Mining User Forums to Evaluate Quality-in-Use of Environmental Software

Yvette D. Hastings, Jeffrey C. Carver, A. Redempta Manzi Muneza, Robert A. Payn, S. A. Ewing, Stephan Warnat, Ann Marie Reinhold

Mining User Forums to Evaluate Quality-in-Use of Environmental Software

Yvette D. Hastings*, Jeffrey C. Carver†, A. Redempta Manzi Muneza*, Robert A. Payn‡, Stephanie A. Ewing‡, Stephan Warnat§, Ann Marie Reinhold*¶

*Gianforte School of Computing, Montana State University, Bozeman, Montana, USA
†Department of Computer Science, University of Alabama, Tuscaloosa, Alabama, USA
‡Land Resources & Environmental Science, Montana State University, Bozeman, Montana, USA
§Mechanical & Industrial Engineering, Montana State University, Bozeman, Montana, USA
¶Joint appointment with Pacific Northwest National Laboratory, Richland, Washington, USA

***Abstract*—Environmental research software applications are foundational tools for modeling and predicting Earth system dynamics. Among these, reactive transport models (RTMs) are particularly important because they simulate critical system processes that affect contaminant transport and water quality. Many RTMs are developed by domain scientists rather than software engineers, and have quality-in-use (QIU) issues that frustrate end users. Addressing these frustrations requires a clear understanding of the problems and challenges that users experience. We characterize the QIU frustrations most concerning to RTM users using a systematic approach. Specifically, we analyze RTM user forums and codify user-reported challenges by operationalizing the QIU characteristics and subcharacteristics defined in the ISO/IEC 25019:2023 standard. We extracted 3,941 forum threads from four widely used RTM user forums. Terms and phrases related to QIU barriers were identified using natural language processing (NLP) and mapped to QIU subcharacteristics to calculate their frequencies. The QIU characteristic of Beneficialness challenged users most frequently. Within Beneficialness, *Usability* caused the most frustration across the four forums, suggesting users struggle to achieve their goals effectively, efficiently, and with satisfaction. These findings highlight the importance of providing clear, accurate user resources, such as documentation and tutorials, to support research goals. By operationalizing the full scope of the QIU standard, we offer a transferable method for identifying end-user challenges across domains. By treating QIU as a first-class citizen, developers can improve user satisfaction and develop user-centric, trustworthy software systems across software domains.**

***Index Terms*—Software Quality, Quality-in-Use, User Forums, Environmental Software Applications**

I. Introduction

Environmental software applications are foundational tools for monitoring and predicting changes in Earth systems (e.g., climate change, water resource degradation) [1], [2]. Because Earth systems are dynamic and inherently complex, scientists have created software subcategories, generally operationalized as modeling frameworks, to assess specific subsets of Earth system processes. Among these modeling frameworks are reactive transport modeling (RTM) software applications. Researchers use RTMs to simulate the complex set of reactions and transport mechanisms governing dissolved contaminants, which are known to degrade water quality across global Earth systems [1]–[4].

As RTM software applications have evolved to model an increasing number of interacting Earth processes, they have become increasingly complex, both mathematically and computationally [1], [3], [5], [6]. These complexities not only increase the challenges of implementing software with high product quality but also affect software quality-in-use (QIU).

Quality-in-use describes the "extent to which the system or product, when it is used in a specified context of use, satisfies or exceeds stakeholders' needs to achieve specified beneficial goals or outcomes" [7]. In the context of RTM software applications, achieving high QIU would allow users to operate the software, interpret outputs, and report scientific findings effectively within their environmental research workflows. Yet, high QIU is the exception rather than the rule.

Because developers often overlook the user when creating RTM software applications, QIU is diminished [8], [9]. This oversight leaves users struggling to use these software applications. Specific difficulties users commonly express include inexperience in operating software applications from a command line, lack of access to the most current documentation, and non-intuitive or outdated user interfaces [10]. These difficulties are widely acknowledged [9] and compounded by the fact that the developers of many RTM software applications are environmental scientists not versed in software engineering quality-assurance and documentation practices [11], [12]. This situation is not unique to RTM software; rather, it is a common situation across research software domains [13], [14]. As such, these difficulties diminish software quality from the user's perspective.

Researchers have only recently and actively focused on improving both product and QIU characteristics of environmental software applications [9], [11], [13], [15]. This attention is especially important for RTM software, which has a notoriously high learning curve. This learning curve hampers user

This material is based upon work supported by the National Science Foundation under the following Grant Numbers: SitS CBET-2034430 and EPSCoR Cooperative Agreement OIA-1757351. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work is also supported in part by funding from the Montana State University Software Engineering and Cybersecurity Laboratory.

SoutheastCon 2026 | 979-8-3315-4642-7/26/$31.00 ©2026 IEEE | DOI: 10.1109/SOUTHEASTCON63549.2026.11476204

confidence and prevents many environmental scientists from using the software at all. Thus, improvements to QIU can increase the ease of learning and use.

To identify improvements and enhance software QIU from the user's perspective, software engineers rely on standards-based development practices and human-centered design principles [16], [17]. The ISO/IEC 25019:2023 standard provides a formal definition of QIU and a set of measurable characteristics and subcharacteristics for evaluating it, including **Beneficialness**, **Freedom from Risk**, and **Acceptability** [7] (characteristic and subcharacteristic definitions are provided in Table S1). Usage of standards-based development practices increases the likelihood that users will adopt and accept software by incorporating their needs and expectations throughout the software lifecycle [16], [17].

To efficiently evaluate QIU from a user perspective, software engineers can synthesize feedback and questions from online user forums and blogs [18]–[20]. These platforms are an easily accessible avenue for users to ask questions, seek support, and share experiences with software. They also offer engineers insight into recurring challenges users face when learning, using, or trusting software.

While these platforms are easily accessible, they pose a challenge for software engineers. Because users are employing forums at a high frequency, the rate of user posts poses a challenge for software engineers to read, assimilate, and respond to pressing user concerns and questions in a timely manner [21]. Automated methods for quickly extracting information from user forums offer a scalable way to detect common frustrations and identify QIU breakdowns among users. Such analysis may facilitate targeted improvements to software product quality and QIU.

In this study, we evaluate the extent to which we can assess the QIU of RTM software from information scraped from four popular RTM user forums: Hydrus 1D, PHREEQC, Geochemist's Workbench, and the Community Earth System Model (CESM) "Land, River Runoff" forums. More specifically, we pose the following research questions:

RQ: Which QIU characteristics and subcharacteristics, as defined by ISO/IEC 25019:2023, are most and least represented in RTM software user forums?

RQ2: What are the dominant types of user questions that reflect unmet QIU concerns in RTM software forums?

We address these questions by analyzing challenges, concerns, and questions in user threads using the characteristics and subcharacteristics defined in ISO/IEC 25019:2023. We assume that the frequency of threads about a QIU characteristic and subcharacteristic indicates the significance of problems with that characteristic and subcharacteristic. Conversely, we assume that an absence of threads about a QIU characteristic and subcharacteristic indicates that the characteristic and subcharacteristic are satisfying user needs and expectations.

The main contributions of this study are:

- A standards-based evaluation of QIU concerns in RTM software, addressing a critical gap in the QIU evaluation literature. Our analysis identifies the QIU characteristics and subcharacteristics that most frequently lead to user frustration or concern and provides actionable guidance for research software developers on where to prioritize usability and quality improvements.
- A systematic and replicable workflow that enables researchers and practitioners to rapidly extract, classify, and interpret user-reported concerns from software forums to inform QIU-driven design and maintenance decisions. The approach is generalizable and applicable across a wide range of software domains.
- A fully FAIR-compliant research artifact comprised of all code metadata, processing pipelines, data products, and supplemental information archived in Zenodo and openly available to the community[1].

II. Background on Quality Models

The evolution of software quality models has shifted from strict product metrics used to assess usability toward the broader QIU concept [7]. Early quality frameworks, such as the McCall (1977) and Boehm (1978) Models, assess usability (an early measure of QIU) during product operation. These frameworks define usability as the "extent of effort required to learn, operate, and understand" the software [22]–[24]. This definition persisted in the ISO/IEC 9126 (1991) standard and FURPS/FURPS+ (1995), both of which grouped usability with human factors and aesthetics [25]. The concept of usability as a measure of QIU was further broadened by the ISO/IEC 25000 SQuaRE series (2011), which includes effectiveness, efficiency, and satisfaction [26].

In 2023, the ISO/IEC 25019:2023 SQuaRE Quality-in-Use standard [7] was introduced, representing a significant shift in how QIU is conceptualized and evaluated. This shift marks the first time QIU is explicitly separated from product quality, positioning it as a distinct concept focused on the user's context and perspective. Here, the QIU standard provides a continuous evaluation framework for software engineers to define their product goals, measure software's success in meeting user needs, and evaluate and improve their software based on user feedback [7]. Quality-in-use is broken into three main characteristics related to the software application's context of use: Beneficialness, Freedom From Risk, and Acceptability. These QIU characteristics are broken up into more atomic subcharacteristics to facilitate software usability evaluation (Table S1). For this study, we operationalize the ISO/IEC 25019 standard because it is the state-of-the-art quality model that focuses solely on QIU characteristics centered around the user perspective.

[1]https://doi.org/10.5281/zenodo.18132283

III. Methods

Our methods follow a three-stage pipeline: 1) web scraping, 2) natural language processing (NLP) for lexicon development (i.e., keyword extraction & selection, and dictionary construction & refinement), and 3) QIU quantification & statistical analysis. In this section, we provide a high-level overview of the methodology employed in this study. A detailed description of all procedures is available in the supplementary information[1].

A. Web Scraping

We analyzed user threads from four established environmental software forums (Table I). These forums were selected based on their focus on RTM, at least 10 years of archival availability, and high user activity.

Using the `rvest` package [27] in R [28], we scraped 3,941 threads. We targeted original posts to isolate the user's initial motivation, excluding developer-initiated threads and non-English content. Extracted metadata (URLs, timestamps, headers) and body text were stored in a local database.

B. Natural Language Processing

We developed a domain-specific lexicon to quantify QIU characteristics of *Beneficialness* and *Acceptability* and subcharacteristics of *Usability*, *Suitability*, *Accessibility*, *Experience*, *Compliance*, and *Trustworthiness*. We excluded the characteristic of *Freedom from Risk* as cost and environmental impact are rarely discussed in technical support forums.

1) Candidate Keyword Extraction & Selection: We identified candidate keywords (terms and phrases) that operationalize the QIU standard. Using the `quanteda` package [29] in R, we performed unigram and trigram analyses to extract the top 50 terms and top 50 phrases for each forum. Following [30], we classified each keyword as "topical and useful" if it mapped to QIU definitions (Figs. S1 & S2).

We then manually validated relevance by reviewing a random sample of ten threads per keyword (3,160 threads total). This refinement left us with a list of 36 total candidate keywords.

2) Dictionary Construction & Refinement: Dictionary construction required three of the authors to manually code keywords to QIU characteristics and subcharacteristics based on their user experiences and software engineering expertise. These three human coders are software engineers, two of whom are also subject matter experts (SMEs) who have experience with a diverse suite of Earth system modeling software. The human coders finalized mappings through unanimous agreement (i.e., κ = 1), resulting in a finalized dictionary of 35 unique keywords (Tables S2 & S3; Fig. S3). Notably, we allowed for one-to-many mappings, as a single keyword often contextualizes multiple QIU concerns. As a result, a thread could pertain to multiple QIU characteristics and subcharacteristics.

TABLE I
Web scraped scientific software user forums.

RTM Software	Date Scraped	No. Threads	Thread Date Range
Hydrus 1D (H1D)[2]	02-12-24	1731	06-01-04 to 02-06-24
PHREEQC[3]	02-13-24	262	04-15-14 to 02-12-24
Geochemist's Workbench[4]	02-14-24	786	12-09-10 to 10-17-23
CESM "Land, River Runoff"[5]	02-17-24	1162	04-07-06 to 02-16-24

Forum Links:
[2]https://www.pc-progress.com/forum/viewforum.php?f=4
[3]https://phreeqcusers.org/index.php/board,25.0.html?PHPSESSID=gnde9illalp8tfl34905obo814
[4]https://forum.gwb.com/forum/23-the-geochemists-workbench/
[5]https://bb.cgd.ucar.edu/cesm/forums/ctsm-clm-mosart-rtm.134/

3) QIU Quantification: We analyzed the corpus of texts using `quanteda` to identify the presence of QIU characteristics and subcharacteristics. For every thread, we assigned a Boolean value of 1 for the presence of each dictionary keyword and a Boolean value of 0 for the absence of each keyword.

We calculated frequency and proportion at three levels:

1) **Keyword Level:** The proportion of threads containing specific keywords using Eqs. 1 and 2.

$$\text{keywordTotal}_i = \sum_j \text{keywordPresence}_{i,j}, \quad (1)$$

where i indexes each of the 35 keywords, j indexes each forum thread, and $\text{keywordPresence}_{i,j}$ equals 1 if keyword i appears in thread j, and 0 otherwise.

$$\text{keywordProportion}_i = \frac{\text{keywordTotal}_i}{N}, \quad (2)$$

where N is the total number of forum threads.

2) **Subcharacteristic Level:** The proportion of threads containing *any* keyword mapped to a given subcharacteristic using Eqs. 3 and 4.

$$\text{subTotal}_s = \sum_j \text{subPresence}_{s,j}, \quad (3)$$

where s indexes the six subcharacteristics and $\text{subPresence}_{s,j}$ equals 1 if any keyword linked to subcharacteristic s appears in thread j.

$$\text{subProportion}_s = \frac{\text{subTotal}_s}{N}. \quad (4)$$

3) **Characteristic Level:** The proportion of threads associated with parent characteristics using Eqs. 5 and 6.

$$\text{charTotal}_c = \sum_j \text{charPresence}_{c,j}, \quad (5)$$

where c indexes the two characteristics and $\text{charPresence}_{c,j}$ equals 1 if characteristic c is present in thread j.

$$\text{charProportion}_c = \frac{\text{charTotal}_c}{N}. \tag{6}$$

This hierarchical approach ensures that a thread containing multiple keywords within the same category is counted only once to avoid artificially inflating proportions.

C. Statistical Analysis

To account for hierarchical clustering of threads within forums and to quantify between-forum variability, we fit a generalized linear mixed-effects model (GLMM) with a binomial distribution using the `lme4` package [31] in R. The dependent variable was subcharacteristic presence (Boolean), the fixed effect was Subcharacteristic, and a random intercept was specified for Forum. The model converged without warnings, and residual and random-effects diagnostics did not indicate misspecification. The intra-class correlation coefficient (ICC) was computed from the fitted model to estimate the proportion of variance attributable to between-forum differences.

Estimated marginal means (EMMs) for each subcharacteristic were then computed using the `emmeans` package [32] with Tukey multiplicity adjustment. EMMs were reported on the response scale, yielding model-adjusted probabilities that a thread mentions a given subcharacteristic, averaged across forums.

All statistical inferences were conducted at a significance level of $\alpha = 0.05$.

IV. Results

Overall, user concerns were dominated by the QIU characteristic of *Beneficialness*, with the *Usability* subcharacteristic emerging as the primary source of frustration. On average, threads related to *Beneficialness* occurred more often than those related to *Acceptability* (73% [range: 69-76%] versus 9% [range: 6-12%], respectively; Fig. 1). While the mean values highlight the disproportionate emphasis between characteristics, the ranges capture the variability in user frustrations across forums.

Our NLP classification revealed that keyword usage associated with *Beneficialness* was primarily driven by the *Usability* subcharacteristic, followed by *Suitability* and *Accessibility* (67% [range: 66-68%], 45% [range: 40-51%], and 13% [range: 4-27%] of threads, respectively; Fig. 1). In contrast, *Acceptability*-related keyword usage was dominated by *Experience*, followed by *Compliance* and *Trustworthiness* (6% [range: 4-9%], 3% [range: 2-4%], and 1% [range: 0.6-1%] of threads, respectively). Similar to the characteristic-level results, the ranges of subcharacteristic proportions indicate that some concerns were consistently prevalent across threads, whereas others were comparatively rare and unevenly distributed across forums. This suggests that certain subcharacteristics varied substantially by forum. For example, *Accessibility* exhibited

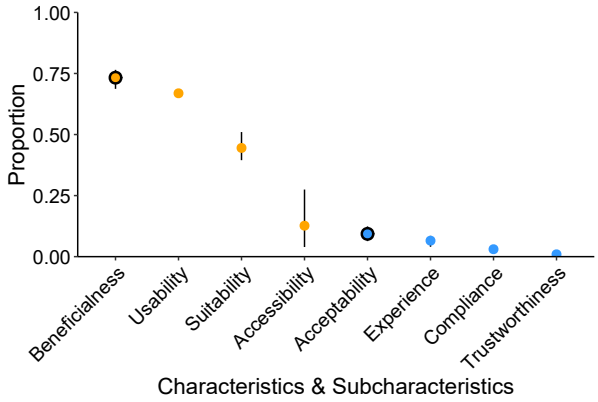

Fig. 1. Proportion of QIU characteristic and subcharacteristic thread mentions across user forums. Black-outlined circles denote QIU characteristics (Beneficialness in orange; Acceptability in blue). Filled circles denote subcharacteristics, colored by their associated characteristic (orange for Beneficialness; blue for Acceptability). Points indicate mean proportions across forums; vertical lines indicate the range of proportions.

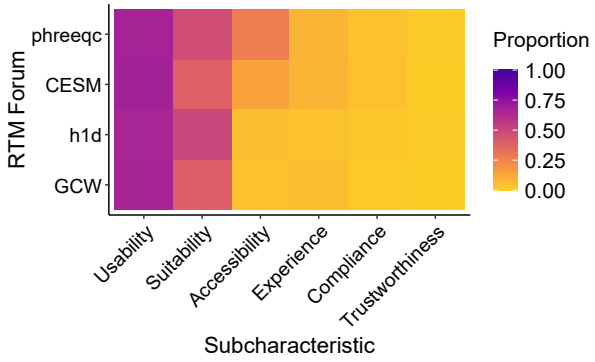

Fig. 2. Proportions of QIU subcharacteristic mentions across the four user forums. Dark (purple) colors indicate higher proportions, and light (yellow) colors indicate lower proportions.

the widest range of mentions (Figs. 1 & 2), indicating forum-specific differences in user concern.

To evaluate whether these patterns persisted after accounting for forum-level clustering, we fit a mixed-effects hierarchical model. Model-adjusted estimates confirmed that *Usability* and *Suitability* were the dominant sources of frustration across all forums (Table II), while forum-level clustering explained only a small proportion of total variance (ICC $\approx$ 1.1%). Thus, although absolute levels differed among forums, the primary drivers of user concern were largely consistent (Fig. 2).

At a finer level of granularity, keyword usage displayed a similar pattern to the subcharacteristic-level results. Subcharacteristic results nested under *Beneficialness* were driven primarily by the repeated use of keywords such as *general use*, *general suitability*, *error*, and *parameters*, which, on average, appeared in 47%, 27%, 18%, and 14% of threads, respectively (Fig. 3). All remaining keywords identified during classification exhibited mean usage rates of $\leq$ 13% across forums. These patterns indicate that users are primarily fo-

TABLE II
ESTIMATED PROBABILITIES OF QIU SUBCHARACTERISTIC MENTIONS. ALL EFFECTS WERE SIGNIFICANT AT $p < 0.001$; VALUES ARE REPORTED AS PERCENTAGES WITH 95% CONFIDENCE INTERVALS; VALUES WERE ADJUSTED FOR FORUM-LEVEL CLUSTERING.

Subcharacteristic	Estimated Probability ± SE	95% CI
Usability	68 ± 2.3	63–72
Suitability	46 ± 2.6	41–51
Accessibility	9.3 ± 1.0	7.6–11
Experience	6.1 ± 0.7	4.9–7.6
Compliance	3.0 ± 0.4	2.3–3.9
Trustworthiness	1.0 ± 0.2	0.7–1.5

cused on learning how to operate RTM software, evaluating whether the software meets their modeling needs, correcting execution errors, and adjusting model parameters. Consistent with subcharacteristic-level results, keyword usage patterns were relatively uniform across forums.

The results above directly address our research questions by identifying which QIU characteristics and subcharacteristics are most frequently discussed in RTM user forums. These findings highlight the dominant sources of user frustration and concern within the RTM software community. We summarize the key results below.

RQ1 answer:

- Users most frequently expressed concerns related to the QIU *Beneficialness* characteristic, indicating that RTM users are often not achieving the intended benefits when using these software applications.
- The *Usability* subcharacteristic accounted for the majority of these concerns, suggesting persistent challenges related to efficiency, effectiveness, and satisfaction.

RQ2 answer:

- The most common user questions involved general software use, general suitability, runtime errors, and parameter configuration.

V. DISCUSSION, IMPLICATIONS, & FUTURE WORK

A. Implications for RTM software applications

User forum activity indicates that RTM software users experience the greatest frustration with the QIU characteristic of *Beneficialness*, driven primarily by the *Usability* subcharacteristic. Our manually coded NLP analysis revealed that a substantial portion of user threads centered on software operation, including general use, input errors, and parameter configuration. Many users struggled to locate clear documentation on how to adjust or interpret key parameters required for simulations, and others expressed confusion regarding software functionality and error resolution. These challenges hinder users' ability to conduct environmental simulations effectively and efficiently, thereby reducing their satisfaction.

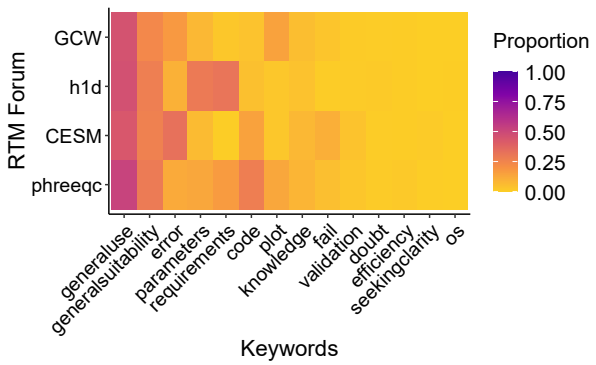

Fig. 3. Proportions of QIU keyword mentions across the four user forums. Dark (purple) colors indicate higher proportions, and light (yellow) colors indicate lower proportions.

Rather than focusing primarily on software defects or faulty outputs, forum discussions suggest that users are often preoccupied with understanding *how* to use RTM software to accomplish their modeling objectives. Previous studies similarly report that technology-related frustrations significantly affect user productivity and satisfaction [33]–[35]. Hertzum [36] found that users waste approximately 11–20% of their time due to technology-related issues, underscoring the substantial cost of poor usability. These findings reinforce the need to prioritize usability in the design and evolution of RTM software applications.

Broader structural issues in environmental software development further compound usability challenges. Kelly et al. [37] observed that domain scientists who develop environmental software rarely perform usability testing beyond their own laboratories, limiting broader user input. This disconnect between end users and developers contrasts with modern software engineering practices, such as Agile and DevOps, which emphasize continuous user involvement throughout the development lifecycle. Indeed, Pagano and Brügge [38] highlight that understanding users' needs and expectations is essential for effective software planning and evolution. These approaches remain underutilized in scientific software development.

In addition to *Usability*, users frequently expressed frustration with *Suitability*, particularly when questioning whether RTM software possessed the capabilities required for their intended simulations. This pattern suggests that practical guidance on appropriate software applications and limitations is often unclear or insufficiently documented. Because RTM software underpins environmental decision-making, QIU failures related to usability and suitability may have downstream implications for model reproducibility and interpretation. Difficulties in configuring parameters or understanding model behavior may contribute not only to inefficiencies but also to the risk of undetected modeling errors. To mitigate both usability and suitability challenges, RTM developers would benefit from systematically incorporating user feedback throughout

the software lifecycle, beginning with early design phases.

Finally, improving user-facing resources remains critical for reducing RTM software frustration. Although documentation, tutorials, and workshops are available for many RTM platforms, multiple forum threads raise questions about their clarity, accessibility, and comprehensiveness. Grossman et al. [39] demonstrated that software documentation indirectly influences user satisfaction through perceived product quality, emphasizing the importance of well-designed learning resources. Together, these findings suggest that enhancing the quality, structure, and accessibility of RTM documentation and training materials could substantially improve users' ability to model complex environmental processes efficiently and effectively [16], [40], [41].

B. Relationship to Prior Work and Implications for Researchers

Systematic issues with software QIU exist across many software application domains. For instance, a study by [20] analyzed reviews of web browsers and found that users were most concerned with the *Freedom from Risk* characteristic followed by the *Satisfaction* characteristic of the ISO/IEC 25010:2011 [26] standard. Although *Satisfaction* is not a named characteristic in the ISO/IEC 25019 standard, it is represented by the *Beneficialness* characteristic and *Usability* subcharacteristic of this new standard. Using the language of the ISO/IEC 25019 standard, [20] findings indicate that *Freedom from Risk* characteristic and the *Usability* subcharacteristic within *Beneficialness* are the predominant concerns for users of web browsers.

Our results are similar to those of [20] in that the most significant concerns for the users of RTM software are also related to *Beneficialness* and *Usability*. Such similarities are not surprising, given that both web browser and RTM domains require usable and easy-to-learn software. This alignment highlights a broader lesson for software engineers: usability and overall user benefit are recurring concerns, regardless of the application domain. Software that lacks these qualities is more likely to be abandoned, even when alternative tools require more technical effort to use. Addressing QIU concerns proactively improve the software lifecycle through increased adoption and support long-term user engagement.

While previous studies have mined user reviews to identify quality issues, few have directly mapped these issues to formal QIU standards. Several studies explore mining software forums or vendor platforms to derive software requirements. For instance, [19] and [42] demonstrate how user generated content can guide software development. However, these studies typically focus on product functionality and ignore QIU frameworks that emphasize user acceptance, risk mitigation, and benefit. As a result, QIU issues (i.e., usability, trustworthiness, experience, etc.) go unaddressed.

Web scraping and text mining of user forums provide a robust way to collect user feedback [19]–[21], [42]–[44]. However, these methods come with limitations. As a result, we recommend that those who apply these methods consider the potential biases they can introduce. First, the timing of web scraping relative to the frequency of users posting to open online forums is a major consideration [21]. If developers do not regularly scrape the forums, they may miss user concerns. Second, developers can miss user comments if users post to third-party forums (i.e., Stack Overflow[6]). Third, interpretation bias can lead to misclassifications of manually coded NLP dictionaries. These misclassifications lead to incorrect conclusions.

To overcome these challenges, involving domain experts when designing classification models and interpreting user feedback is critical. Equally important is the regular scraping of multiple online platforms to capture user concerns in a timely and comprehensive manner. To accelerate this process, many practitioners have started employing off-the-shelf large language models (LLMs) to quickly gather and analyze user feedback and questions. While these LLMs offer speed and broad language capabilities, they often lack the contextual understanding required to detect QIU issues in domain-specific applications. Without expert input, these models are prone to misclassifying feedback or overlooking critical concerns that would be obvious to experienced practitioners. By incorporating domain expertise into the design and tuning of NLP tools, software engineers can significantly enhance the accuracy, relevance, and practical value of QIU assessments.

In summary, QIU is not just a post-release concern. By integrating QIU assessments throughout the software development process, and treating it as a first-class citizen, software engineers can develop products that meet user expectations and needs, thus delivering long-term value to end-users across software domains.

C. Future Research

Traditional social science methods can complement web scraping and text mining to identify software application QIU concerns. [38], [17], and [44] distributed surveys or conducted interviews to determine the user experience and relate it to the QIU of software applications. The authors of all studies agree that directly consulting users through surveys or interviews is an effective method for determining software QIU from the user perspective [17], [38], [44].

Our ongoing research includes surveying several members of the RTM and the broader environmental software community to directly gather their experiences with software QIU. The surveys will allow us to understand whether user threads reflect a small subset of frustrated users or if they broadly represent the RTM and environmental user community as a whole.

VI. Threats to Validity

Construct validity addresses the extent to which theoretical QIU concepts are accurately operationalized through the selected keywords. Keyword selection is foundational to this study and directly influences all downstream analyses. A

[6]https://stackoverflow.com/

limited dictionary risks omitting terms that represent important QIU concerns, while an overly expansive dictionary risks introducing noise by including tangential or weakly related terms. To mitigate this threat, we combined objective term-frequency and n-gram analyses with expert-driven manual review. Three software engineers, including two subject matter experts, independently evaluated and mapped candidate terms to QIU subcharacteristics, with unanimous agreement on final assignments. Nevertheless, some degree of subjectivity in keyword selection is unavoidable in NLP-based studies, and different research teams may reasonably select different sets of terms depending on domain familiarity and user context.

Internal validity addresses whether observed patterns are attributed to the causality rather than selection bias. A primary threat arises from the limited availability of RTM software forums, which constrained the pool of potential systems for inclusion. Many RTM tools lack public user forums entirely. To mitigate this limitation, we selected widely used RTM software based on citation impact and expert consensus. Two systems were selected from prior bibliometric work identifying the most cited RTM applications, while two additional systems were selected based on sustained professional use by five domain experts. While this strategy maximizes coverage of active user communities, it does not eliminate the possibility that omitted tools exhibit different QIU concern profiles.

Conclusion validity relates to the appropriateness of inferences drawn from the observed data. This study assumes that the presence of a keyword reflects a user concern related to the corresponding QIU subcharacteristic, while absence indicates lower salience rather than actual software adequacy. This assumption represents the primary threat to conclusion validity. Users may refrain from discussing a QIU concern not because it is fully satisfied, but because their immediate support need lies elsewhere. However, repeated omissions across thousands of threads and multiple forums provide reasonable evidence that certain QIU aspects are less dominant in the user forums at the time of posting.

External validity concerns the generalizability of findings beyond the studied context. This work focuses exclusively on RTM software, which involves high technical complexity and steep learning curves. These characteristics differ substantially from commercial, financial, or other domains with different user expectations and risk constraints. Consequently, the specific QIU concern patterns identified here may not transfer directly to other software domains. However, because our methods are grounded in ISO/IEC 25019 and use a systematic standards-based classification pipeline, the framework itself remains transferable with domain-specific lexical adaptation. Additionally, the data likely overrepresent vocal or less experienced users while underrepresenting satisfied or experienced users.

VII. Conclusion

Mining user forums can help assess software quality, especially for QIU. While our focus was on mining user forums to evaluate the QIU of environmental software applications (and RTMs, in particular), our systematic qualitative approach applies to other types of software where user forums are a common means of communication. The results of applying this approach may vary depending on the software domain, user base, and developer community; however, the process is transferable. Because user forums are created for and easily accessed by users, these platforms provide a rich source of information for software engineers to determine frequently encountered obstacles to QIU. Thus, mining user forums presents an opportunity to extract information on QIU characteristics and subcharacteristics and may even help developers improve the QIU of environmental software applications.

Finally, our study demonstrates the importance of considering the end-user perspective on RTM QIU. We emphasize that QIU should be treated as a first-class citizen, where software applications developed with the end-user in mind will enhance the software lifecycle and improve the overall user experience. For software engineers, our results highlight the value of systematically incorporating user feedback throughout the development and maintenance phases, rather than treating usability, experience, or other QIU concerns at the final stages of release. By prioritizing user-centered measures of QIU, engineers can proactively address potential adoption barriers, reduce user frustration, and enhance the long-term sustainability of software applications across domains. Ultimately, applying QIU frameworks helps bridge the gap between functional correctness and real-world usefulness, thereby creating high-value, beneficial software products for end-users across diverse domains.

Acknowledgment

We thank several individuals in the Montana State University Software Engineering and Cybersecurity Laboratory for their feedback and review of this work: Ryan Cummings, Garrett Perkins, Madison Munro, and Eric O'Donoghue.

References

[1] C. I. Steefel, S. B. Yabusaki, and K. U. Mayer, "Reactive transport benchmarks for subsurface environmental simulation," *Computational Geosciences*, vol. 19, pp. 439–443, 2015.

[2] L. Li, K. Maher, A. Navarre-Sitchler, J. Druhan, C. Meile, C. Lawrence, J. Moore, J. Perdrial, P. Sullivan, A. Thompson *et al.*, "Expanding the role of reactive transport models in critical zone processes," *Earth-science reviews*, vol. 165, pp. 280–301, 2017.

[3] C. I. Steefel, D. J. DePaolo, and P. C. Lichtner, "Reactive transport modeling: An essential tool and a new research approach for the earth sciences," *Earth and Planetary Science Letters*, vol. 240, pp. 539–558, 2005.

[4] H. Vereecken, A. Schnepf, J. W. Hopmans, M. Javaux, D. Or, T. Roose, J. Vanderborght, M. Young, W. Amelung, M. Aitkenhead *et al.*, "Modeling soil processes: Review, key challenges, and new perspectives," *Vadose zone journal*, vol. 15, no. 5, pp. vzj2015–09, 2016.

[5] J. Carrera, M. W. Saaltink, J. Soler-Sagarra, J. Wang, and C. Valhondo, "Reactive transport: a review of basic concepts with emphasis on biochemical processes," *Energies*, vol. 15, no. 3, p. 925, 2022.

[6] D. Wang, T. Janjusic, C. M. Iversen, P. E. Thornton, M. Karssovski, W. Wu, and Y. Xu, "A scientific function test framework for modular environmental model development: Application to the community land model," in *2015 IEEE/ACM 1st International Workshop on Software Engineering for High Performance Computing in Science (SE4HPCS)*. Florence, Italy: Institute of Electrical and Electronics Engineers Inc., 2015, pp. 16–23.

[7] "Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE)-Quality-in-use model," International Organization for Standardization, Geneva, CH, Standard, Nov. 2023.

[8] Y. D. Hastings and A. M. Reinhold, "Applying software quality in use standards to improve scientific software selection," in *49th Euromicro conference series on Software Engineering and Advanced Applications (SEAA 2023)*, ser. Works in Progress in Embedded Computing Journal (WiPiEC), vol. 9, Durres, Albania, 2023, pp. 22–25.

[9] M. A. Heroux, D. E. Bernholdt, L. C. McInnes, J. R. Cary, D. S. Katz, E. M. Raybourn, and D. Rouson, "Basic research needs in the science of scientific software development and use: Investment in software is investment in science," US Department of Energy (USDOE), Washington, DC (United States). Office of Science, Tech. Rep., 2023.

[10] F. J. R. Meysman, J. J. Middelburg, P. M. J. Herman, and C. H. R. Heip, "Reactive transport in surface sediments. I. Model complexity and software quality," *Computers and Geosciences*, vol. 29, pp. 291–300, 2003.

[11] J. Koehler Leman, B. D. Weitzner, P. D. Renfrew, S. M. Lewis, R. Moretti, A. M. Watkins, V. K. Mulligan, S. Lyskov, J. Adolf-Bryfogle, J. W. Labonte *et al.*, "Better together: Elements of successful scientific software development in a distributed collaborative community," *PLoS computational biology*, vol. 16, no. 5, p. e1007507, 2020.

[12] L. Rampersad, S. Blyth, E. Elson, and M. M. Kuttel, "Improving the usability of scientific software with participatory design: a new interface design for radio astronomy visualisation software," in *Proceedings of the South African Institute of Computer Scientists and Information Technologists*. Thaba 'Nchu, South Africa: Association for Computing Machinery, 2017.

[13] E.-M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, and J. C. Carver, "Software engineering practices for scientific software development: A systematic mapping study," *Journal of Systems and Software*, vol. 172, p. 110848, 2021.

[14] J. C. Carver, N. Weber, K. Ram, S. Gesing, and D. S. Katz, "A survey of the state of the practice for research software in the united states," *PeerJ Computer Science*, vol. 8, p. e963, 2022.

[15] R. Milewicz and P. Rodeghero, "Position paper: Towards usability as a first-class quality of hpc scientific software," in *2019 IEEE/ACM 14th International Workshop on Software Engineering for Science (SE4Science)*, Montral, Quebec, Canada, 2019, pp. 41–42.

[16] J. Singh and N. B. Kassie, "User's perspective of software quality," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 1958–1963.

[17] L. Souza-Pereira, N. Pombo, and S. Ouhbi, "Software quality: Application of a process model for quality-in-use assessment," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, pp. 4626–4634, 2022.

[18] J. Tizard, "Requirement mining in software product forums," in *Proceedings of the 27th IEEE International Conference on Requirements Engineering*, vol. 2019-September. Jeju Island, South Korea: IEEE Computer Society, 2019, pp. 428–433.

[19] J. A. Khan, "Mining requirements arguments from user forums," in *Proceedings of the IEEE International Conference on Requirements Engineering*, vol. 2019-September. Jeju Island, South Korea: IEEE Computer Society, 2019, pp. 440–445.

[20] I. Atoum, "A novel framework for measuring software quality-in-use based on semantic similarity and sentiment analysis of software reviews," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, pp. 113–125, 2020.

[21] Y. Jiang and T. Hu, "Software's quality-in-use mining for user's comments," in *2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI)*, Kunming, China, 2019, pp. 37–46.

[22] B. Singh and S. P. Kannojia, "A review on software quality models," in *2013 International Conference on Communication Systems and Network Technologies*, Gwalior, India, 2013, pp. 801–806.

[23] GeeksforGeeks, "McCall's Quality Model," 2024, Accessed 11 Sept 2024. [Online]. Available: https://www.geeksforgeeks.org/mccalls-quality-model/

[24] ——, "Boehm's Software Quality Model," 2024, Accessed 11 Sept 2024. [Online]. Available: https://www.geeksforgeeks.org/boehms-software-quality-model/

[25] Quality Management, "FURPS," 2018, Accessed 11 Sept 2024. [Online]. Available: https://qualitymanagement387741893.wordpress.com/2018/01/10/first-blog-post/

[26] "Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE)-System and software quality models," International Organization for Standardization, Geneva, CH, Standard, 2010.

[27] H. Wickham, *rvest: Easily Harvest (Scrape) Web Pages*, 2024, R package version 1.0.4, https://github.com/tidyverse/rvest. [Online]. Available: https://rvest.tidyverse.org/

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: https://www.R-project.org/

[29] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo, "quanteda: An r package for the quantitative analysis of textual data," *Journal of Open Source Software*, vol. 3, no. 30, p. 774, 2018.

[30] C. Izurieta, N. Woods, and A. M. Reinhold, "A brief of distributed data processing," in *49th Euromicro conference series on Software Engineering and Advanced Applications (SEAA 2023)*, ser. Works in Progress in Embedded Computing Journal (WiPiEC), vol. 9, Durres, Albania, 2023, pp. 14–17.

[31] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[32] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024, r package version 1.10.2. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[33] K. Bessière, J. E. Newhagen, J. P. Robinson, and B. Shneiderman, "A model for computer frustration: the role of instrumental and dispositional factors on incident, session, and post-session frustration and mood," *Computers in Human Behavior*, vol. 22, no. 6, pp. 941–961, 2006.

[34] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman, "Determining causes and severity of end-user frustration," *International Journal of Human–Computer Interaction*, vol. 17, no. 3, pp. 333–356, 2004.

[35] J. Lazar, A. Jones, and B. Shneiderman, "Workplace user frustration with computers: an exploratory investigation of the causes and severity," *Behaviour & Information Technology*, vol. 25, no. 3, pp. 239–251, 2006.

[36] M. Hertzum and K. Hornbæk, "Frustration: Still a common user experience," *ACM Trans. Comput.-Hum. Interact.*, vol. 30, no. 3, 2023.

[37] D. Kelly and R. Sanders, "The challenge of testing scientific software," in *Proceedings of the 3rd annual conference of the Association for Software Testing (CAST 2008: Beyond the Boundaries)*. Citeseer, 2008, pp. 30–36.

[38] D. Pagano and B. Bruegge, "User involvement in software evolution practice: A case study," in *35th International Conference on Software Engineering (ICSE)*, San Francisco, California, USA, 2013, pp. 953–962.

[39] T. Grossman, G. Fitzmaurice, and R. Attar, "A survey of software learnability: metrics, methodologies and guidelines," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 649–658.

[40] E. Aghajani, C. Nagy, O. L. Vega-Márquez, M. Linares-Vásquez, L. Moreno, G. Bavota, and M. Lanza, "Software documentation issues unveiled," in *IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, Montreal, Quebec, Canada, 2019, pp. 1199–1210.

[41] E. H. Weiss, "The retreat from usability: user documentation in the post-usability era," *SIGDOC Asterisk J. Comput. Doc.*, vol. 19, no. 1, p. 3–18, 1995.

[42] M. Stade, M. Oriol, O. Cabrera, F. Fotrousi, R. Schaniel, N. Seyff, and O. Schmidt, "Providing a user forum is not enough: First experiences of a software company with crowdre," in *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, 2017, pp. 164–169.

[43] B. Lin, N. Cassee, A. Serebrenik, G. Bavota, N. Novielli, and M. Lanza, "Opinion mining for software development: A systematic literature review," *ACM Trans. Softw. Eng. Methodol.*, vol. 31, no. 3, pp. 1–41, 2022.

[44] J. Bragge, H. Merisalo-Rantanen, and P. Hallikainen, "Gathering innovative end-user feedback for continuous development of information systems: a repeatable and transferable e-collaboration process," *IEEE Transactions on Professional Communication*, vol. 48, no. 1, pp. 55–67, 2005.