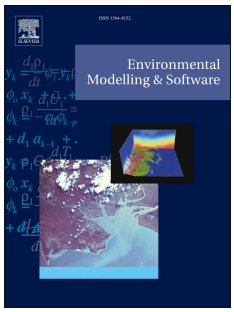
Accepted Manuscript



Model validation: A bibliometric analysis of the literature

Sibel Eker, Elena Rovenskaya, Simon Langan, Michael Obersteiner

PII: S1364-8152(18)31276-3

DOI: https://doi.org/10.1016/j.envsoft.2019.03.009

Reference: ENSO 4413

To appear in: *Environmental Modelling and Software*

Received Date: 21 December 2018

Accepted Date: 19 March 2019

Please cite this article as: Eker, S., Rovenskaya, E., Langan, S., Obersteiner, M., Model validation: A bibliometric analysis of the literature, *Environmental Modelling and Software* (2019), doi: https://doi.org/10.1016/j.envsoft.2019.03.009.

Model validation: A bibliometric analysis of the literature

Sibel Eker[a,*],

Elena Rovenskaya[a,b],

Simon Langan[a],

Michael Obersteiner[a]

[a] *International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A2361 Laxenburg, Austria*

[b] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia*

[*] Corresponding author, eker@iiasa.ac.at

Software and data availability: The dataset of academic publications used in this paper is obtained from the Scopus database, and the analysis is implemented in an IPython notebook. Both the dataset and the analysis scripts are available via https://github.com/sibeleker/Validation.

Model validation: A bibliometric analysis of the literature

Highlights

- We conduct citation and text-mining analyses on a broad model validation literature.
- *Data* and *predict* are the most common words in the studied publication dataset.
- The most-cited publications are not similar to the rest in terms of their content.
- Validation practices of different modeling fields are closed to each other.

Abstract

Validation is a crucial step in environmental and economic modeling that establishes the reliability of models to be used in decision-making contexts. It is often said that validation approaches proposed in the literature are not widely adopted, and different modeling fields do not benefit from each other. This study analyses a broad academic literature on model validation, mainly in environmental and decision sciences, by using an innovative combination of bibliometric and text-mining tools. The results show that a data-driven validation practice is prevalent. Although most publications in the studied dataset resemble each other, the most-cited ones tend to be different from the rest in terms of their abstracts' content. Furthermore, the validation practices in different modelling areas are distinct, and do not extensively cite each other. In future, validation approaches can extend beyond data-oriented reliability for a wider acceptance of modelling in decision-making, and can synthesize the methods and views from various fields.

Keywords

Model validation, model evaluation, model testing, citation analysis, text-mining analysis

1 Introduction

Modelling has long assisted the management of and decision-making in socio-economic and environmental systems. The reliability of models has long been debated, too, with criticisms that tend to cluster around the following issues: Models do not utilize high quality data, or they extrapolate the past data to predict future; models fail to include relevant and important processes in their scopes; or models include false assumptions such as averages and linearity (Maslin and Austin, 2012; Pilkey and Pilkey-Jarvis, 2007; Saltelli and Funtowicz, 2014).

In line with these critiques that pinpoint data use, model conceptualization, boundaries and assumptions as the most important issues, Smith and Petersen (2014) distinguish between three dimensions of a model's reliability. *Statistical* reliability refers to the subjective or objective probability distributions communicated in the model-based findings. It covers the concepts of data and behavior (model output) validity. Statistical tests that compare the output of a model to empirical data support this type of reliability. *Methodological* reliability results from the consideration of model purpose, and it refers to whether the model fits its purpose

conceptually and technically. Related to the concepts of conceptual, logical and structural validity, methodological reliability is established by several tests. The commonly used examples of these tests are stress tests (extreme-conditions tests) which check whether the model generates observed or anticipated output when parameters are set to extreme values, or sensitivity analyses which check whether the model outputs are sensitive to its inputs (Balci, 1994; Barlas, 1996). *Public* reliability indicates the extent of public trust in scientists in general and modelers in particular. This is often proposed to be established by 'soft' and participatory approaches (van der Sluijs, 2002).

Validation is a crucially important modeling step to establish the reliability of models and expel criticism. In environmental and economic modeling, validation deals mostly with statistical and methodological reliability with several approaches and techniques developed in different areas of environmental science. Whether they focus on model output or structure, these techniques address the representation power of a model, i.e. how well it represents reality. For instance, Matott et al. (2009) present an extensive review of software-based evaluation methods and tools with a focus on statistical reliability, data quality, sampling, input and output uncertainty. Validation approaches in biophysical modeling (Bellocchi et al., 2010), ecological modeling (Augusiak et al., 2014), and environmental modeling (Bennett et al., 2013) acknowledge that validity extends beyond representation, especially beyond an accurate representation of empirical data by model output. Yet, these studies still focus on quantitative, data-oriented techniques that aim to reduce the uncertainty in model outcomes.

It has been recognized that although such realism in validation has served well, it has major philosophical and pragmatic flaws (Beven, 2002; Oreskes and Belitz, 2001; Oreskes et al., 1994). Following this, several studies offer integrated validation frameworks that consider different types of validity at different stages of model development. For instance, the evaluation step in Jakeman et al. (2006)'s ten-stepped model development framework acknowledges the extension of fitness for purpose to 'softer' criteria beyond representation accuracy, like accommodating unexpected scenarios, diverse categories of interests and time frames. Schwanitz (2013) incorporates approaches from various fields such as operations research and simulation to integrated assessment modeling, and proposes a validation framework that iteratively evaluates conceptual, logical, data, behavior and structure validity to ensure methodological reliability. van Vliet et al. (2016) review the validation practice in land-change modeling, and discuss validity as a broader concept extending to usefulness, transparency and salience.

As for public reliability, Risbey et al. (2005) provide a checklist that can guide participatory model evaluation approaches. Applied to the TIMER global energy system model, this checklist covers a wide variety of issues to be discussed by stakeholders, e.g. whether the right outcome indicators are chosen, whether the model can be used for different value systems, and whether the model output is sensitive to the parameter values as well as alternative model structures. Based on this checklist, van der Sluijs et al. (2008) present a good practice guidance that focuses on problem framing, involvement of stakeholders, selection of performance indicators, appraisal of knowledge base, and assessing and reporting relevant uncertainties. Refsgaard et al. (2005) review technical and non-technical guidelines for modeling and model use in the hydrology and

water management domain. These guidelines contribute to public reliability directly by facilitating the interaction between modelers and water managers.

Despite such a variety of validation approaches, it is often said that these approaches are not widely adopted by practitioners, i.e. modelers and analysts who develop and evaluate models. For instance, van Vliet et al. (2016) find that calibration or validation approaches are not even mentioned in a large portion of the publications on land-use modeling. Furthermore, many publications focus on a single area of environmental modeling, hence may not benefit from the validation approaches developed in other modelling areas or in different fields such as operations research and simulation. For instance, different validity types and various validation issues that are recently discussed in ecological modelling (Augusiak et al., 2014) were discussed earlier in the decision sciences literature (Landry et al., 1983).

In line with these two issues of uptake and connection across modeling fields, the objective of this study is to examine the extent of the adoption and acknowledgement of validation in the environmental and economic modelling publications, and to investigate the relations between the validation practices in different modelling areas. For this purpose, we employ a combination of citation and text-mining analyses on a large dataset of academic publications. The specific questions we aim to answer are: (i) What are the prevalent concepts in the publications in this dataset? (ii) How related are these publications in terms of their content? (iii) How does this relatedness reflect on their citation scores as an indicator of their uptake? (iv) Can this relatedness be explained by different topics that refer to different areas of environmental modeling?

In the remainder of this paper, Section 2 describes the bibliometric and text-mining methods we use. Section 3 presents the results of these analyses and answers the abovementioned questions. Section 4 discusses the implications of these findings for the current and future validation research. The paper ends with conclusions in Section 5.

2 Methods

Bibliometrics, broadly defined as a quantitative analysis of published units (Broadus, 1987), is increasingly used to investigate the temporal, content, collaboration or citation trends in scientific fields or journals (Cancino et al., 2017; Laengle et al., 2017; Merigó et al., 2018). In this study, we combine a bibliometric and text-mining analysis to provide an overview of the academic literature on validation in environmental and economic modeling. Although validation literature has been reviewed extensively in several modelling areas (Augusiak et al., 2014; Bellocchi et al., 2010; Bennett et al., 2013; Tsioptsias et al., 2016), our approach with bibliometrics and text-mining is more comprehensive since it analyses a much broader literature. This bibliometric approach also provides quantitative information that relates the content to uptake of the publications measured by citation scores.

In particular, we employ a data visualization technique to map the publications based on their content similarities, merge this mapping with citation analyses, and with the main topics identified by another text-mining technique called topic modeling. To have flexibility and customization opportunities, we use script-based algorithms instead of a software package such

as VOSviewer (van Eck and Waltman, 2010). Below, we describe the specifications of the publication dataset and explain the mapping and topic modeling methods we use.

Dataset

The publication dataset we analyse in this study is retrieved from the Scopus database with the search keyword *model validation* and similar terms such as *evaluation, assessment* or *testing.* The search focuses mainly on environmental science, economics and decision sciences, and the related fields of sustainability science such as agriculture and energy. Table 1 lists the predefined Scopus fields included in our study. The search results are limited to these fields by excluding all other predefined Scopus fields such as chemistry, engineering and psychology. This implies that, if an article is classified in multiple subjects, for instance in environmental science and chemistry, it is not included in this dataset. Table 1 summarizes these search criteria, which returned 10,739 publications in total between the publication dates of 1980 and 2017. The final dataset contains 10,688 of these publications, after the duplicate items or items with insufficient content have been removed. Figure A.1 in the Appendix shows how this publication dataset is distributed over the years.

Table 1: Search criteria used to retrieve the publication dataset

Search field	Search criteria
Any of the title, abstract or keywords include	"model validation" OR "model validity" OR "model evaluation" OR "model assessment" OR "model testing"
Language	only English
Predefined Scopus fields	Environmental science Computer science Agricultural and biological sciences Mathematics Energy Social sciences Economics, econometrics and finance Decision sciences Multidisciplinary

The bibliometric analysis is based on the citation scores of these publications reported by Scopus (as of 11 May 2018), and the references they cite to determine the citation relations within the dataset. For the text-mining analysis, their abstracts are used to examine the content similarity between the publications, and to identify the main topics. Prior to text-mining, all general stopwords are removed from the abstracts, as well as the words that have no significant meaning in this case, such as *model*, *validation*, *research*, *analysis*. All words are stemmized, implying that the words with the same root, for instance *predicting* and *prediction*, are reduced to their stem (*predict*) and considered the same. This preparation of the textual data is done by using the Natural Language Toolkit (NLTK) (Bird and Loper, 2004), which is a Python-based natural language processing software for English.

Relatedness of the validation publications: Nonlinear mapping

One question addressed in this study is how the validation publications from various fields are related to each other in terms of content similarity and in terms of citation scores. We investigate the content relatedness of publications by mapping them on two-dimensional space. In bibliometric analysis, there are two main approaches to mapping, being graph-based and distance-based (van Eck and Waltman, 2010). We use a distance-based mapping technique, so that similar articles are positioned closer to each other. In particular, we use a nonlinear dimensionality reduction and data visualization technique called t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten, 2014; Maaten and Hinton, 2008), as implemented in Python's machine learning library *scikit-learn*.

The t-SNE algorithm builds a map of data points on which the distances between the points depend on similarities between them. In our case, each data point is a publication, represented in a multidimensional space by the words in its abstract. Each word corresponds to a dimension in this space. Similarity between two publications is then defined based on the distance between them in this multidimensional space. The algorithm assigns a small number of data points to each data point based on their similarity, and then constructs an undirected graph with reduced dimensions. This layout technique tends to spread the data points locally, but positions the dissimilar points further away. In other words, the publications similar to each other in terms of the content of their abstracts are positioned closer to each other. Therefore, the dense regions of the resulting map correspond to the clusters of similar work.

Main topics in the model validation publications

Mapping the profile of academic literature helps to identify various clusters of work. However, although potential clusters formed by it are based on content similarity, t-SNE is a visualization and dimensionality reduction algorithm that does not aim to search for topics precisely. Therefore, we use another text-mining method that enables discovering the main topics in a collection of documents with the aid of statistical techniques that are generally named topic modelling (Cunningham and Kwakkel, 2016). In this study, we use topic modelling to investigate whether relatedness observed on the map aligns with the major topics discussed in the abstracts of the model validation publications. In particular, we adopt the most commonly used topic modelling method, which is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and we use its open source implementation in a Python package (lda Developers, 2014).

An LDA implementation starts with a user-defined number of topics, i.e. bags, and the algorithm then probabilistically allocates each document to one of these bags to a certain extent. This extent signifies the topic probability of a document. In other words, it is not an exclusive allocation where each document is placed in only one bag, but each document is assigned to a bag by a percentage. In that way, LDA forms document-topic and topic-word pairs based on the words included in each document. In this study, when we divide the dataset into subsets based on the identified topics, we associate each publication with the topic it is assigned to by the highest topic probability. For instance, if publication A's topic probabilities are 22%, 35%, 18%, 25% for Topics I, II, III, IV, respectively, then it is associated with Topic II. This choice of assigning a document to only one topic based on the highest topic probability carries the risk of over-distinguishing the topics. However, the document-topic pairs (Figure A.4 in Appendix)

show that the topics identified by LDA are quite distinct, meaning that most publications can be exclusively associated with one of the topics.

While mapping and topic modelling enable covering a large number and wide variety of publications, they cannot analyze and interpret the content as precisely as a human reviewer can do. They identify the relationships between documents based on co-occurrence of words, and main themes based on word frequency. For instance, the publications deemed similar in terms of the word content by the mapping algorithm may not be using very similar validation approaches. The similarity of the validation approaches can only be inferred, because the publications inputted to the data mining algorithms are selected based on their focus on validation. Therefore, the methods used in this study do not single out the differences between different validation approaches and different modelling fields precisely and definitively. They provide information about the general themes, trends and relations.

3 Results

Overview of the model validation publications: Prevalent concepts and journals

Figure 1 lists the most frequent words in the abstracts of the publications in our dataset, which contain 'model validation' explicitly in their title, abstract or keywords. *Data* is the most common word, indicating that the validation practice is strongly associated with data in general, whether it is used as model input or to match the model output. *Prediction* receives the second rank, which can be interpreted as a prediction-orientation in these modeling studies. Furthermore, the emergence of water and soil among the most common words indicates that our dataset contains mostly ecosystems and hydrology studies.

Figure 2 shows the top 20 publication sources in the model validation literature. In other words, it shows the journals that published the highest 20 number of model validation articles, together with their citation scores in 2017. Citation scores represent the CiteScore metric of the Scopus database, which is computed as the ratio of total citations of a journal in 2017 to the total number of documents published in it between 2014 and 2016. This list of publication sources is dominated by the environment and ecosystems journals (1339 articles, 12% of the dataset) and hydrology journals (978 articles, 9% of the dataset), which relates to the previous finding that *water* and *soil* are two of the most common words. There are also several energy and environment journals among the top journals. An unexpected observation is that this list does not contain any journals that focus on the simulation methodology from a decision sciences perspective. This finding can be related to the prominent weight of environmental science in the publication dataset. Over 5000 articles in the dataset are labeled with environmental science, whereas only around 900 articles are labeled with decision sciences and economics. Furthermore, the sources which contain the highest number of validation publications are not the ones with the highest citation scores.

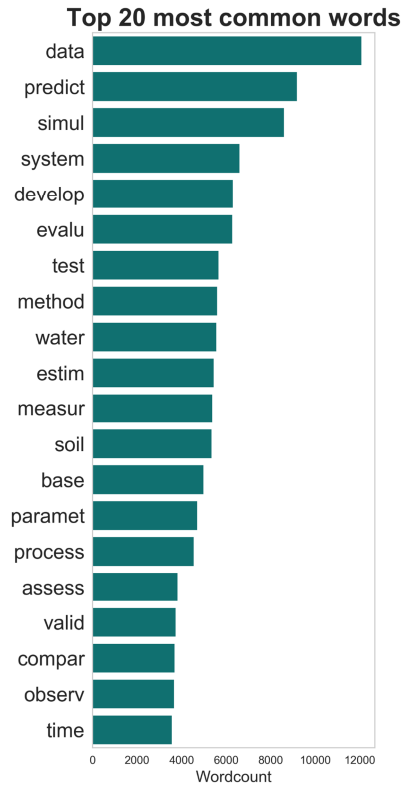


Figure 1: Top 20 most common words in the model validation publications

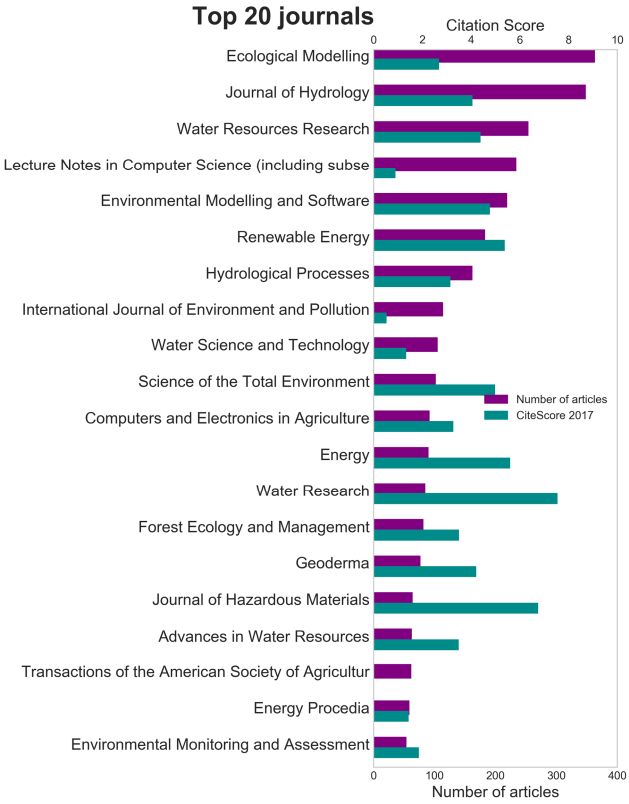


Figure 2: Top 20 journals where model validation studies are published

Relatedness of the validation publications: Nonlinear mapping

Figure 3 visualizes the relatedness of model validation publications resulting from the t-SNE mapping. Instead of scattering individual data points (publications), we plot a density map that shows where most articles accumulate. The darker a region is in this figure, the higher the number of articles there. The presence of a central dense region indicates that there is a large number of articles, which are very similar to each other in terms of their abstracts' word content compared to the rest of the publication dataset, hence positioned in close proximity. There are also several small and distinct clusters around this core with varying degrees of density, demarcated by white rectangles for visualization purposes. These clusters indicate groups of publications that are clearly distinguished from the central one, yet similar within the cluster. Also, the top five words of the articles falling into the corresponding rectangle are listed in the ranked order. *Data* is the top word in the core region and it is among the top five words in all demarcated clusters, yet in lower ranks. *Predict* is also among the top words in some of these clusters, yet not in the core one. Application areas, such as ecosystems and water (bottom two and rightmost clusters) seem to play a role in distinguishing the clusters. Methodological differences are also visible. For instance, the upper left cluster more dominantly contains *data*-oriented *tests* for the model output, while the central right cluster next to the core focuses on *parameter estimation* and *uncertainties*.

A few of the well-known and highly cited publications in the validation literature are marked on Figure 3, too. Oreskes et al. (1994) state briefly that model validation in a purely positivist way

is impossible; therefore, models should be used as heuristics. This article is considerably distant from dense regions of the map, indicating that its rather philosophical content does not have a strong resemblance to most articles. In particular, while Oreskes et al. (1994) contains common words such as *predict, evaluate, observe*, it also has several uncommon words such as *impossible, heuristic* and *logic*. The other two well-known articles (Bennett et al., 2013; Jakeman et al., 2006) address environmental modeling domain specifically and they are positioned relatively close to the central and dense region on the map. Therefore, it can be said that their contents are highly related to the majority of model validation publications in our dataset. In addition to the common words such as *data, test, calibrate*, these two articles contain the words *aim, purpose, tailor, custom* frequently, indicating a validation approach based on model purpose, i.e. fit for purpose. Another peripheral article is Schwanitz (2013), which stresses the importance of an *integrated* validation approach, *documentation* and *communication* with *stakeholders* for *transparency*, especially for the models used to assess the impacts of climate change on socioeconomic systems, and hence heavily concern public decision-making. Table A.1 in the Appendix contains the entire word list of these four articles used in this analysis.

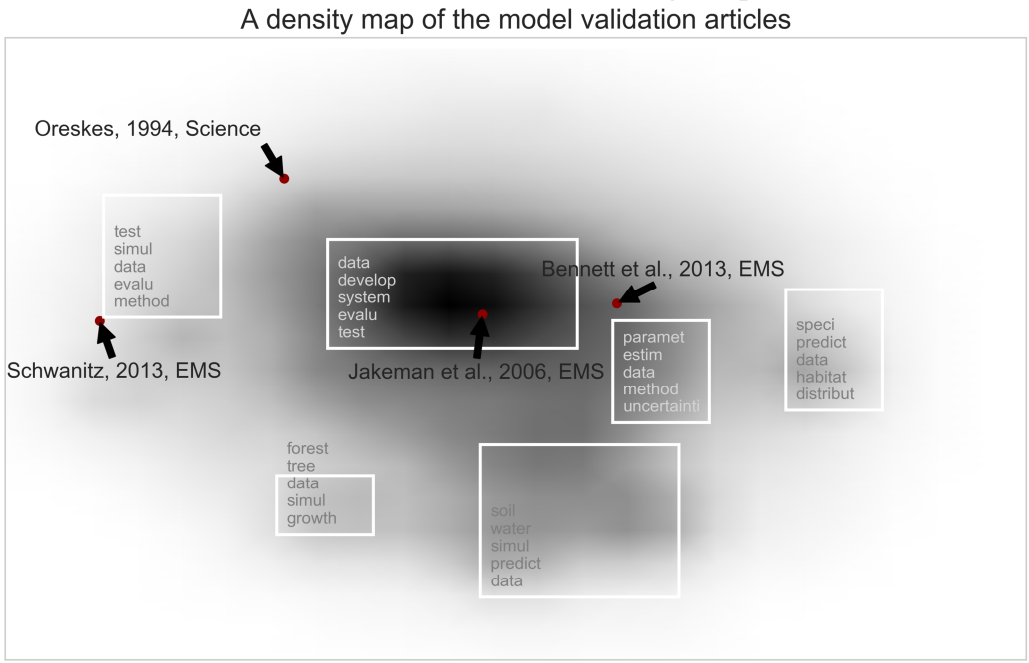
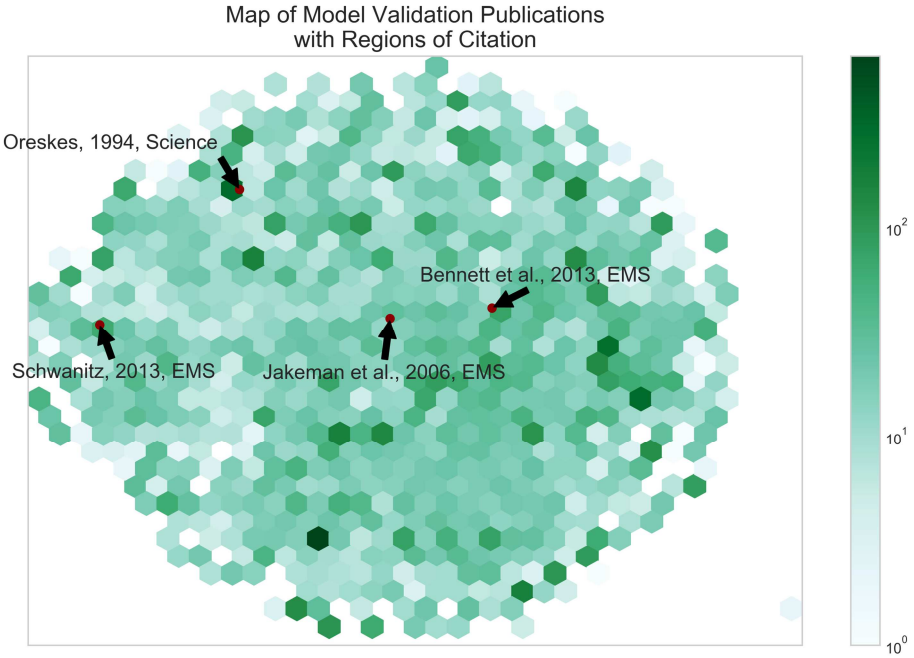


Figure 3: A density map of the model validation publications resulting from the t-SNE application

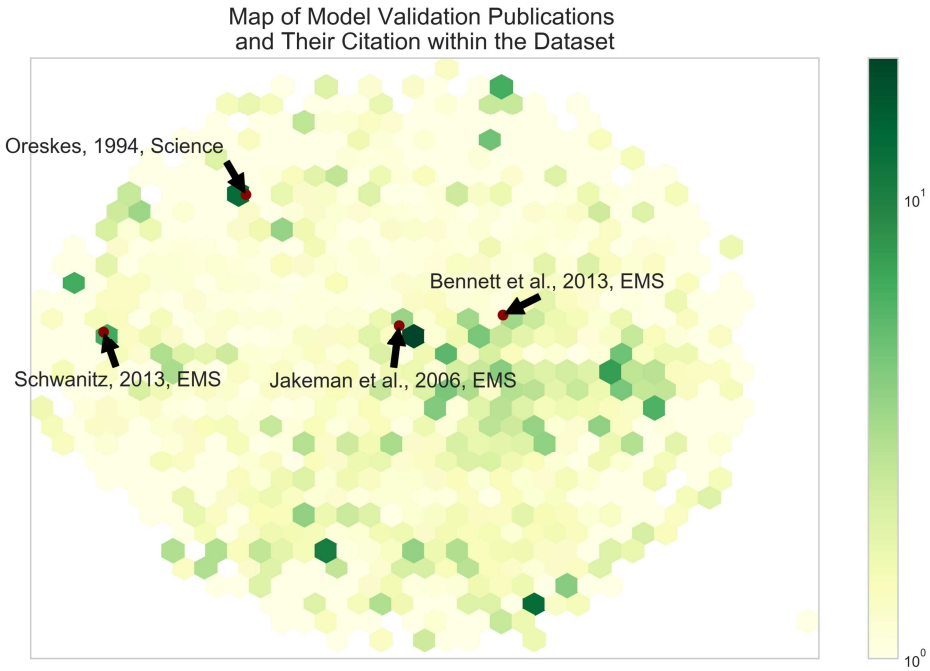
This visualization of publications raises two questions: Does the relatedness shown on this map reflect the citation scores of the articles? Do the density-based clusters on the map represent distinct topics? Figure 4 answers the first question by aligning the citation scores of the articles with their positions on the map. In Figure 4a and 4b, the density map shown in Figure 3 is divided into small hexagons. The color of each hexagon represents the average citation score of the articles falling into this hexagon. The darker the color, the higher the average citation score. Figure 4a visualizes the total number of citations recorded in the Scopus database, whereas Figure 4b is based on exclusive citation scores, i.e. the number of citations an article received only from the articles in our dataset.

According to Figure 4a, the densest regions of the map contain many highly cited articles, yet do not necessarily contain the most-cited ones. Instead, the most-cited articles are located rather in the periphery of the clusters. (See Figure A.2 for an alignment of Figure 3 and Figure 4a). If the peripheral articles are considered different in their content, it can be said that the most-cited articles tend to be different in their content and presumably innovative. Oreskes et al. (1994), which has 1699 citations on Scopus, fall into a highly-cited region in Figure 4a. Jakeman et al. (2006) and Bennett et al. (2013), of which citation scores on Scopus are 532 and 541 respectively, are in moderately cited regions.

The first observation on Figure 4b is the considerable reduction in citation scores. This implies that the articles in our dataset are cited mostly by the articles that are not included in this dataset, for instance the articles that might have applied a validation procedure but not necessarily used the terms such as *model validation* and *evaluation* in their title, abstract or keywords. Many of the dark regions in Figure 4a remain dark in Figure 4b. Hence, it can be said that the highly-cited articles are acknowledged not only in the general modeling literature but also in the specific validation literature. Oreskes et al. (1994) remain highly-cited in Figure 4b, while the relative citation scores of Jakeman et al. (2006), Bennett et al. (2013) and Schwanitz (2013) increase compared to Figure 4a. Therefore, it can be said that the latter two articles are highly recognized specifically in the model validation literature.



(a)



(b)

Figure 4: Map of the validation publications and their citation scores: (a) According to the total number of citations, (b) According to the number of citations only from the publications within our dataset

Main topics in the model validation literature

The second question raised by the density map in Figure 3 is whether the clusters on this map correspond to distinct topics. To answer this question, we first identify the main topics in our dataset as explained in the Methods section. The four main topics found by the topic-modeling algorithm are named as *Agriculture*, *Ecosystems*, *Hydrology*, and *Methods*, based on their most frequent and most descriptive words. The total topic probabilities are 17%, 16%, 26% and 40% for these topics, respectively, meaning, for example that, the total probability of all publications being associated with the *Ecosystems* topic is 16%. Figure A.3 illustrates the contents of these topics in terms of the most frequent, hence the most descriptive words they contain.

To investigate if the clusters on the density map correspond to these four topics, Figure 5 presents the map of the publications colored according to the topics they are associated with. In other words, each point in Figure 5 correspond to an article in our dataset, and its color represents the topic this article is associated with.

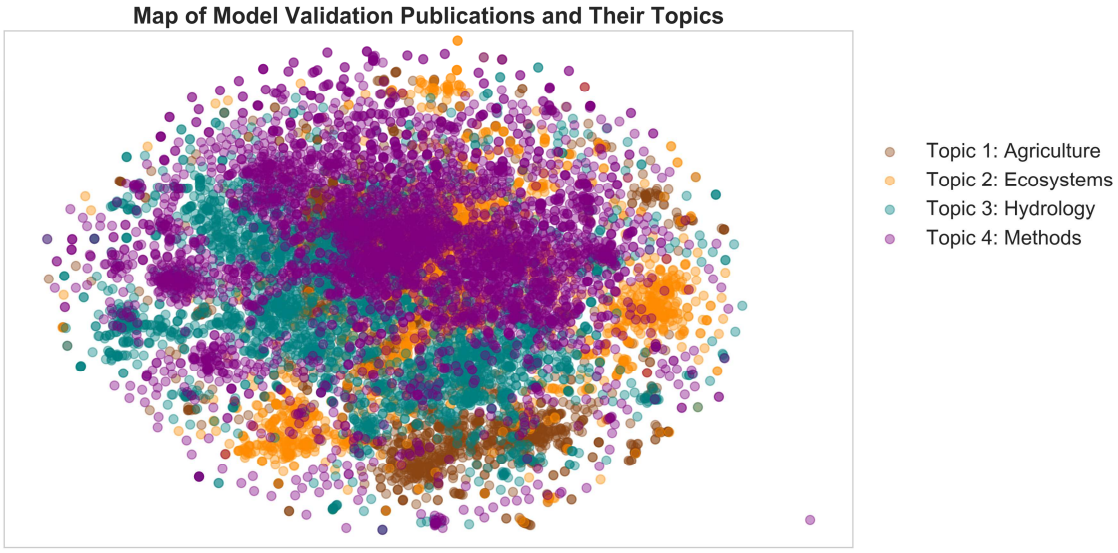


Figure 5: The map of the model validation publications colored according to the four main topics

Figure 5 shows that the four topics are not strictly distinct from each other on the map, and there are several overlaps. Still, the articles in the central dense region of Figure 3 belong mostly to the *Methods* topic, meaning that they have the highest resemblance to each other and most articles fall into this category. The clusters in the lower region of Figure 3 and Figure 5 are formed mostly by the *Ecosystems* and *Agriculture* publications. This means that the validation literature especially in the ecosystems and agriculture fields is distinctive from the others. This does not necessarily mean that the validation techniques in the *Ecosystems* or *Agriculture* field are different, since this analysis is based on the resemblance of word content, which can be attributed to the content that is unrelated to validation. Still, the compact clusters of these two topics indicate that the studies associated with them are clearly distinguished by the ones in other modelling fields. The publications in the *Hydrology* group are relatively dispersed, i.e. they do not form dense clusters. Located mostly at the lower part of the map, these publications can be said to have similarities with the Agriculture and Methods topics, yet they are quite dissimilar from the publications in the *Ecosystems* topic.

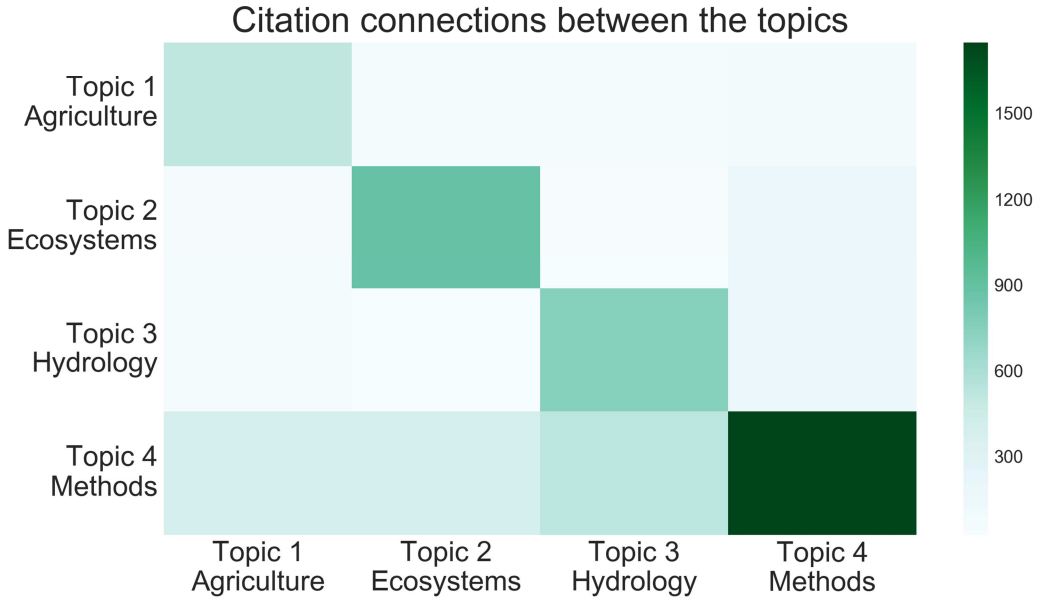


Figure 6: Number of citations between the main topics in the validation literature, from the topics in the rows to the topics in the columns

Having the content-based similarities and dissimilarities between the four main topics as discussed above, a complementary analysis can show whether these topics are related in terms of the citations between them. For such an analysis, we count the total number of citations made by the articles categorized in one topic to the articles in another topic. The grid in Figure 6 visualizes the results, where each cell is colored according to the total number of citations from the articles in the row's topic to the articles in the column's topic. Figure 6 shows that the articles in each topic cite the articles in the same topic most. This tendency of topic categories to self-citation indicates that the validation literatures of these modeling areas are closed to each other. In other words, they do not acknowledge each other in terms of widespread cross-citations, and they are not considerably connected when citation score is a proxy for connectedness. Furthermore, the highest number of citations are between the articles in the *Methods* topic. This can be explained not only by the high resemblance and relatedness of the articles (based on Figure 5), but also by the high number of articles in this category.

4 Discussion

This paper presents an overview of the model validation literature based on a combination of bibliometric and text-mining analyses. We are interested in the validation of environmental and economic models used in various decision-making contexts. Therefore, our analysis is on a large dataset of more than 10,000 publications from various fields related to sustainability science such as environmental science, economics, energy, social sciences and decision sciences. This breadth of the dataset is helpful in covering general issues in model validation, as well as similarities and differences between the validation practices in different modeling fields. However, such an analysis can as well be conducted on more customized publication datasets to obtain information about specific fields, such as only hydrological modeling or decision sciences.

The mapping of publications in terms of the similarity of their contents, where similarity is defined by the commonality of words in their abstracts, resulted in several clusters of work in different sizes (Figure 3). The most-cited publications, however, were not in the centers of these clusters but rather in the peripheries (Figure 4 and Figure A.2 in Appendix). Therefore, it can be said that the most-cited and most widely acknowledged publications in the model validation literature are not the ones that are highly similar to a large body of work, but the ones that are different from the majority, and presumably innovative. Oreskes et al. (1994) is an example of this, because they discordantly argue that validation based on representation accuracy is impossible. This argument is based on the idea that a match between the model output and observational data does not demonstrate the reliability of a model or hypothesis, it only supports its probability. Therefore, since models can never accurately represent reality, they should not be used for predicting the future but for sensitivity analyses, exploring what-if scenarios, and for challenging our biases and assumptions.

Based on its high citation score, this view of Oreskes et al. (1994) is widely acknowledged, yet might not be followed in practice. Our results show that the most common words in the abstracts of model validation publications are *data* and *predict*, and most of these publications were published after Oreskes et al. (1994). (See Figure A.1 for the number of publications in each year in our dataset.) This finding can be interpreted as the prevalence of a prediction-oriented modeling, i.e. models being used to predict the future as opposed to the view of Oreskes et al. (1994) on using them to explore scenarios or to test different assumptions. Furthermore, the validation practice seems to be strongly associated with data. This analysis alone cannot definitively conclude that the common validation techniques are based on historical data. Yet, it can conclude that data is heavily emphasized in the validation literature, indicating that the validity is related to the representation of reality and replicating empirical data. Therefore, statistical and methodological reliability can be said to be the main concern of validation practice. When the content of individual articles are scanned, the words that relate to public reliability, such as *stakeholder*, *user*, *decision-maker*, *credibility* appear, for instance in the exemplary articles studied (Jakeman et al., 2006; Schwanitz, 2013). However, they are not common in the larger literature, and do not appear among the frequent words.

The prominent role of data-based approaches in validation is shown by Eker et al. (2018), who investigated the practitioners' view on validation. Practitioners report that the comparison of model output and historical data is one of the most commonly used techniques, and a match between the output and data is a reliable indicator of a model's predictive power. Furthermore, a large majority of practitioners participated in (Eker et al., 2018)'s study disagree that models cannot be used for prediction purposes, indicating a strong support for using models to predict.

The clusters observed in the mapping of publications could be partially explained by their topics. These topics identified by a text-mining analysis correspond to the main areas of sustainability science in our case, such as *Ecosystems*, *Agriculture* and *Hydrology*, as well as a general *Methods* topic. Among these groups of publications, especially the ecosystems and agriculture/land use studies were distinct from the others. A more striking distinction between the topics is in terms of the number cross-citations between them. The publications in each topic cite mostly the publications in their own topic. This analysis cannot conclude on the context of citations, therefore we cannot say if the citation scores indicate the sharing of validation approaches. Still, since the dataset is constituted by the validation literature, this finding

indicates that the validation research in other fields is acknowledged relatively less. This finding supports the previous finding that the validation literatures of different fields are distinct from each other, and may not be benefitting from each other effectively.

These findings lead to two main recommendations for future research. The prevalence of words like *data* and *predict* indicate a strong focus on statistical and methodological reliability. There is no indication among the most frequent words about public reliability, which relates to the acceptance of model-based conclusions by decision-makers and stakeholders. Therefore, future research can further investigate how public reliability is addressed in the broad model validation literature. Future research can also extend validation approaches beyond data-oriented reliability to public reliability. Secondly, since different areas of environmental modeling, such as hydrological, ecosystem and agricultural modeling are found to be distinct in terms of not only contents but also cross-citations, future studies can synthesize the methods and views from various areas. Such an integration can enhance the methods and create a coherent validation practice.

5 Conclusion

This paper investigated the model validation practice across a large body of scientific publications by adopting several data analysis techniques. This overview of model validation literature led to a number of conclusions: Firstly, *data* plays an important role in the current validation practice, appearing as the most frequent word in publications. This is considered a prevalent discussion of statistical and methodological reliability ensured by data-driven techniques. Yet, whether the practice relies on data-driven validation methods cannot be concluded based on this analysis. Secondly, the most-cited publications on model validation are the ones that do not strongly resemble the others in content, where resemblance is defined based on the commonality of words in the abstracts. In other words, different and presumably innovative publications, which appeal to a wider scientific audience, are acknowledged more. Thirdly, the validation literature in the main areas of environmental modeling, such as hydrology, ecosystems and agriculture, are distinct from each other as indicated by their contents, and not strongly connected to each other when cross-citation scores between the fields is considered as a proxy for connectedness.

The current validation practice is strong in ensuring statistical and methodological reliability. Therefore, future studies can provide a deeper analysis on how public reliability addressed in the current validation practice. Furthermore, future validation studies can focus on soft and participatory approaches to establish public reliability, in order to enhance the acceptance and adoption of model-based conclusions in decision-making contexts. Future validation studies in any area of environmental modeling, such as hydrological, ecosystem and energy systems modeling, can also benefit from the validation approaches in other fields. A synthesis of methods, views and experiences from various fields can strengthen the model validation practice in line with the requirements of future decision-making challenges.

Acknowledgements: The research was funded by IIASA and its National Member Organizations in Africa, the Americas, Asia, and Europe.

References

Augusiak, J., Van den Brink, P.J., Grimm, V., 2014. Merging validation and evaluation of ecological models to 'evaludation': a review of terminology and a practical approach. *Ecological Modelling* 280 117-128.

Balci, O., 1994. Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operations Research* 53(1) 121-173.

Barlas, Y., 1996. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12(3) 183-210.

Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K., 2010. Validation of biophysical models: issues and methodologies. A review. *Agronomy for Sustainable Development* 30(1) 109-130.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. *Environmental Modelling & Software* 40 1-20.

Beven, K., 2002. Towards a Coherent Philosophy for Modelling the Environment. *Proceedings: Mathematical, Physical and Engineering Sciences* 458(2026) 2465-2484.

Bird, S., Loper, E., 2004. NLTK: the natural language toolkit, Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, p. 31.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan) 993-1022.

Broadus, R., 1987. Toward a definition of "bibliometrics". *Scientometrics* 12(5-6) 373-379.

Cancino, C., Merigo, J.M., Coronado, F., Dessouky, Y., Dessouky, M., 2017. Forty years of computers & industrial engineering: a bibliometric analysis. *Computers & Industrial Engineering* 113 614-629.

Cunningham, S.W., Kwakkel, J.H., 2016. Analytics and Tech Mining for Engineering Managers. Momentum Press.

van der Sluijs, J.P., 2002. A way out of the credibility crisis of models used in integrated environmental assessment. *Futures* 34(2) 133-146.

van der Sluijs, J.P., Petersen, A.C., Janssen, P.H.M., Risbey, J.S., Ravetz, J.R., 2008. Exploring the quality of evidence for complex and contested policy decisions. *Environmental Research Letters* 3(2) 024008.

van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84(2) 523-538.

Eker, S., Rovenskaya, E., Obersteiner, M., Langan, S., 2018. Practice and perspectives in the validation of resource management models. *Nature Communications* 9(1) 5359.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software* 21(5) 602-614.

Laengle, S., Merigó, J.M., Miranda, J., Słowiński, R., Bomze, I., Borgonovo, E., Dyson, R.G., Oliveira, J.F., Teunter, R., 2017. Forty years of the European Journal of Operational Research: A bibliometric overview. *European Journal of Operational Research* 262(3) 803-816.

Landry, M., Malouin, J.-L., Oral, M., 1983. Model validation in operations research. *European Journal of Operational Research* 14(3) 207-220.
lda Developers, 2014. lda: Topic modeling with latent Dirichlet Allocation. <https://pythonhosted.org/lda/>.
Maaten, L.v.d., 2014. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* 15(1) 3221-3245.
Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine Learning research* 9(Nov) 2579-2605.
Maslin, M., Austin, P., 2012. Uncertainty: Climate models at their limit? *Nature* 486(7402) 183-184.
Matott, L.S., Babendreier, J.E., Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. *Water Resources Research* 45(6).
Merigó, J.M., Pedrycz, W., Weber, R., de la Sotta, C., 2018. Fifty years of Information Sciences: A bibliometric overview. *Information Sciences* 432 245-268.
Oreskes, N., Belitz, K., 2001. Philosophical issues in model assessment, In: Anderson, M.G., Bates, P.D. (Eds.), *Model validation: Perspectives in hydrological science*. John Wiley and Sons.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263(5147) 641-646.
Pilkey, O.H., Pilkey-Jarvis, L., 2007. Useless Arithmetic: Why Environmental Scientists Can't Predict the Future? Columbia University Press, New York, USA.
Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholten, H., Kassahun, A., 2005. Quality assurance in model based water management – review of existing practice and outline of new approaches. *Environmental Modelling & Software* 20(10) 1201-1215.
Risbey, J., van der Sluijs, J., Kloprogge, P., Ravetz, J., Funtowicz, S., Corral Quintana, S., 2005. Application of a checklist for quality assistance in environmental modelling to an energy model. *Environmental Modeling & Assessment* 10(1) 63-79.
Saltelli, A., Funtowicz, S., 2014. When all models are wrong. *Issues in Science and Technology* 30(2) 79-85.
Schwanitz, V.J., 2013. Evaluating integrated assessment models of global climate change. *Environmental Modelling & Software* 50 120-131.
Smith, L.A., Petersen, A.C., 2014. Variations on reliability: Connecting climate predictions to climate policy, In: Boumans, M., Hon, G., Petersen, A.C. (Eds.), *Error and Uncertainty in Scientific Practice*. Pickering & Chatto: London.
Tsioptsias, N., Tako, A., Robinson, S., 2016. Model validation and testing in simulation: a literature review, OASIcs-OpenAccess Series in Informatics. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
van Vliet, J., Bregt, A.K., Brown, D.G., van Delden, H., Heckbert, S., Verburg, P.H., 2016. A review of current calibration and validation practices in land-change modeling. *Environmental Modelling & Software* 82 174-182.

Appendix

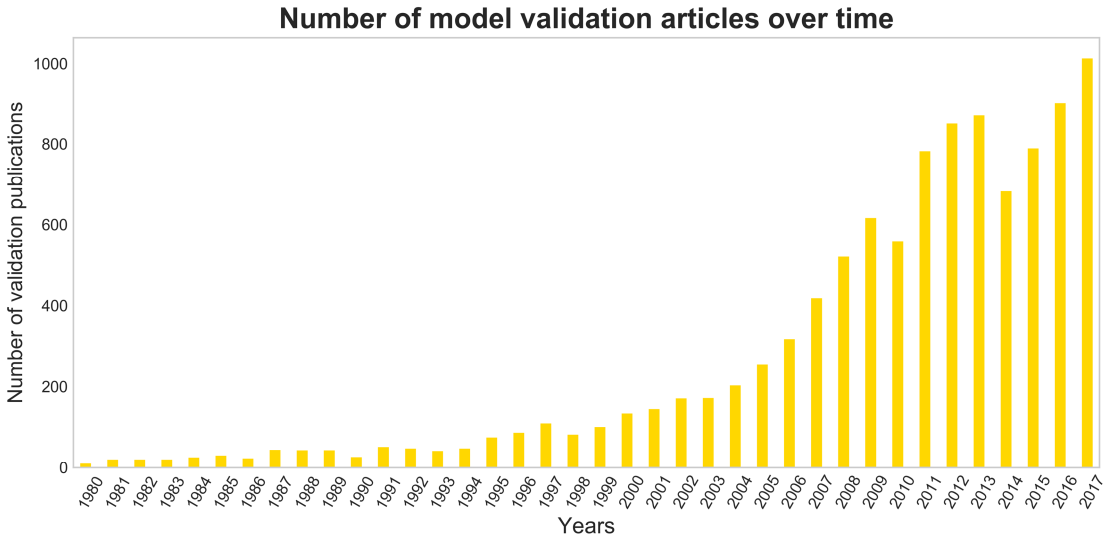


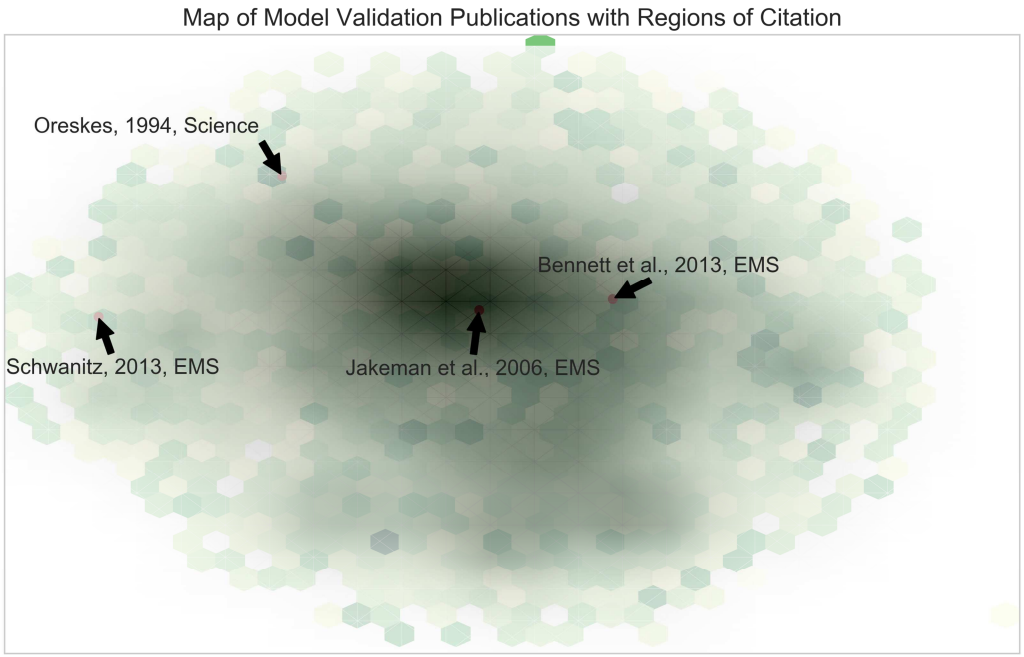
Figure A.1: The distribution of model validation articles across the years 1980-2017

Table A. 1: Word lists of the four exemplary articles in the order of decreasing frequency

Oreskes et al. 1994	**Bennett et al. 2013**	**Jakeman et al. 2006**	**Schwanitz 2013**
confirm	data	natur	evalu
natur	environment	practic	behavior
predict	method	review	global
system	valu	test	iam
imposs	characteris	resourc	climat
primari	nonmodel	manag	chang
evalu	confid	step	system
access	level	strong	document
demonstr	calibr	data	commun
observ	field	user	framework
agreement	key	limit	discuss
partial	establish	disciplin	use
verif	vital	client	natur
complet	model'	scope	step
incomplet	techniqu	credibl	tool
numer	order	end	public
preclud	depend	support	integr
consequ	real	stage	complex
rel	aim	basic	experi
phenomena	visual	featur	build
question	qualit	rang	assess
open	systemat	altern	uncertainti
close	problem	report	human
heurist	comparison	improv	import
nonuniqu	procedur	applic	offer
inher	detect	peopl	test
logic	effect	identifi	standard
affirm	suggest	choic	histor
term	select	provid	establish
	test	quantit	futur
	observ	make	understand

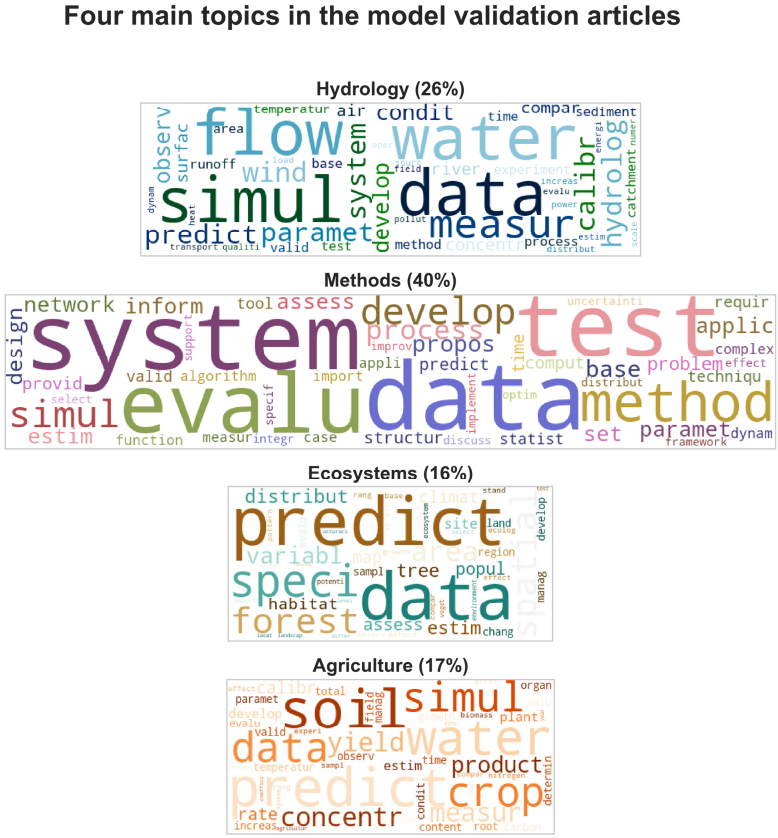
	criteria	purpos	demonstr
	requir	aim	systemat
	evalu	trend	process
	base	interest	observ
	implement	techniqu	problem
	direct	process	verif
	purpos	util	model'
	reassess	calibr	styliz
	focu	accuraci	set
	pattern	critic	sensit
	overview	incorpor	miss
	scale	object	stepbystep
	consider	discuss	advis
	diverg	sceptic	plausibl
	element	assumpt	policymak
	gain	prior	overcom
	preserv	awar	stakehold
	tailor	famili	urgent
	workflow	increasingli	transpar
	decisionmak	outlin	answer
	indirect	justifi	way
	advanc	confront	open
	transform	ten	question
	coupl	quantiti	fundament
	discuss	revis	code
	scope	impli	insight
	numer	modelbuild	element
	paramet	rational	reflect
	manag	statement	wide
	basic	parti	pattern
	combin	open	unknown
	graphic	wider	deriv
	behaviour	entail	inform
	practic	broader	hierarchi
	handl	inform	exampl
	class	construct	challeng
	review	exercis	conceptu
	metric	learn	
	inform	endus	
		qualiti	
		right	
		reserv	
		encompass	
		background	
		reli	
		knowledg	
		document	
		constitut	
		partnership	
		develop	

526



527

528 *Figure A.2: The density (Figure 3) and citation score (Figure 4a) maps of the model validation articles overlaid*



529

530 *Figure A.3: The four main topics and their content in the model validation publications*

531 The LDA algorithm used in this study to identify the main topics in the validation literature
532 allocates each publication to a topic with a calculated probability. This figure visualizes these

topic probabilities, where each line represents a document. The darker this line in the corresponding topics' segment (column), the higher the probability. Having heterogeneity across the columns in these figures indicate that the topics identified by the algorithm are distinct from each other.

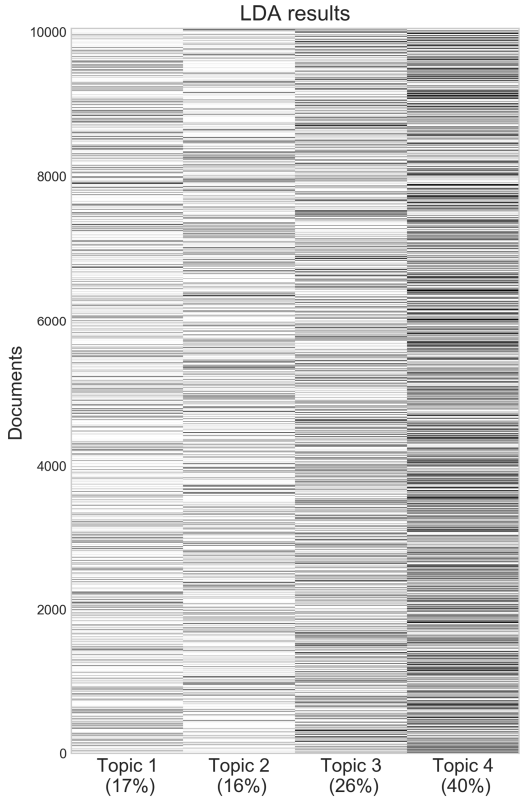


Figure A. 4: Document-topic pairs resulting from the LDA implementation for topic modelling