Research Calculator: how the checks are tested

Methods, testing and limitations · Version 1.3 (beta) · November 2026

Researchers need to know whether a checking tool can itself be trusted. This document sets out what each check is based on, how the add-in is tested, and what its results can and cannot tell you.

Principles

- **Published methods.** Every plausibility check implements a method from the peer-reviewed literature or a standard statistical result. The references are listed below and shown in the add-in.
- **Transparent.** Each result comes with its working, so it can be checked by hand.
- **Deterministic, no AI.** The same inputs always give the same results. No machine-learning or language model is involved.
- **Local.** All calculations run on the user's computer; nothing is sent anywhere.

How it is tested

The tests below were run on the Word checks for version 1.0 (October 2026).

1. **Against the reference implementation.** The checks are also published as web tools at rafiki.earth (checks and tools pages). The add-in was run side by side with the website's own code on the same inputs: 23 comparisons for the checks and 34 for the tools, covering typical values, edge cases, missing values and error messages. In every case the numbers, verdicts, tables and messages were identical.
2. **Equation regression suite.** 80 test equations (including matrices and transposes, subscripts, derivatives, integrals, sums, products and complex numbers) are rerun after every change and compared with the previous version. A result may change only when the change is an intended improvement.
3. **Languages.** Every check is run in English, Kiswahili and Spanish to confirm the calculations and results are the same in each.
4. **Word integration.** Reading equations, colouring them, jumping to them and adding comments are tested against simulated Word documents, for desktop Word and Word on the web.

References in Word, the Excel spreadsheet checks, the tools in Excel, the `=RA.` cell functions and R are not yet covered by this document.

What the results mean

> **A plausibility check is not a proof.** A check that passes shows that the reported numbers are consistent with each other and with the method; it does not show that the analysis is right. A check that flags a problem shows that the numbers cannot all be true as reported; the cause may be a typo, a different dataset or scale, or a genuine error.

- The equation checker tests equations numerically at random values. This is strong evidence, but not a symbolic proof, and very unusual equations may not be read correctly. Anything it cannot

read is reported as such rather than guessed.

- Checks for independence, statistical assumptions and model comparison are guided checklists that support your judgement; they do not give an automatic verdict.
- Results are only as good as the values entered. Check that units and scales match the document.

Methods and references

Area	Check	Basis
Equations	Equation checker Paste or select equations to catch unreadable or undefined expressions, rearrangements that don't follow, and worked examples that don't reproduce.	Numerical evaluation with math.js: each side is evaluated at many random values of the variables; rearrangements are checked by substituting one equation into another; worked examples are evaluated at the stated values.
Model accuracy and validation	Area from a confusion matrix Unbiased class areas with confidence intervals, instead of pixel counts.	Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42–57. doi:10.1016/j.rse.2014.02.015
Model accuracy and validation	Spatial cross-validation and leakage Is the validation optimistic? A guided checklist.	Roberts, D.R. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. doi:10.1111/ecog.02881 Meyer, H., Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12, 1620–1633. doi:10.1111/2041-210X.13650 Milà, C., Mateu, J., Pebesma, E., Meyer, H. (2022). Nearest neighbour distance matching leave-one-out cross-validation for map validation. Methods in Ecology and Evolution 13, 1304–1316. doi:10.1111/2041-210X.13851
Model accuracy and validation	Model evaluation and PICP Paste observed and predicted values: statistics, a 1:1 plot and a coverage curve.	Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. Malone, B.P., McBratney, A.B., Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626. Burrough, P.A., van Gaans, P.F.M., Hootsmans, R. (1997). Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115–135.
Model accuracy and validation	R^2, RMSE and SD Can the reported R^2, RMSE and spread all be true?	No paper needed: for the same validation data, RMSE ≥ SD × $\sqrt{(1 - R^2)}$ follows from the definitions of R^2 and RMSE, and no model can predict more precisely than the soil was measured.
Model accuracy and validation	Agreement metrics Can a table of R^2, CCC, RMSE and bias all be true together?	Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. The bounds follow from the definitions: CCC ≤ $\lvert r \rvert$, and $RMSE^2$ = $bias^2$ + the variance of the errors.

Area	Check	Basis
Model accuracy and validation	Classification accuracy Confidence interval, chance baseline and comparison with another map.	Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. J. Am. Stat. Assoc. 22, 209–212. doi:10.1080/01621459.1927.10502953 Foody, G.M. (2009). Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. Remote Sens. Environ. 113, 1658–1663. doi:10.1016/j.rse.2009.03.014
Spatial structure and sampling	Effective sample size How much clustered samples are really worth.	Cressie, N.A.C. (1993). Statistics for Spatial Data, revised ed. Wiley. Section 1.3. Webster, R., Oliver, M.A. (2007). Geostatistics for Environmental Scientists, 2nd ed. Wiley.
Spatial structure and sampling	Finding the right pixel size Coarsest legible, finest legible and recommended resolution for a map.	Hengl, T. (2006). Finding the right pixel size. Computers & Geosciences 32, 1283–1298. doi:10.1016/j.cageo.2005.11.008
Spatial structure and sampling	Kriging error vs. the variogram Is a reported kriging RMSE possible with the stated variogram and sampling?	Burgess, T.M., Webster, R., McBratney, A.B. (1981). Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. J. Soil Sci. 32, 643–659. doi:10.1111/j.1365-2389.1981.tb01737.x Webster, R., Oliver, M.A. (2007). Geostatistics for Environmental Scientists, 2nd ed. Wiley.
Spatial structure and sampling	Samples for a map (from the variogram) How many samples on a grid give a chosen kriging error.	Burgess, T.M., Webster, R., McBratney, A.B. (1981). Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. J. Soil Sci. 32, 643–659. doi:10.1111/j.1365-2389.1981.tb01737.x McBratney, A.B., Webster, R. (1983). How many observations are needed for regional estimation of soil properties? Soil Science 135, 177–183.
Spatial structure and sampling	Can the change be detected? Smallest change in soil carbon a sampling design can detect.	Garten, C.T., Wullschleger, S.D. (1999). Soil carbon inventories under a bioenergy crop (switchgrass): measurement limitations. J. Environ. Qual. 28, 1359–1365. doi:10.2134/jeq1999.00472425002800040041x
Soil data and units	Carbon stock arithmetic Rebuild a stock from its inputs and catch unit and stone-fraction slips.	Poeplau, C., Vos, C., Don, A. (2017). Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. SOIL 3, 61–66. doi:10.5194/soil-3-61-2017 Pribyl, D.W. (2010). A critical review of the conventional SOC to SOM conversion factor. Geoderma 156, 75–83. doi:10.1016/j.geoderma.2010.02.003
Soil data and units	Back-transformation Was a log-scale result reported as the median or the mean?	Aitchison, J., Brown, J.A.C. (1957). The Lognormal Distribution. Cambridge University Press. Journel, A.G. (1980). The lognormal approach to predicting local distributions of selective mining unit grades. Mathematical Geology 12, 285–303. doi:10.1007/BF01029417 Webster, R., Oliver, M.A. (2007). Geostatistics for Environmental Scientists, 2nd ed. Wiley.

Area	Check	Basis
Soil data and units	Lab method conversions Walkley–Black, organic matter factors, pH in $CaCl_2$ vs. water, oxides.	Walkley, A., Black, I.A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Science 37, 29–38. Pribyl, D.W. (2010). A critical review of the conventional SOC to SOM conversion factor. Geoderma 156, 75–83. doi:10.1016/j.geoderma.2010.02.003 Ahern, C.R., Baker, D.E., Aitken, R.L. (1995). Models for relating pH measurements in water and calcium chloride for a wide range of pH, soil types and depths. Plant and Soil 171, 47–52. doi:10.1007/BF00009563 Oxide and unit conversions are exact, from molar masses.
Soil data and units	Coordinates and distances Degrees treated as metres, and how far a degree really is.	No paper needed: standard spherical geometry (haversine formula, mean Earth radius 6,371 km).
Study design and statistical reasoning	Independence and study design Are the sampling unit, analysis and validation aligned? How much does grouping shrink the effective sample size?	Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54, 187–211. doi:10.2307/1942661 Roberts, D.R. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. doi:10.1111/ecog.02881 Meyer, H., Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12, 1620–1633. doi:10.1111/2041-210X.13650 Kish, L. (1965). Survey Sampling. Wiley, New York. Moulton, B.R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Review of Economics and Statistics 72, 334–338. doi:10.2307/2109724
Study design and statistical reasoning	Statistical assumptions check Which assumptions matter for this method? Paste residuals for skewness, outliers, unequal spread and autocorrelation.	This beta tool is a structured checklist, not a test of model assumptions. The relevant assumptions depend on the exact statistical procedure and intended inference. An unchecked item only means that you have not marked the check as reported. The residual diagnostic screens one set of residuals; it does not replace a full model diagnosis. Jarque, C.M., Bera, A.K. (1987). A test for normality of observations and regression residuals. International Statistical Review 55, 163–172. doi:10.2307/1403192 Durbin, J., Watson, G.S. (1950). Testing for serial correlation in least squares regression. I. Biometrika 37, 409–428. doi:10.1093/biomet/37.3-4.409

Area	Check	Basis
Study design and statistical reasoning	Model comparison audit Are competing models evaluated on comparable terms? Test the difference with paired fold scores or McNemar.	This beta check examines the comparability of model evaluations based on your selected answers, and tests a difference only from the paired scores or counts you enter. For spatial, temporal or grouped data, validation design should reflect the intended prediction task (Roberts et al. 2017, Ecography 40, 913–929). Nadeau, C., Bengio, Y. (2003). Inference for the generalization error. Machine Learning 52, 239–281. doi:10.1023/A:1024068626366 Bouckaert, R.R., Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In: Advances in Knowledge Discovery and Data Mining (PAKDD 2004), LNCS 3056, 3–12. Springer. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157. doi:10.1007/BF02295996 Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10, 1895–1923. doi:10.1162/089976698300017197
Study design and statistical reasoning	F tests and degrees of freedom Does each p follow from F and its degrees of freedom? Which df does the reported p imply, and which does the design allow?	Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54, 187–211. doi:10.2307/1942661 Kenward, M.G., Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983–997. doi:10.2307/2533558 Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin 2, 110–114. doi:10.2307/3002019 Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S., Wicherts, J.M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). Behavior Research Methods 48, 1205–1226. doi:10.3758/s13428-015-0664-2