24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

Application of topic modelling to Integrated Water Resource Assessment domain

M.G. Erechtchoukova [a], **A.K. Kohli** [a]

[a] *School of Information Technology, Faculty of Liberal Arts and Professional Studies, York University, Canada*
Email: marina@yorku.ca

Abstract: Integrated Water Resource Assessment (IWRA) is a multidisciplinary analysis of cause-effect interactions of environmental, social, and economic processes which play an important role in an investigated scenario or undertaking related to a water resource. The goal of this analysis is to generate new knowledge in support of decision or policy making. The concept of IWRA covering natural processes in a water resource with at least one of the areas of economic or social aspects of modern society had been articulated almost a half a century ago. Nowadays, it became mature to the extent of supporting environmental sustainability and promoting UN Sustainable Development Goals through offering well-accepted approaches, frameworks, simulation models and computational techniques upholding the assessment. Nevertheless, there is steady interest to the issues of IWRA among researchers and practitioners because new technologies open opportunities for advanced computational techniques and comprehensive analysis. This study presents exploratory analysis of the corpus of scientific publications using text mining techniques with the aim to identify salient topics and potential gaps in the IWRA research area. The analysis was conducted based on the topic modelling approach. Topic modelling is a form of text mining that allows to find a representation of information from a collection of documents called corpus. Any text document can be viewed as a collection of several themes which are present in the document and reflect the document contents in a meaningful to its readers way. A theme or a topic is represented via an array of words that have a high tendency of co-occurrence when a particular theme underlying a document is being discussed. The most salient characteristic of topic models is that they automate the process of extracting these underlying (latent) themes in large corpora of texts without any human intervention excluding text pre-processing. Given that a topic model operates with a fixed vocabulary, domain specific analysis is expected to be more informative. Therefore, careful selection of documents included into a corpus is required. Application of topic modelling to multidisciplinary areas such as IWRA carries more importance because it helps to automate the process of extraction of salient topics relevant to a document and categorize the documents into themes for targeted analysis and knowledge extraction. The corpus of abstracts of 89726 papers published from 1970 to 2020 in peer-reviewed journals representing leading outlets in the areas of water resources and integrated environmental assessment was assembled. It was analysed using basic bibliometric statistics. After that, the corpus was pre-processed following conventional topic modelling framework and fed into LDA mallet algorithm to identify salient topics. Hyperparameters of the selected topic modelling algorithm were identified based on exploratory computations and evaluation of several topic models performance using a coherence score and qualitative evaluation of the identified topics. The model producing 20 topics was considered satisficing and used as a basis for the qualitative analysis of clusters of words forming topics. The analysis revealed two categories of latent topics presented in the corpus: methodological and environmental. The latter describes various aspects of utilization, protection, and restoration of a natural water resources. No theme reflecting assessment of socio-economic processes was uncovered despite the fact, that these processes play critical role in the environmental state of a water resource.

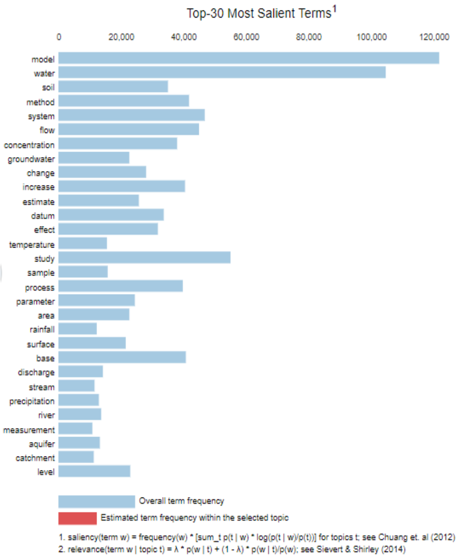

Figure 1. Top 30 salient terms

Keywords: *Topic modelling, LDA mallet algorithm, Integrated Water Resource Assessment, text mining*

1. INTRODUCTION

Integrated Water Resource Assessment (IWRA) is an integral part of Integrated Environmental Assessment which can be defined as an interdisciplinary analysis of cause-effect interactions of environmental and socio-economic processes relevant to an investigated problem with the aim to generate knowledge in support of decision and/or policy making (Rotmans and Dowlatabadi 1997). Even though IWRA deals with water related problems, it remains complex and multi-disciplinary due to the crucial role of water in sustaining humans and wildlife, economic and social activities. The concept of integrated assessment covering natural processes in a water resource with at least one of the areas of economic or social aspects of modern society had been articulated almost a half of a century ago (Toth and Hizsnyik 1998), and had matured enough to offer well-accepted approaches, frameworks, simulation models and computational techniques supporting the assessment (Kelly et al. 2013). Nowadays IWRA should be conducted in a way supporting environmental sustainability and promoting UN Sustainable Development Goals (UN 2015). This requirement adds complexity to the assessment process due to the necessity to evaluate not only immediate effects of a policy or an undertaking, but its future impact on the environment and society.

The significance of IWRA is supported by the growing research activities in this area offering enhanced methodological approaches and computational techniques for their implementation. There were several attempts to determine the up-to-date status of the research field through identification its dominant perspectives. Kelly et.al (2013) analyzed five modelling approaches supporting IWRA and suggested a framework for selection of the most appropriate one. Comprehensive analysis of methodological and applied aspects of Integrated Environmental Assessment and identified key dimensions of the assessment and modelling were proposed in (Hamilton et al. 2015). Li and Zhao (2015) conducted bibliometric analysis of publications on Environmental Assessment which were indexed by Web of Science. The study revealed patterns in distribution of articles and overall trends at global and regional levels. Bibliometric analysis was applied to 11,000 papers on Integrated Water Assessment and Modelling to investigate landscape of this research field through scientific publications (Zare et al. 2017). The study identified trending themes, commonly used keywords, and the most influential papers.

Bibliometrics provides general quantitative characteristics of publications reflecting major categorization, keywords, geographic distribution of the papers and their popularity among research communities. Notwithstanding the importance of these studies, it is necessary to stress out that the quantitative nature of the analysis focused on a few keywords and other articles' meta-data limits substantially critical scrutiny of the corpus of published works. At the same time, Natural Language Processing (NLP) techniques have been successfully applied for automated text summarization in various knowledge domains with very few studies related to Environmental Sustainability. Thus, Cheng et al. (2018), applied topic modelling technique to investigate the domain of Ecology, Environment and Poverty Nexus. Nine dominating topics were identified in the corpus of 4335 papers selected using 'ecology, environment and poverty' keyword. Topic modelling analysis of Hydropower domain was performed by Jiang et al. (2016) using Topicmodels R package.

This paper presents exploratory analysis of the corpus of scientific publications using topic modelling techniques with the aim to identify salient topics and potential gaps in the IWRA research area. The paper provides a brief overview of the techniques used, describes the way the corpus was constructed, the analysis performed and the obtained results.

2. BACKGROUND

NLP is a branch of Artificial Intelligence that automatically processes human language to understand meaning of speech or text and convey conclusions. One of the branches of NLP deals with semantic or text mining. The main aim of text mining is to extract meaning from unstructured text. Topic modelling is a form of text mining that allows to find a representation of information from a collection of documents called corpus. Any text document can be viewed as a collection of several themes which are present in the document and reflect the document contents meaningful to its readers. A theme or a topic is represented via an array of words that have a high tendency of co-occurrence when a particular theme underlying a document is being discussed. The most salient characteristic of topic models is that they automate the process of extracting these underlying (latent) themes in large corpora of texts without any human intervention excluding text pre-processing (Blei 2012; Mohr and Bogdanov 2013). The distribution of topics for each document from the corpus can also be identified. This helps the researchers to understand whether the document had one dominating latent theme or multiple latent themes. Topic modelling becomes of paramount importance to a vast expanse of studies in diverse fields ranging from sociology, humanities to natural sciences due to its ability to automatically determine document relevance to a subject of investigation and correct categorization of large collections of documents.

There are several algorithms and their implementations available for topic modelling. Topic modelling techniques are classified into Non-Probabilistic (Algebraic) and Probabilistic Approaches (Kherwa and Bansal, 2020). Non-Negative Matrix Factorization (NNMF) and Latent Semantic Analysis (LSA) identify the underlying themes by performing dimensionality reduction using algebraic methods (Deerwester et al. 1990; Lee and Seung 1999). NNMF and LSA utilize the Bag of Words (BoW) methodology in which the text is converted into the term - document matrix and the sequence in which the words appeared in the document is neglected. Only the term frequency is taken into account. Probabilistic topic modelling considers document as an output of a random generative process, in which each topic is represented as a distribution of its words over a fixed vocabulary (Blei, 2012). Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) is one of the most popular probabilistic topic modelling algorithms.

Given that a topic model operates with a fixed vocabulary, domain specific analysis is expected to be more informative. Therefore, careful selection of documents included into a corpus is required. Application of topic modelling to multidisciplinary areas such as IWRA carries more importance because it helps to automate the process of extraction of salient topics relevant to a document and categorize the documents into themes for targeted analysis and knowledge extraction. At the same time, topic modelling of a multidisciplinary area such as IWRA is even more challenging.

3. FRAMEWORK FOR TOPIC MODELLING

In general, the quality of identified topics depends on the data sets, e.g., corpora of documents, and the text pre-processing step. A generic framework for topic modelling consists of six steps presented in Figure 1.

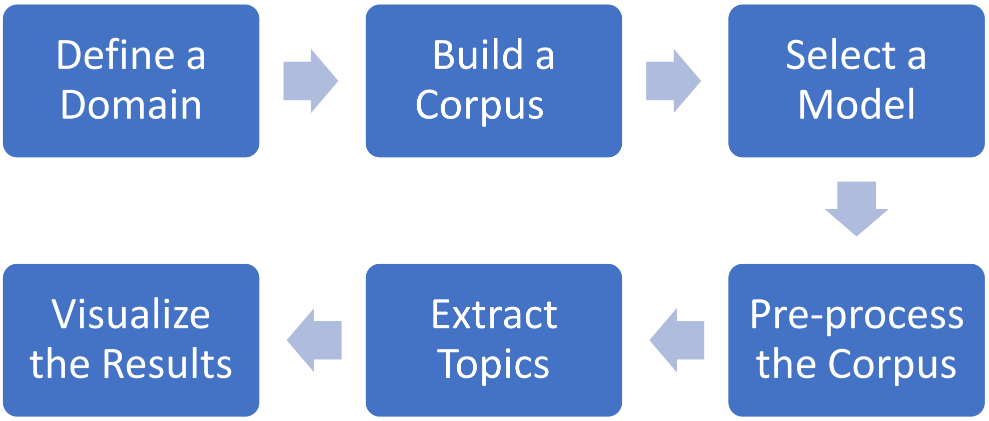

Figure 1. Conventional Framework for Topic Modelling

This framework is usually applied to short text documents, e.g., online tweets or brief reviews. In many cases all documents forming a corpus are selected from the same source, e.g., a review system or online thematic forum. Rarely, it is used to process abstracts of research publications with the length not exceeding 500 words. In majority of cases, all abstracts come from the same outlet. It can be explained by technical limitations making automated document downloading from the sources with different layout and formatting very challenging. In addition, topic modelling requires to process high dimensional matrices which actual size depends on the size of the corpus. Standard computer resources may not be sufficient for such computations.

This framework were followed in the current study. The specifics of its application are described further. Well-defined research domain helps to identify sources and selection criteria for relevant and credible documents. Often, documents are selected to a corpus based on specific words which may appear in the title, abstract or keywords of a publication. Given that the goal of the current study was to obtain a general picture on the research area in IWRA including themes currently dominating, other selection criteria were employed. To guarantee credibility of the corpus, online sources were restricted to peer-reviewed journals. Among many highly reputable journals only those which scope covers all or some aspects of IWRA were selected. Abstracts of all papers published in these journals over the last 50 years were extracted and analysed.

Selection of a topic modelling techniques is an important step of the analysis affecting further steps of the framework. A probabilistic approach was adopted in this study with an attempt to take into account not only frequencies of terms, but their co-occurrence as well. The analysis was conducted using LDA mallet algorithm and its Python implementation in Gensim package (Rehurek and Sojka 2011). The selected algorithm requires text in BOW representation. Further transformation into tf-idf (e.g., Sammut and Webb 2011) representation is not required.

4. PRELIMINARY DATA ANALYSIS

The conventional framework for topic modelling was applied to 14 peer-reviewed journals representing leading outlets in the areas of water resources and integrated environmental assessment. A corpus of abstracts of all papers published from 1970 to 2020 in these journals was assembled. Journal titles and document distribution among journals are presented in Figure 2.

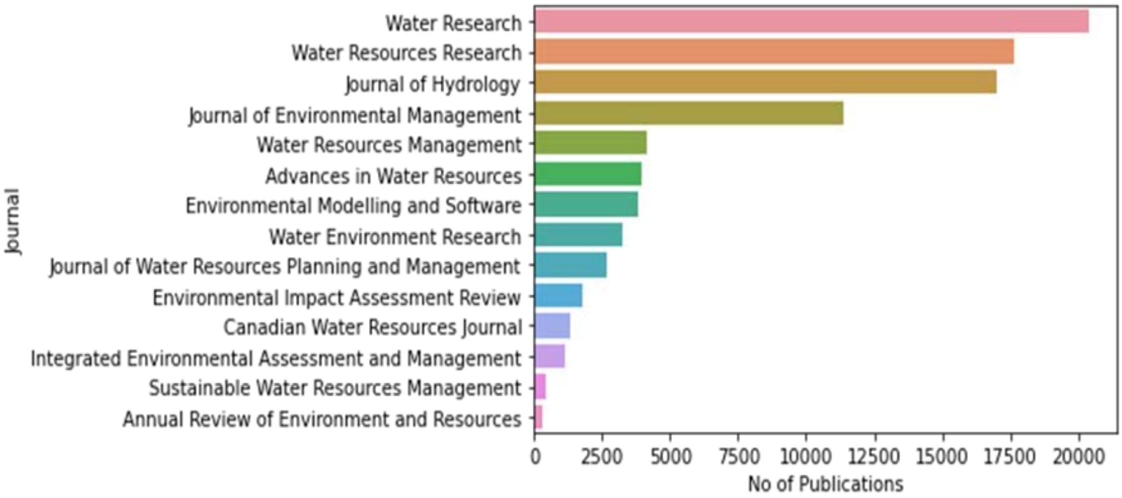

Figure 2. Selected journals and document distribution among them

The corpus consists of 89726 documents. The number of publications on IWRA is growing with significant increase in recent years. This increase is a testament of a raising awareness of environmental and water related issues. Document distribution over time clear indicates exponential grows of publications relevant to IWRA (Figure 3).

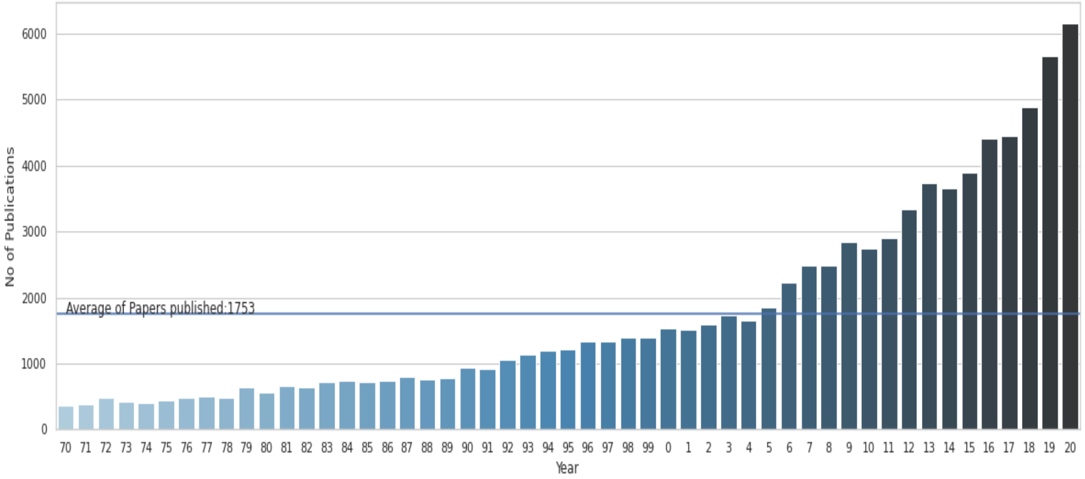

Figure 3. Document distribution over time in the constructed corpus

5. EXPLORATORY COMPUTATIONS

The corpus has been pre-processed using standard techniques: tokenization, stop word removal, filtration, and lemmatization. After that, the corpus was fed into the selected algorithm to identify clusters of tokens representing latent themes. The LDA algorithm has several hyperparameters which can be tuned for better performance. The number of the most salient topics which will be uncovered via computations is the most important one. Obviously, this number was not known beforehand and was determined through computational experiments. The goal of these experiments was to determine the values of the number of identified topics which supports 'good-enough' algorithm performance. The algorithm performance was evaluated based on formalized measure – coherence score (values) and the qualitative analysis of topics generated during each run. The model performance was considered 'good-enough', if further increase in the number of topics did not lead to notable improvement the coherence score. The details of these experiments can be found in (Kohli 2021). The results of these computations suggested that 20 topics are sufficient to ensure that coherence score of the uncovered topics is high enough (Figure 4). At the same time, each identified topic represents a distinct specific area of IWRA.

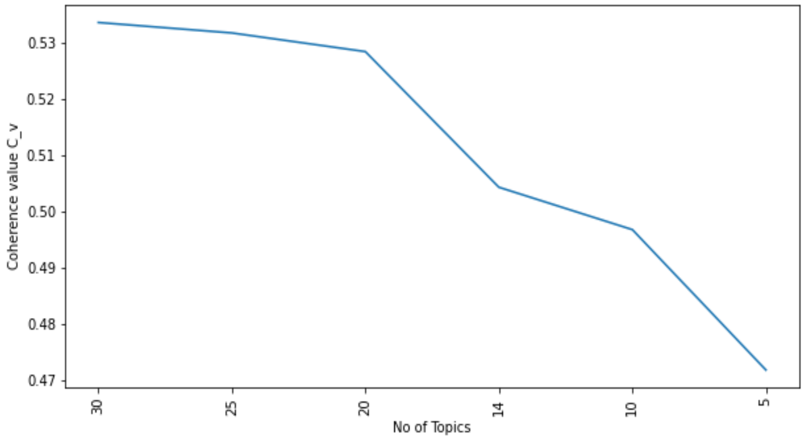

Figure 4. The number of topics evaluation

The informativeness of extracted clusters of words was analysed based on the authors' expert opinion. To support topic interpretation and comparison, topic visualization was performed using pyLDAvis. This is the python extension of the original LDAvis tool implemented in R programming language (Sievert and Shirley 2014).

Each of the 20 identified topics is represented as clusters of words. From each topic, top 30 words with the highest probability were used for the qualitative analysis and interpretation. In addition, top 30 of the most salient words and their distribution in the corpus were analysed (Figure 1).

6. DISCUSSION AND CONCLUSION

Analysis of the top 30 words of each topic suggested that all identified topics can be classified into methodological themes and areas of investigation. Topics 4, 9, 10, 16, and 18 belong to the methodological category covering applied frameworks and approaches to modelling and data analysis. Topics 18, 16, 4, and 9 notably dominate other topics identified in the corpus. The most salient topic 18 is devoted to methodological approaches for investigation of multifactor impact on water resources. Topic 16 covers parameterization of natural processes. Topic 4 represents simulation modelling theme (Figure 5), while topic 9 deals with the analysis of observation data. All other topics describe various subdomains representing different aspects of IWRA reflected in the corpus. These subdomains deal with remediation of water pollution, wastewater treatment, water sanitation, soil moisture, Water-Energy nexus, water flow modelling, sediment pollution, water salinization, optimization of reservoir operations, sediment transport, water in agriculture, aquatic toxicology, extreme hydrological events, and land use change. Aquatic toxicology topic appeared to be least represented in the corpus, followed by aquatic microbiology and Water-Energy nexus. Some topics overlap with others from the same or different categories. Overlap of topics from different categories justifies the fact that analysed publications described methodological approaches and their application to specific areas of IWRA. Where word distribution patterns representing different themes partially coincide, it suggests that the published studies were interdisciplinary.

It is worth noting that methodological theme 4 pointing to simulation modelling approach and theme 9 describing approaches to data analysis have significant overlap between each other and with the topic 16 corresponding to parameterization of natural processes. This suggests that model development was conducted following both deterministic process-based and statistical data-driven approaches.

Topic 10 from the methodological category represents frameworks and approaches to policy and decision support. Although the very idea of the concept of ‘integrated assessment’ assumes integration with the policy or alternative decisions assessment, this theme has no overlap with any other dominating topic identified in the corpus.

The most salient word used in the investigated corpus is ‘model’ which implies that ‘modelling’ was the most popular methodological approach over the last 50 years. ‘Modelling’ is a fundamental concept of the classical universal framework for problem solving introduced in the Information Science. Top 30 salient words presented in Figure 1 indicate that majority of studies were conducted to imitate natural processes in water objects and on their catchments using observation data. Even though modelling techniques were developed to support decision and/or policy making, and the very idea of the concept of ‘integrated assessment’ assumes integration with the policy or alternative decisions assessment, words representing this theme did not appear on the top terms list.

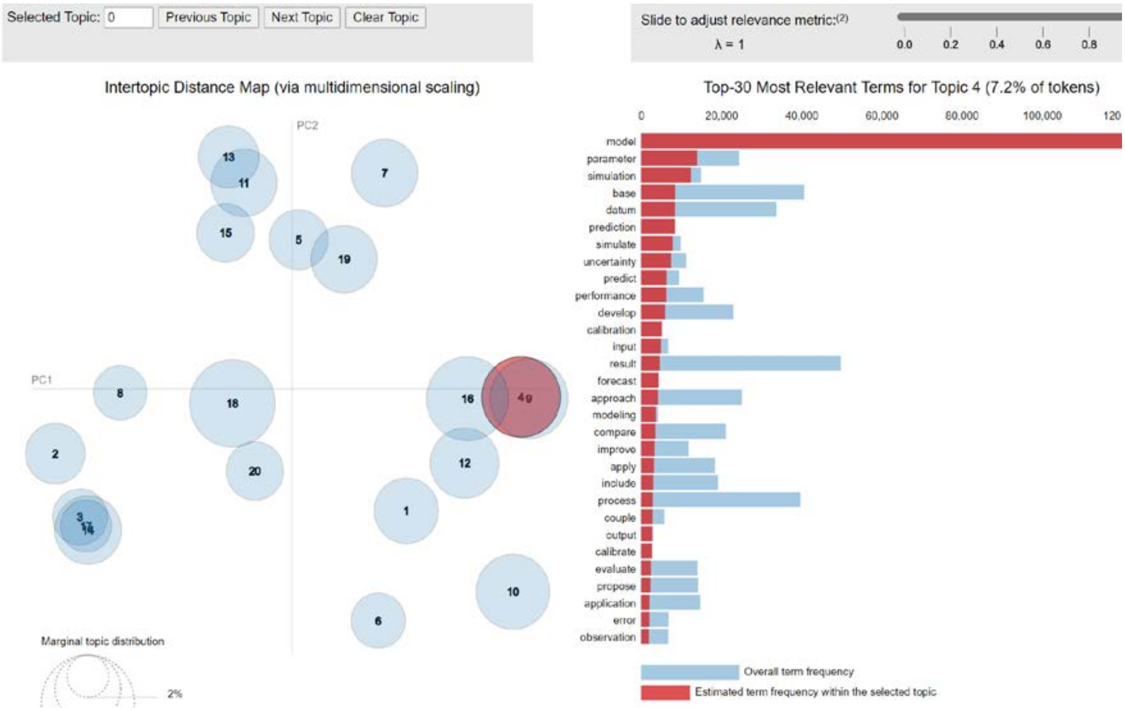

Figure 5. Salient topic representation via dominating terms

The conducted topic modelling allowed to identify the most represented subdomains of IWRA. The analysis demonstrated that solid expansive knowledge on IWRA is represented by scientifically valid frameworks and methodologies supported by reliable mathematical and computational tools that model natural processes and can be used in evaluation of management scenarios relevant to different aspects of IWRA. At the same time, no theme reflecting formalized assessment of socio-economic processes was uncovered despite the fact that these processes play critical role in the impact assessment of a policy or undertaking on the environmental state of a water resource. Apparently, the investigated publications describe frameworks for policy assessment and decision making at the higher level of abstraction while it was expected that the majority of modelling exercises would be implemented to provide policy and decision makers with important and reliable information including automated support for evaluation of alternative decisions. The analysis showed solid interest of the research community to this theme. However, relatively few success stories on end-to-end integration of a policy assessment framework with the real-world problem of the evaluation of the policy impact on the aquatic state were reported.

It should be noted that further improvement of corpus pre-processing and modelling techniques may identify more informative topics. The study is currently conducted in this direction. Nevertheless, the presented results of exploratory computations clearly suggest that the multifaceted nature of IWRA problems is yet to be addressed with the full integration of all levels of the assessment.

The presented topics were found in the corpus covering more than 50-year period. Even though the preliminary analysis of the corpus of abstracts determined the temporal distribution of the number of publications and revealed the exploding numbers of publications on IWRA in the recent years, the experiments conducted do not allow for identification of temporal trend in dominating topics. The identification of such trend and changes in dominating latent topics present in modern publications are in the focus of further investigations.

ACKNOWLEDGMENTS

The authors are grateful to editors and anonymous reviewers for their thoughtful suggestions and helpful comments on the manuscript improvement.

REFERENCES

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Blei, D. M., 2012. Probabilistic topic models. Communications of the ACM, 55(4):77–84.

Cheng, X., Shuai, C., Liu, J., Wang, J., Liu, Y., Li, W., Shuai, J., 2018. Topic modelling of ecology, environment, and poverty nexus: An integrated framework. Agriculture, Ecosystems & Environment, 267, 1–14.

Deerwester, S., Furnas, G.W., Landauer, T.K, 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Hamilton, S.H., Guillaume, J., Elsawah, S., Jakeman, A.J., Pierce, S.A., 2015. Integrated assessment and modelling: overview and synthesis of salient dimensions. Environmental Modelling and Software 64, 215–229.

Jiang, H., Qiang, M., Lin, P., 2016. A topic modeling based bibliometric exploration of hydropower research. Renewable and Sustainable Energy Reviews, 57, 226–237.

Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E., van Delden, H., Voinov, A.A., 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. Environmental Modelling and Software 47, 159–181.

Kherwa, P., Bansal, P., 2020. Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems, 7(24).

Kohli, A., 2021. Exploring topic modelling in the domain of Integrated Water Resource Management. Master thesis, York University, Canada.

Lee, D. D., Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791.

Li, W., Zhao, Y., 2015. Bibliometric analysis of global environmental assessment research in a 20-year period. Environmental Impact Assessment Review, 50, 158–166.

Mohr, J., Bogdanov, P., 2013. Introduction – topic models: What they are and why they matter. Poetics, 41, 545–569.

Rehurek, R., Sojka, P., 2011. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Rotmans, J., Dowlatabadi, H., 1997. Integrated assessment modelling. In: Rayner, S., Malone, E.L. (eds) Human Choice and Climate Change. Vol. 3, pp. 291-377.

Sammut, C., Webb G.I., 2011. Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_832.

Sievert, C., Shirley, K., 2014. LDAvis: A method for visualizing and interpreting topic. Proc. of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA. doi: 10.3115/v1/W14- 3110.

Toth, F. L. and Hizsnyik, E. (1998). Integrated environmental assessment methods: Evolution and applications. Environmental Modeling & Assessment, 3(3), 193–207.

United Nations, 2015. The 2030 Agenda for Sustainable Development. https://sdgs.un.org/goals (Accessed Aug. 10, 2021).

Zare, F., ElSawah, S., Iwanaga, T., Jakeman, A.J., Pierce, S.A., 2017. Integrated water assessment modelling: A Bibliometric analysis of trends in the water resource sector. Journal of Hydrology, 552, 765-778.