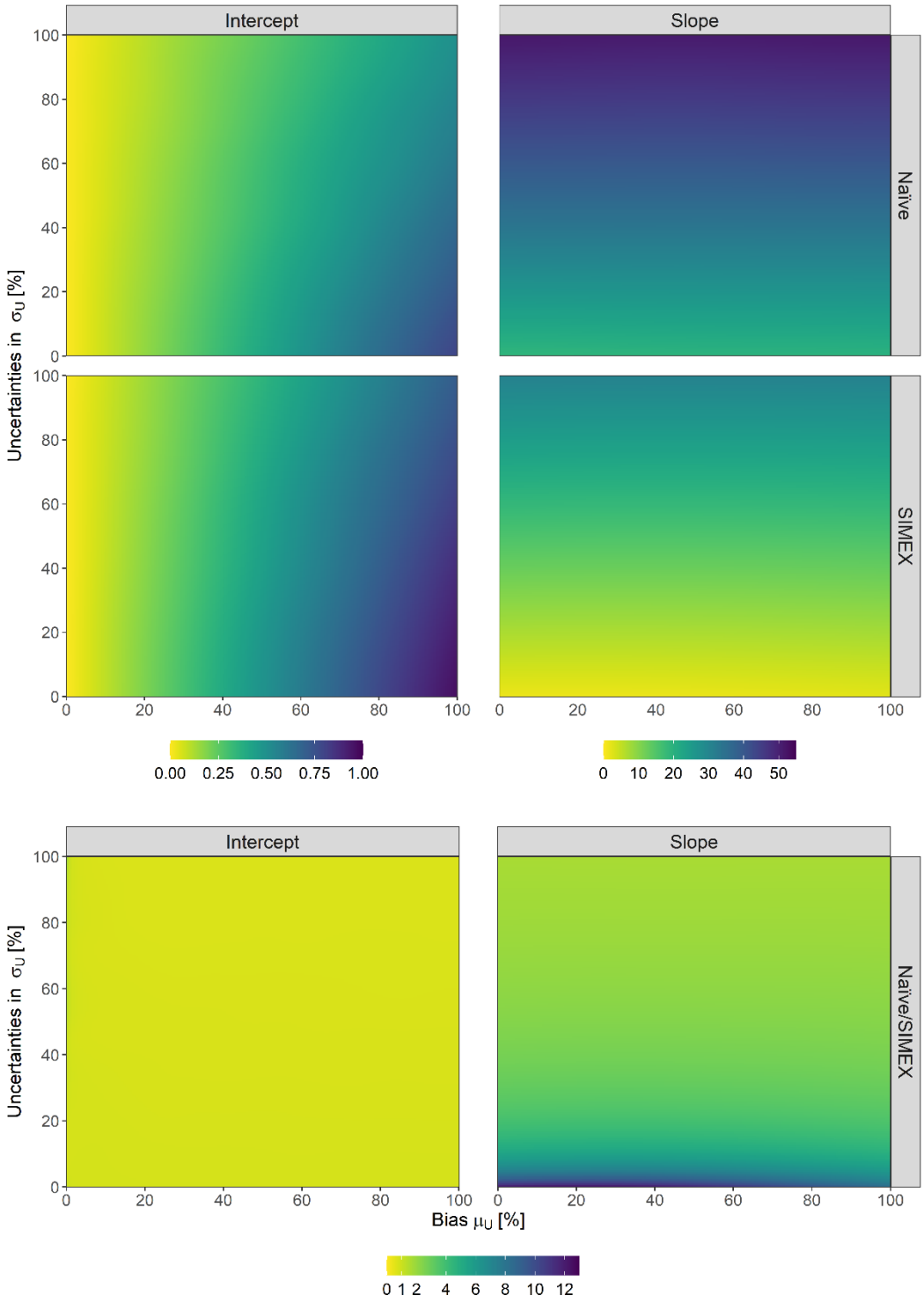

Figure 3. Sensitivity of the intercept (left panels; $\Delta_{\beta_0,\text{naïve}}$ and $\Delta_{\beta_0,\text{SIMEX}}$, respectively) and slope (right panels; $\Delta_{\beta_1,\text{naïve}}$ and $\Delta_{\beta_1,\text{SIMEX}}$, respectively) estimates based on the OLS (top row) and SIMEX (middle row) methodologies to mis-specification of the mean and standard deviation of the measurement error U. The value of the true slope β_1 is 0.3, while

the value of the true intercept β_0 is equal to 0. The panels in the bottom row show $\Delta_{\beta_0,\text{naïve}}/\Delta_{\beta_0,\text{SIMEX}}$ (bottom-left) and $\Delta_{\beta_1,\text{naïve}}/\Delta_{\beta_1,\text{SIMEX}}$ (bottom-right), which represent the ratios of the two panels above.

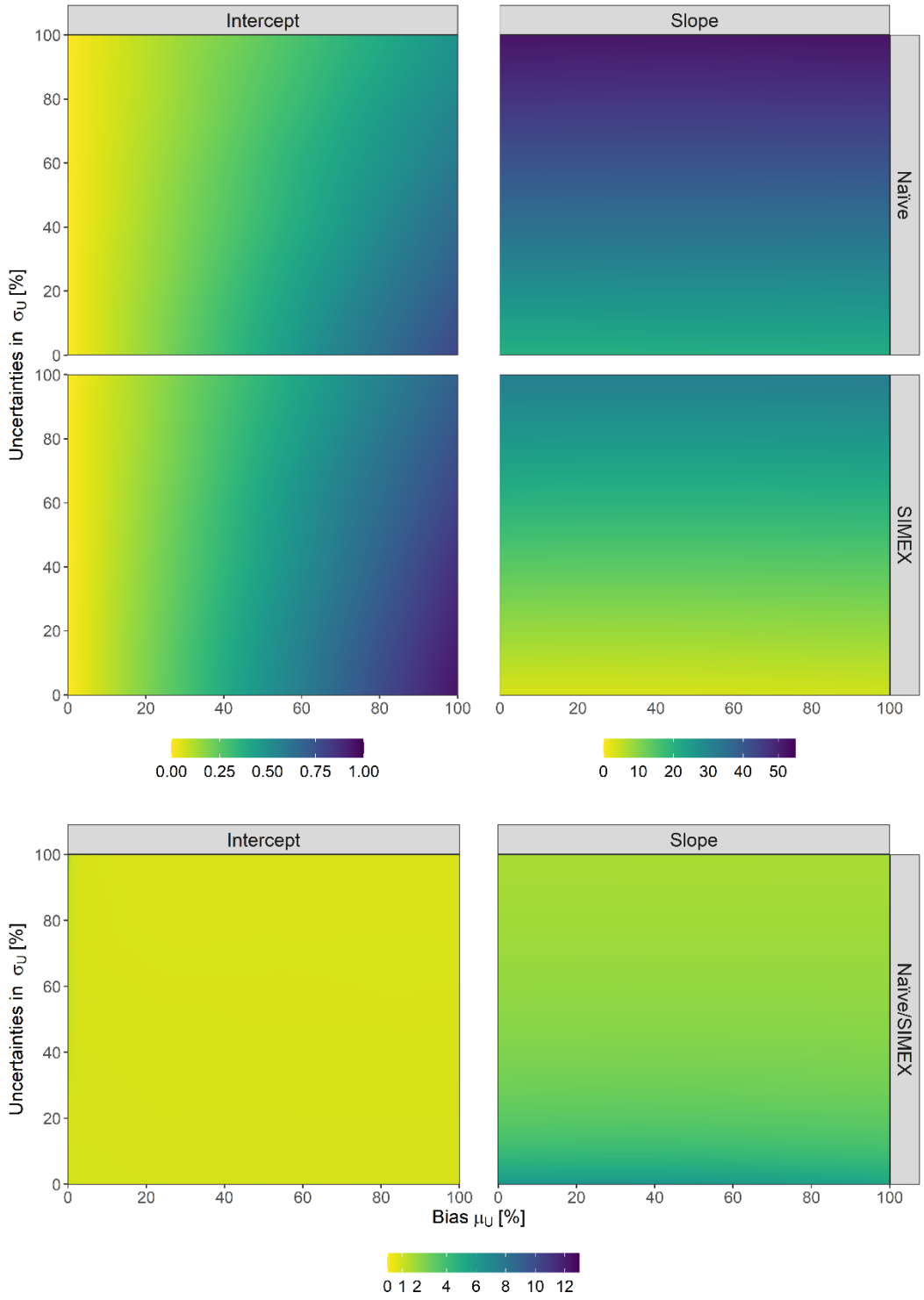

Figure 4. Same as Figure 3, but for case in which the true slope β_1 is 0.8, while the value of the true intercept β_0 is equal to 0.

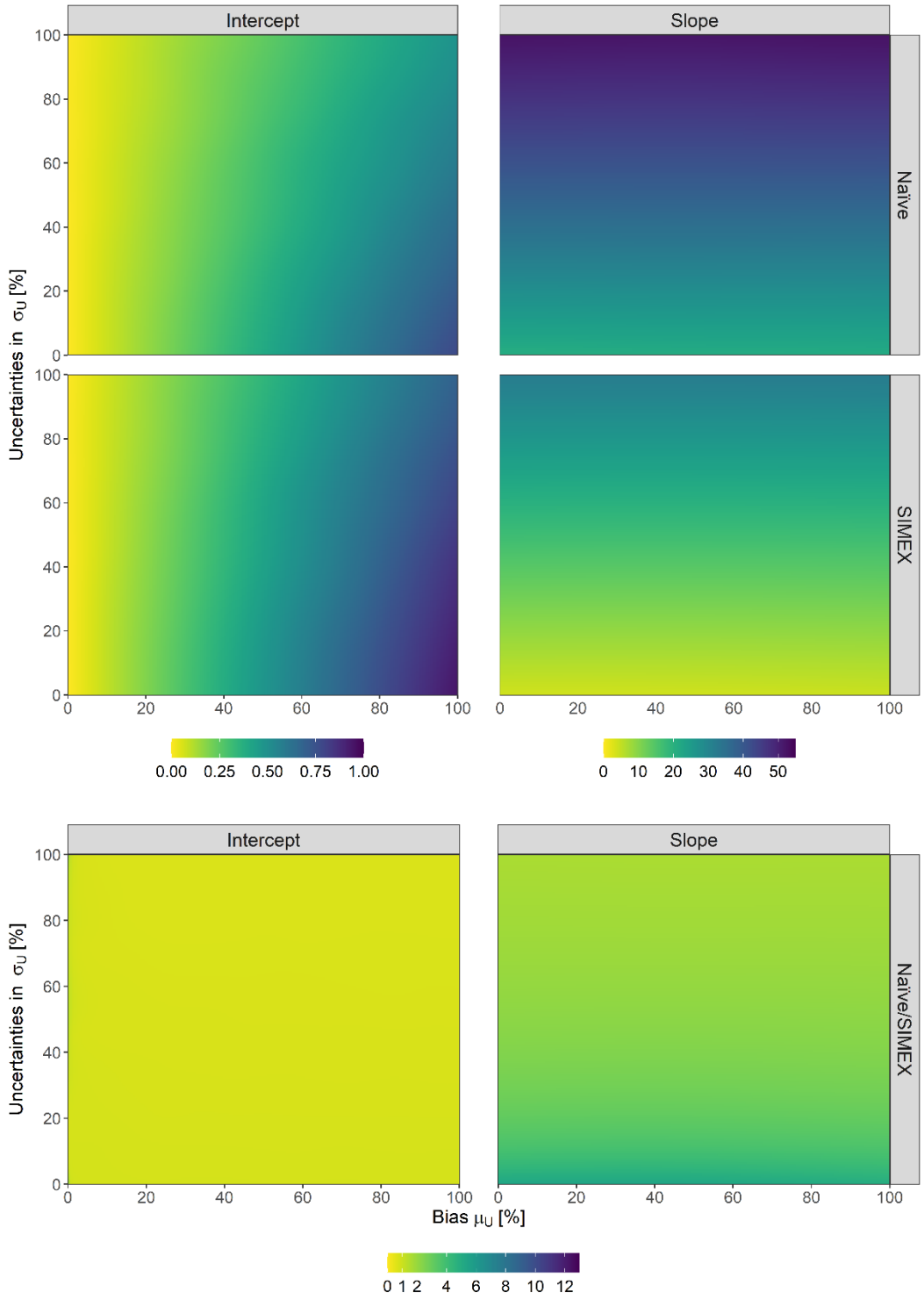

Figure 5. Same as Figure 3, but for case in which the true slope β_1 is 2.0, while the value of the true intercept β_0 is equal to 0.

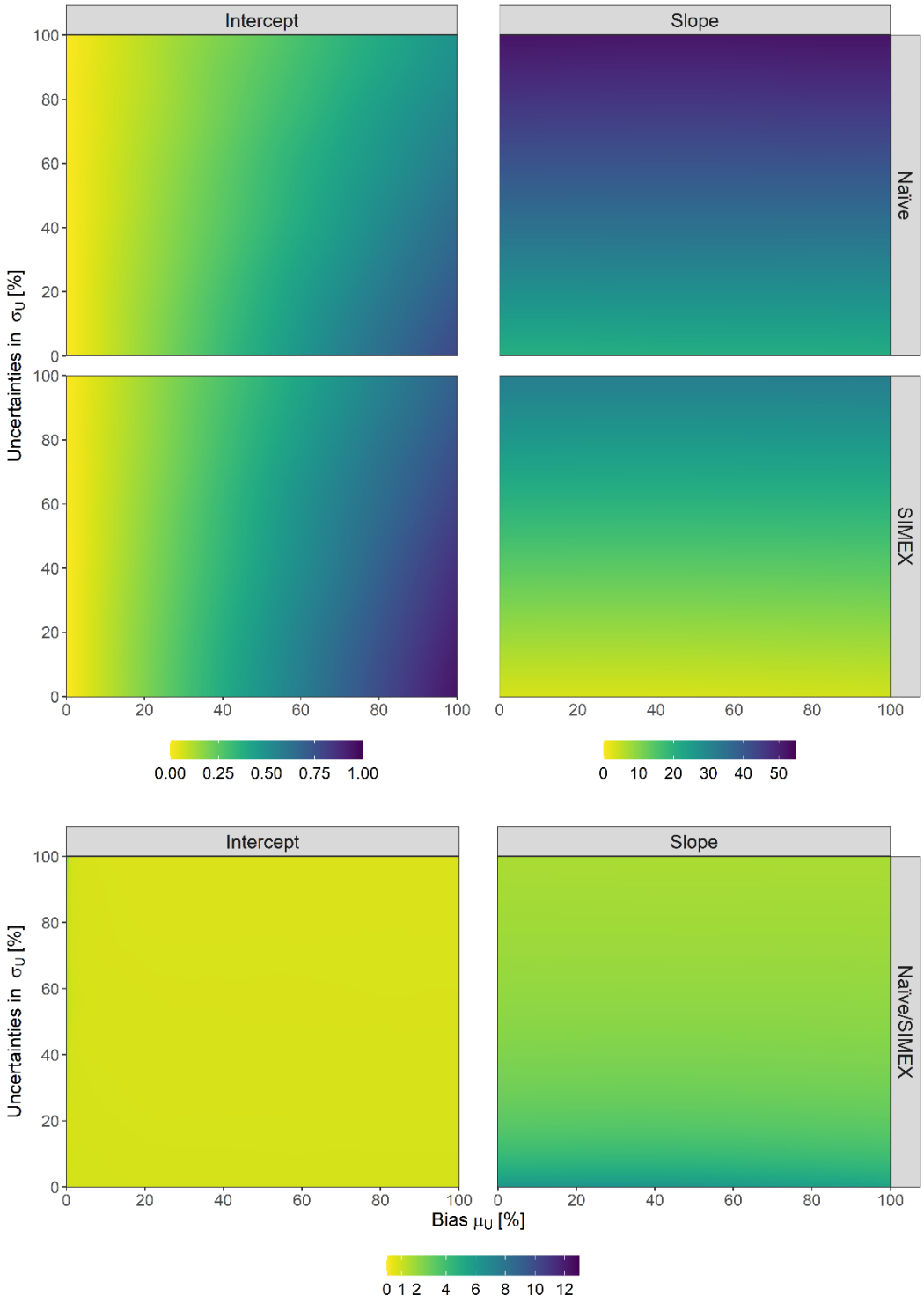

Figure 6. Same as Figure 3, but for case in which the true slope β_1 is 0.8, while the value of the true intercept β_0 is equal to 0.8.

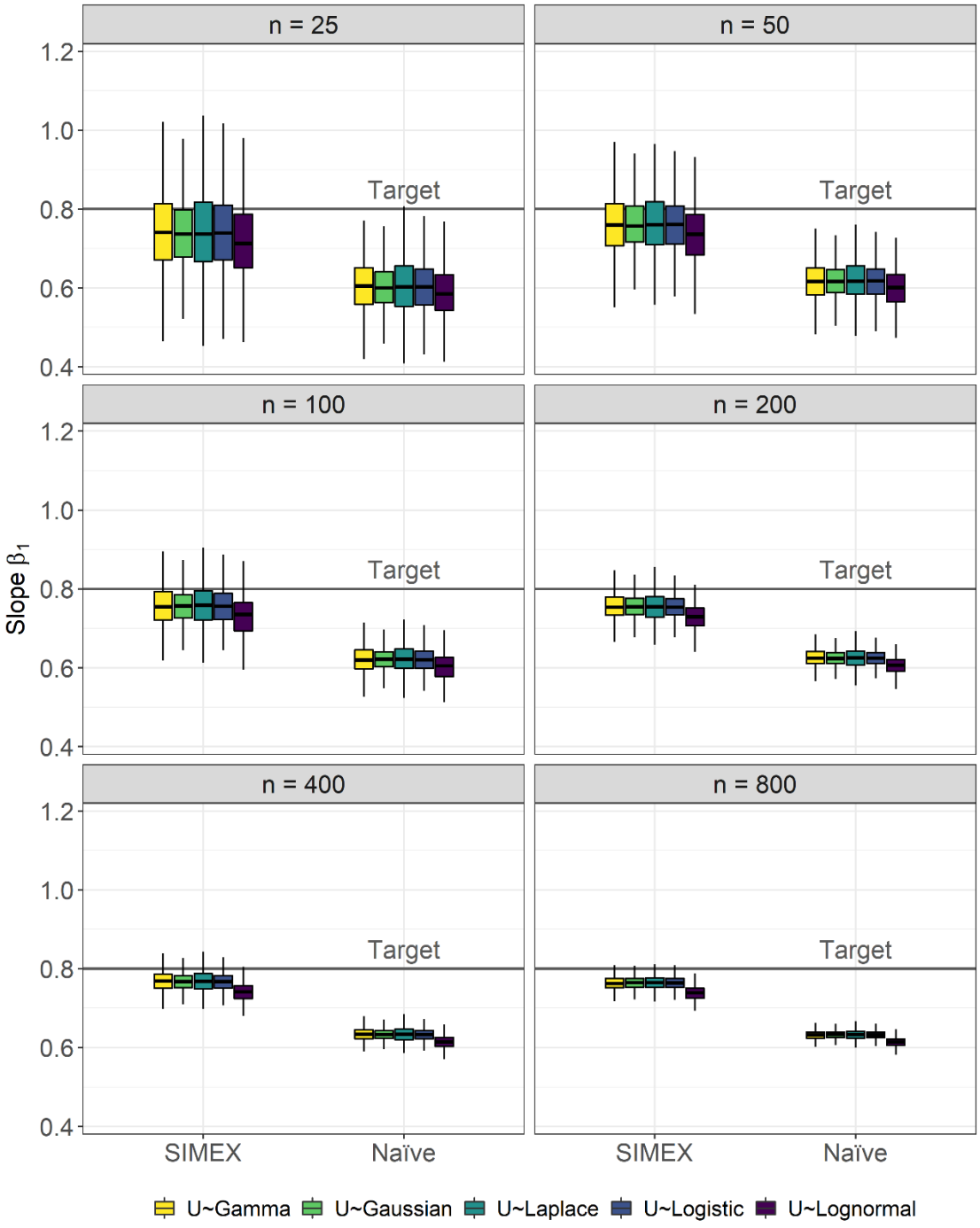

Figure 7. Boxplots showing the performance of SIMEX and the naïve estimation of the slope β_1 as a function of sample size and for different error distributions. In each boxplot, the solid line within the box represents the median, while the limits of the box the 25^{th} and 75^{th} percentiles; the limits of the whiskers indicate the 5^{th} and 95^{th} percentiles.