オープンソフトウェアRによる再現可能な数理モデル解析
ー水稲用除草剤の水中残留データ解析を例にー

一般財団法人残留農薬研究所
化学部
近藤　圭

はじめに

自然科学におけるモデルは複雑な現象を仮定に基づき簡略化することでその解釈を支援するツールである。現象のモデル化は，そのプロセスやメカニズムを記述するのに十分な要素のみを抽出し，ダイヤグラムやフローチャートで表現する概念モデルを構築することから始まる。この過程は，仮説を立て，それに基づきデータを観測する実験計画にも共通していることから，モデルはあらゆる科学的手法において必須である。概念モデルは，続いて各要素を変数として扱い，対象とするプロセスやメカニズムを微分方程式により定式化することで数理モデルとなる。ここで，得られたデータからモデルを構築する統計モデルや機械学習モデルも広義の数理モデルであるが，本稿では上述の定義のものに限定して数理モデルと呼ぶこととする。数理モデルは，与えられたルール（パラメータや境界条件）に対する挙動（出力）を調べることで，現象に由来するデータ生成メカニズムを理解することに有用である。一般に，数理モデルは観測データに基づきそのパラメータをキャリブレーションすることで，データ間の補間あるいはデータ範囲外への外挿が可能となる。さらに，実験では観測できない要素の定量化や新たな実験を行うことなくその結果を予測することもできる（Soetaert and Herman 2009)。

上述の利点から，これまでに農薬の環境中動態を評価する様々な数理モデルが開発・利用されており（Ippolito and Fait 2019)，我が国においても水田環境中における農薬動態を予測する数理モデルの研究が盛んに行われてきた(Inao and Kitamura 1999; Watanabe *et al*. 2006)。これらの多くは，MS Excel® を用いたスプレッドシート上で実装されており（以下，スプレッドシートモデル)，数理モデルに不慣れなユーザーにも配慮した作りとなっている。

ところで近年，研究のオープンサイエンス化が全世界で急速に進んでいる。例えばFAIR原則（Wilkinson *et al*. 2016）は，データのオープンネス(openness) を進めるうえで，優れたデータ管理とは何かを具体的に記述したもので，この考え方に則ったデータ整備や提供のサービス化が進められている（青木 2021)。一方，統計解析や数理モデリングといったデータ解析では，これに加えてさらに透明性（transparency）や再現性（reproducibility）が求められている（Choi *et al*. 2021)。前者はステークホルダーのモデルの構造，支配方程式，パラメータおよびモデルの仮定を認識しやすいことであり（Eddy *et al*. 2012)，後者は同じ入力データ，計算手順，方法，コード，解析条件を用いて一貫した結果を得ることである(National Academies of Sciences and Medicine 2019)。

これらの点で，上述のスプレッドシートモデルは，生データの前処理，モデル解析条件また出力結果の作図等の操作における逐次記録性や追跡性が弱く，それらがユーザー間での再現性に大きく影響する。そのため，最近ではスクリプトベースでの解析環境が好まれる傾向にあり，オープンソフトウェアRおよびその統合開発環境（IDE）であるRStudioを用いたデータ解析がスプレッドシートモデルの代替手段となりつつある (Alarid-Escudero *et al*. 2019)。くわえて，近年の計算機科学の発展，特にモンテカルロシミュレーション等の確率論的アプローチが広く利用されるようになり，それらの実装の困難さや膨大な計算量とデータの取り扱いの煩雑さからも，スプレッドシートモデルから多数の計算パッケージやアプリケーション例が多いRへの転換が求められている（Baio and Heath 2017)。

本稿ではオープンサイエンスを進めるうえでの3課題（オープンネス，透明性および再現性）の内，再現性について，Rで構築した数理モデルによる水稲用除草剤の水中残留データ解析の模擬実演を通じた話題共有を主題とする。また実演内容は，読者の方が実際にソースコードをダウンロードして演習的に体験することも可能となっていることを申し添えておく。

1. 問題設定と解析手順

本稿で取り扱う問題として，任意の除草剤Xを有効成分として8% 含有する乳剤Aの水質汚濁性試験データ(試験区名：LC）をもとに，標準的な流域における河川水中濃度を予測し，

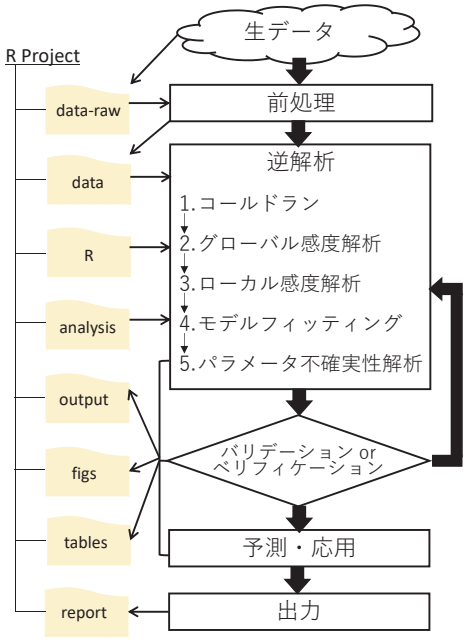

図 -1　解析フローチャートおよび R プロジェクトでのフォルダ構成（黄色枠は各フォルダを表す）

その影響要因を調査することを取り上げる。データ解析には水田環境中での農薬動態を予測する数理モデル PCPF (Pesticide Concentration in Paddy Field)-1R モデルを用いた (Kondo *et al*. 2019; Kondo *et al*. 2020; Kondo 2022)。オリジナルモデルは，東京農工大学で開発されており，MS Excel® を用いたスプレッドシートモデルとして設計されている。PCPF-1R モデルはオリジナルモデルにおける基礎理論はそのままに，オープンソフトウェア R で再構築したモデルである。このモデルを R の逆解析パッケージである “FME”（Soetaert and Petzoldt 2010）を組み合わせることにより，実験データをもとにしたオートキャリブレーションが実装可能となる。

解析手順を図 -1 に示す。解析は，R の IDE である RStudio を用いた。本稿で紹介する RStudio のインストール方法や機能の詳細については高橋（2018）の書籍を参照されたい。RStudio では R Project 機能が備わっており，データ解析をプロジェクトベースで管理することができる。これにより，PC 上での作業ディレクトリ指定が絶対パスから相対パスに簡略化されるため，解析データの出入力が容易になる。その結果，図 -1 に示す通り用途毎にフォルダ分けをすることができ，解析データの散逸を防止できる。

解析ではまず，必要とするデータを収集し，それらを生データとしてフォルダ（“data-raw”）に保存する。次に，これらの生データを R 上で取り扱いやすく前処理してから解析に供する。解析では，前処理済みデータ（“data”），解析に必要なユーザー定義関数（“R”）およびそれらを処理するコード（“analysis”）を各フォルダから取り出して逆解析を実行する。得られた解析結果やそれらをもとに作成した図表についてもそれぞれ対象フォルダ（“output”, “figs” および ”tables”）に保存する。必要に応じて後述する R マークダウン機能を用いて解析レポート（“report”）を作成することもできる。これらの一連の作業は細かな違いはあるが，統計解析および数理モデリングのいずれの場合にも共通の推奨操作となっている（Alarid-Escudero *et al*. 2019; BES and Cooper 2017)。

2. ハンズオン

(1) 解析準備〜データの前処理

PCPF-1R モデルを用いた具体的な解析について説明する。モデルのソースコードは，分散型バージョン管理システム Git を用いて維持・管理され，そのオンラインプラットフォームである GitHub が提供するリモート環境 (リポジトリ) で共有可能な状態にしている（図 -2)。はじめに GitHub 上のリポジトリ（https://github.com/k-kondo-IET/PCPF-1R）から PCPF-1R のプロジェクトソースをダウンロードし，自身の PC にローカルリポジトリを作成する。次に解析のために必要な生データを入力する。実験データから対象農薬の水中残留濃度（“data_conc_label_1.csv”），水収支（“data_wb_1.csv”), 実験圃場の情報（“input_experiment.csv”）および処理農薬製剤情報（“input_pesticide_label_1.csv”）を入力する。対象農薬中有効成分の物理的化学的パラメータ (“input_pesticide_physchem.csv”) は，分子量，水溶解度，蒸気圧，土壌吸着性および分解情報が必要になる。これらの情報に関するデータソースは様々あるが，本稿では原則として，全てが一括で入手可能である農薬抄録あるいは農薬の審査報告書からの取得を推奨する。生データの入力が完了したら，”analysis” フォルダから解析スクリプト（“Hands-on.R”）を開き，解析に使用するモデルや収集した生データを読み込み，表 -1 に示す PCPF-1R モデルの入力パラメータセットを前処理済みデータとして作成する。

(2) 逆解析〜バリデーション

続いて逆解析の工程に移る。"data" フォルダから前処理済みデータを呼び

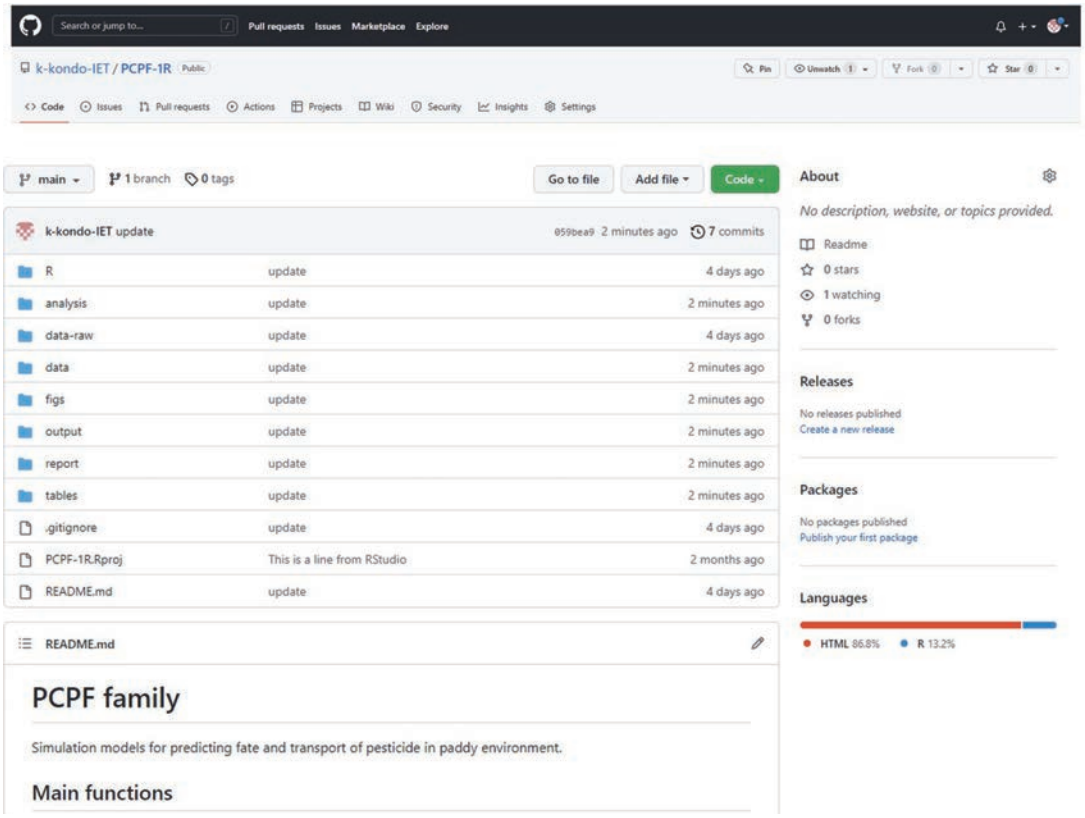

図 -2　PCPF-1R モデルのソースコードが共有可能な GitHub 上のリモートリポジトリ（https://github.com/k-kondo-IET/PCPF-1R にアクセスし，画面に表示される緑色の “Code” ボタンをクリックして “Download ZIP” を選択してダウンロードする。Rstudio から "PCPF-1R.Rproj" を開くと図 -1 に示した R プロジェクトが表示される。）

出し，表 -1 の入力パラメータをキャリブレーションせずにシミュレーションを行う（コールドラン）。これにより，実験データに対するモデルの挙動を視覚的・統計的に確認し，モデルの仮定が実験データに対して妥当であるか，表 -1 の入力パラメータを実験データに合わせて調整するパラメータキャリブレーションを必要性があるか確認する。パラメータキャリブレーションを行う場合には，除草剤 X の水中残留濃度に関する実測データとモデル出力との誤差の平方和を目的関数として定義する。ここで各データに対して 1/（実測データ）の重みを付与したが，これは重みなしのデータで解析を進めた結果，後述するバリデーションまたはベリフィケーション工程で得られた結果が不適と判断されたためである。FOCUS（2006）のガイダンスでも推奨されている通り，はじめは重みなしでの解析を進めることが適当である。

続いて実施するグローバル感度解析では，比較的広いパラメータ範囲を

表 -1　PCPF-1R モデルのパラメータとその算出根拠

パラメータ	Rでの名前	コールドランでの値	算出に用いた生データ
圃場面積（m^2）	Area	1	実験情報
土壌の仮比重（g/cm^3）	bulk	1.04	土壌情報
飽和体積含水率（cm^3/cm^3）	SatWC	0.62	土壌情報
農薬散布量（g/m^2）	AppR	0.04	実験情報
攪拌速度定数（1/day）	alp	1	—
水溶解度（mg/L）	CSLB	74	水溶解度
土水平衡分配係数（L/kg）	Kd	30.56	実験情報，Freundlichパラメータ
見かけ吸着フラクション（—）	f	1	—
拡散速度定数（m/day）	kdiff	0.0026	分子量，土壌情報
吸脱着速度定数（1/day）	ksorp	0.028	土壌情報
揮発速度定数（m/day）	kvol	0.00019	分子量，水溶解度，蒸気圧
水中分解速度定数（1/day）	kbulk	0.12	半減期（加水分解，水中光分解，土壌動態）
土壌中微生物分解速度定数（1/day）	kbios	0.036	半減期（土壌動態）

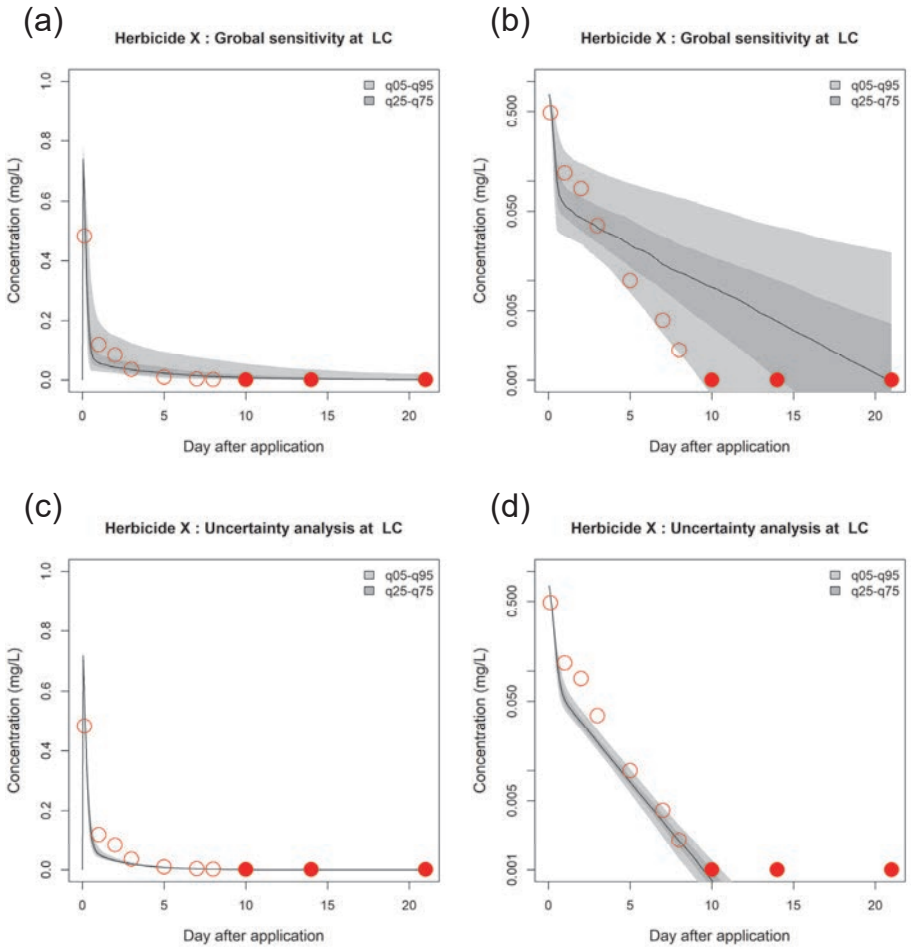

図 -3　PCPF-1R モデルによる除草剤 X 解析結果の線形および対数スケールプロット（(a) および (b) グローバル解析結果，(c) および (d) パラメータ不確実性解析の結果；q05-q95 は 5-95 パーセンタイル，q25-q75 は 25-75 パーセンタイル，実線は 50 パーセンタイル，赤丸は実測値であり，塗りつぶしは定量限界未満であることを表す）

ランダムサンプリングして繰り返し計算を行うモンテカルロ法により探索し，目的関数に対するモデルパラメータの影響度合いを定量化する。ここではモデルパラメータの内，変動が予想される分解，吸脱着，拡散に関わる速度定数や土壌吸着係数等，計 8 パラメータを選択し，1/X×M ～ X×M（M はコールドランで使用したパラメータ値，X は任意の倍数）の範囲でラテンハイパーキューブ法により 250 組のパラメータセットを生成してモンテカルロシミュレーションを行う。この結果から，R パッケージ ”sensitivity”（Iooss *et al.*）を用いて，重回帰モデルを作成し，その標準化ランク偏回帰係数（SRRC）を算出することで目的関数に対するパラメータ感度を求める（Boulange *et al.* 2012; Kondo *et al.* 2012）。ここでは SRRC の絶対値が 0.01 以上である 6 パラメータ（“ksorp”, “Kd”, “f”, “kvol”, “kbulk”, “kbios”）を抽出（後述する図 -8 中のグラフも参照），再度モンテカルロ法を行い，実測データがモデルパラメータの事前不確実性に起因するモデルの挙動範囲に含まれていることを確認する（図 -3 (a) および (b) のグレーバンド）。視覚的評価では，一般的に農薬が指数関数的な減衰挙動をとるという観点から，図 -3 に示すように，線形および対数スケールの両面で確認することが望ましい。

グローバル感度解析で選定したパラメータについて，ローカル感度解析を行う。ローカル感度解析ではパラメータを微小変化させ，その鋭敏比を算出し，その情報からパラメータの同定性（collinearity）を確認することができる。“FME” のパッケージでは，キャリブレーションするパラメータの組み合わせでの collinearity が 20 未満であれば安定して同定可能であるとされており，これにより本解析では最終的にキャリブレーションするパラメータを 3 つ（“ksorp”, “Kd”, “kbulk”）に絞り込んだ。これで全ての条件設定が完了となり，モデルフィッティングを行うことで，3 パラメータの最適化を行う。“FME” では勾配ベースの様々なアルゴリズムが利用可能であるが，ここでは Price（1977）が開発した非勾配型で初期値に依存しないランダムベースの方法（“pseudo”）により実行した。

モデルフィッティングによって得られた結果をもとにバリデーションを行うことでもよいが，得られたパラメータが局所解に最適化されていないかまたキャリブレーションによりパラメータ不確実性がどの程度削減されたかについては確認することができない。そこでマルコフ連鎖モンテカルロ（MCMC）法によるパラメータ不確実性解析により上述の問題点を克服する。“FME” では標準的な方法である Metropolis-Hastings 法に加え，パラメータ生成のための共分散行列を更新する Adaptive Metropolis（AM）法，パラメータ棄却率を調整する Delay-

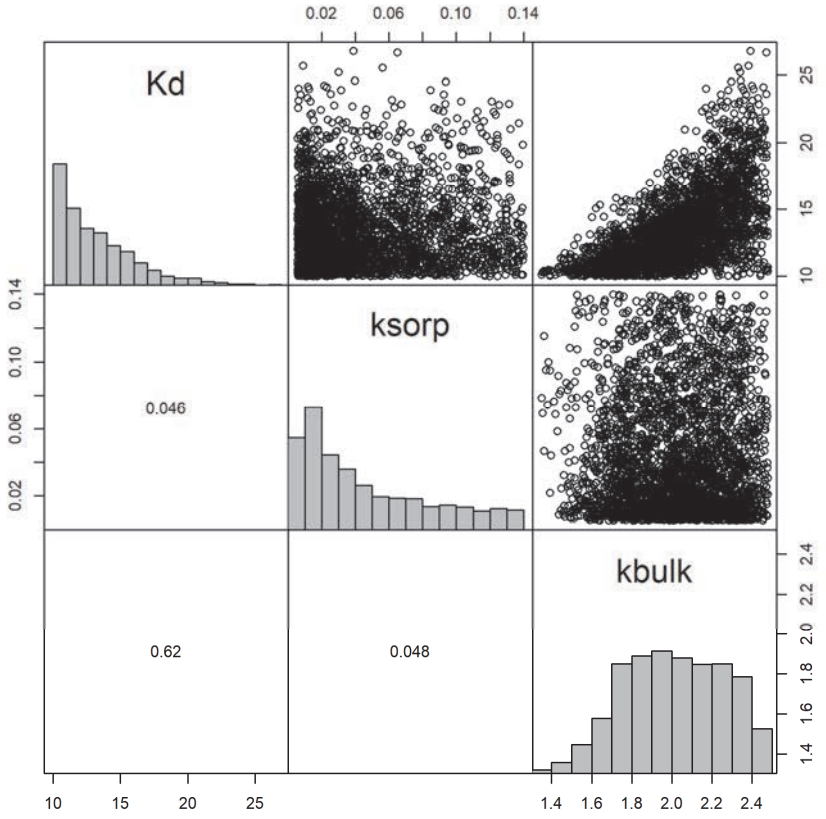

図 -4　マルコフ連鎖モンテカルロ（MCMC) 法によるパラメーターの事後分析。（右上の散布図はパラメータ間の相関図，左上から右下のヒストグラムが事後分布として採択されたパラメータの度数，左下の数値がパラメータ間の相関係数をそれぞれ示す）

rejection（DR）法およびそれらを組み合わせた DRAM 法（Haario *et al*., 2006）が利用できる。これらを駆使してパラメータの採択率をうまく調整することで，キャリブレーションパラメータを事後分布として得ることができる。本解析では総試行回数を 6000 回，0 ～ 1000 回目までの試行を burn-in 期間として削除するシングルチェインの DRAM 法として実行した。得られた事後分布を図 -4 に示す。これらの事後分布から 100 回サンプリングして得られたパラメータセットを用い，実測データに対するパラメータの事後不確実性を確認する（図 -3 (c) および (d) のグレーバンド）。その結果，図 -3(a) および (b) の結果と比較してパラメータ不確実性が大幅に削減されていることが確認できる。

MCMC 法によるキャリブレーション結果から得られたシミュレーション結果を用いてバリデーションを実施する。バリデーションの考え方およびその実施方法は利用するモデリング手法によって様々あるが，本解析では，モデルがキャリブレーションに使用した実測データを，キャリブレーション

表 -2　R パッケージ "hydroGOF" による各種統計指標の算出結果

統計指標	Rでの名前	コールドラン	パラメータ不確実性解析	目標値
ナッシュ-サトクリフ指標	NSE	0.74	0.91	1
相対ナッシュ-サトクリフ指標	rNSE	-360.85	0.96	1
重み付きR^2	bR2	0.71	0.83	1
パーセントバイアス	PBIAS %	54.9	0.8	0
実測値標準偏差に対するRMSE比	RSR	0.47	0.28	0

表 -3　PCPF-BR モデルのパラメータとモンテカルロ法実行時の変動幅

パラメータ	Rでの名前	標準シナリオ	変動幅
水田割合（%）	pfr	5	1 - 20
農薬普及率（%）	usage	10	5 - 30
河川流量（m^3/s）	hflow	3	1 - 10
農薬散布日の標準偏差	AppStdev	1	1 - 5
止水期間（day）	WHP	3	0 - 7
日畦畔浸透量（cm/day）	dseep	0.1	0.05 - 1
日水田排水量（cm/day）	ddrain	0.3	0.1 - 1
日降下浸透量（cm/day）	dperc	0.5	0.1 - 1
田面水深（cm）	Hmax	5	1 - 10

表 -4　PCPF-1R モデルのパラメータ算出に必要な農薬の主な物理的化学的性状および EPI SUITE と VEGA による予測結果

パラメータ	Rでの名前	実験値	データソース	主なメタデータ	予測値	
					EPISUITE	VEGA（適用範囲[a)]）
水溶解度 (mg/L)	CSLB	74	OECD 105	実験温度，pH	12.59	12.59（〇）
蒸気圧 (Pa)	VP	0.00065	OECD 104	実験温度	0.00053	NA
有機炭素補正土壌吸着係数[b)]	KOC	1346	OECD 107	土壌情報	1890	463（△）
加水分解半減期 (day)	DT50_HYD	200	OECD 111	実験温度，pH	NA	20.93（×）
水中光分解半減期 (day)	DT50_PHT	32.4	OECD 316	光照射条件，実験温度，pH，媒体情報	NA	NA
水中微生物分解半減期 (day)	DT50_BIOW	7	OECD 307	実験温度，土壌情報，実験系情報（乾土重，土壌水相の厚さ，実験容器形状），微生物活性	NA	26（△）
土壌中微生物分解半減期 (day)	DT50_BIOS	19	OECD 307		NA	70（△）

NA：該当モデルなし

a) 〇：モデルの適用範囲内である，△：モデルの適用範囲内か疑わしい，×：モデルの適用範囲外である

b) 実際のモデル解析ではFreundlich土壌吸着係数およびFreundlich次数を使用．ここでは比較のため４土壌の中央値を記載．

の結果どの程度の精度で再現できているかを確認する内部バリデーション（ベリフィケーション）により行う。実測データに対するモデルの出力を図示し，視覚的に得られた結果を評価したところ，パラメータキャリブレーションにより，モデルが実測データの消失過程を正確に再現できていることが判る（図 -3 (c) および (d) の実線）。また R のパッケージである ”hydroGOF”（Zambrano-Bigiarini 2017）を用いて多種多様な統計指標を算出し，統計的な評価を実施する。この内，本解析では表 -2 に示す５種の指標を参考とし，コールドランと比較して，各指標が大幅に改善されていることを確認できる。以上から視覚的・統計的に問題ないと解析者が判断すればキャリブレーションモデルを得ることができる。

(3) 予測

前項で得られたキャリブレーションモデルを用いたケーススタディとして，標準流域における河川水中濃度を予測し，その影響要因を調査した例を示す。河川水中濃度を予測するため，PCPF-1R モデルをベースに Phong *et al.*（2011）が開発した PCPF-B モデルに拡張して使用する。実際の解析では，R フォルダから ”PCPF-BR.R” および ”WBcalc.R” を読み込み，表 -3 に示す水田内での水管理，農薬使用および流域特性に関わるパラメータを ”data” フォルダから読み込み実行することで簡易的に河川水中濃度が計算できる。ここで，問題設定した流域の規模やその標準パラメータは，環境中予測濃度（PEC）算出のためのモデル流域で採用されているものを一部流用している。

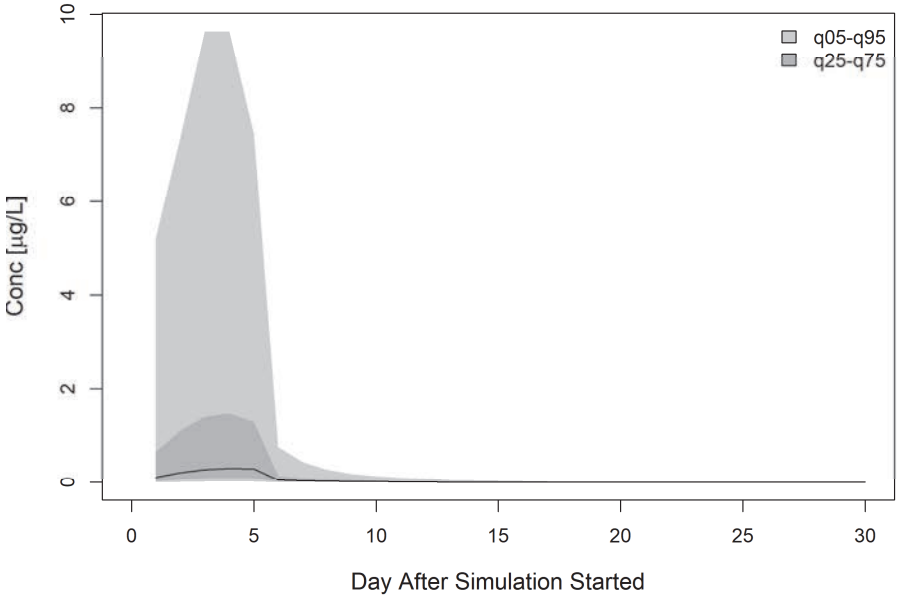

図 -5　PCPF-BR モデルによる乳剤 A 中有効成分除草剤 X の河川水中予測濃度（q05-q95 5-95 パーセンタイル，q25-q75 は 25-75 パーセンタイルおよび実線は 50 パーセンタイルを表す）

表 -3 の標準シナリオ下におけるパラメータとともに示した変動幅に基づき，ラテンハイパーキューブ法で 500 のパラメータセットを作成した。このパラメータセットと前項で得られた事後分布からサンプリングした 500 のキャリブレーションパラメータを結合し，モニタリング期間を 30 日間として河川水中濃度を 500 回繰り返し計算して求めた。この計算結果から予測された河川水中濃度を図 -5 に示す。今，対象の除草剤 X の河川水中最高濃度を，任意のエンドポイント（3 μg/L）と比較することとする。まず河川水中最高濃度に対するパラメータセット感度解析を行う。すると図 -6 に示す通り，水田内での田面水深（“Hmax”），流域内での水田

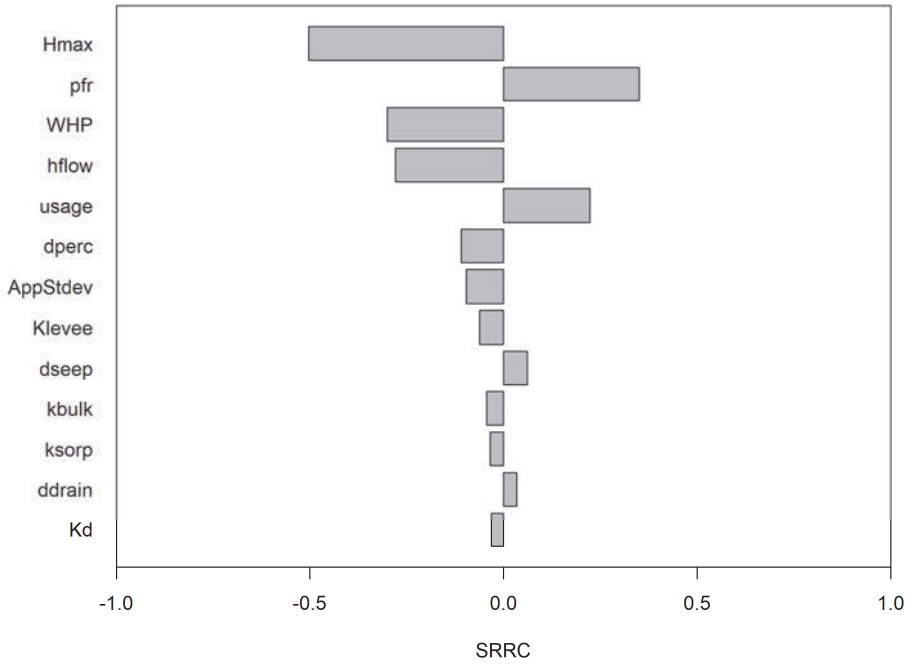

図 -6　PCPF-BR モデルによるパラメータ感度解析結果（各パラメータ記号の対応は表 -1 および表 -3 を参照）

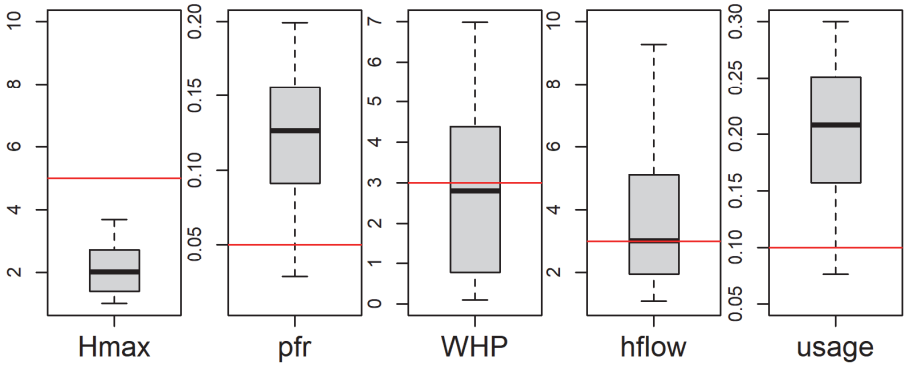

図 -7　除草剤 X のエンドポイント（3μg/L）を超過した計算（n=79）における感度解析上位 5 パラメータの変動（各パラメータ記号の対応は表 -3 を参照，また赤線は同表中の標準シナリオにおける値を示す。注：“pfr” と “usage” は小数表記）

割合（“pfr”），河川流量（“hflow”），止水期間（“WHP”）および普及率（“usage”）の 5 パラメータが大きく影響していることが分かる。この結果をもとに，予測される河川水中最高濃度がエンドポイントを超過した試行のみを抽出し，全 500 試行中 79 試行における 5 パラメータの傾向を図示すると，図 -7 の様になる。この結果から，水田割合が 10% 以上の流域で，8% 乳剤の普及率が 20% を超える場合，田面水深が 3 cm 以下の浅水管理や散布後の止水期間が 3 日程度となる水田が多いとエンドポイントを超過する傾向が確認できる。

上述の通り，水質汚濁性試験のような水中残留データがあれば，本解析によって事前に河川水中濃度を把握すること，またエンドポイントとの比較により特定の条件に絞り込むことが可能になる。これらの情報は，新規にモニタリング試験を行う際の採水地点や採水頻度の設定等，実験計画の立案に有益な情報となる。さらに PCPF-BR によるモデル解析では，モニタリングで取得した河川水中濃度の分析結果，河川流量，気象データ等を入力することで，採水地点における対象農薬の濃度推移を再現可能であるため，実験の事後評価も可能となる。このように，実験と数理モデル解析を組み合わせることにより，実験コストの最適化をより効率的に行うことができる（Holvoet *et al.*, 2007）。

3. より再現性の高い数理モデル解析に向けての諸課題

(1) ユーザー主観

ユーザー主観（user subjectivity）によるモデル操作は，モデルや扱う問題に対するユーザー自身の理解度に依存しているため，数理モデル解析における再現性に大きく影響することが知られている（Beulke *et al.* 2006; Boesten 2000）。そのため，開発者はユーザー主観が発生する操作を予め特定し，文書化しておくなどして，この影響を可能な限り排除することが望ましい。PCPF-1R モデルの様な農薬の環境中動態を予測する数理モデルでは，入力パラメータセット構築あるいはそのための生データ収集，とりわけ物理的化学的性状の選択においてユーザー主観が影響する恐れがある。これはユーザーがどのようなデータソースにアクセスして生データを取得するかという要因と，データソースの中から適切な生データを取得できているかという要因に分解される。前者は第 2 項の (2) で示したように，取得するデータソースを予め限定しておくことで概ね解決可能である。一方後者では，対

象データを説明するデータ（メタデータ）を正しく読み解けるかが重要である。例えば，実験温度が 25℃と 50℃の蒸気圧データがある場合，より解析したいデータを取得した実験温度に近い 25℃のものを選択することが適当である。このようにメタデータが少ない場合はよいが，好気的湛水土壌動態試験（OECD307）のように，メタデータが膨大にあり，かつ，記載内容が一様でない資料を読み込まなくては取得しづらいデータもある。そのため，農薬抄録や農薬の審査報告書等で情報公開が進んでいる農薬については，可能な限り関連するデータを予め抽出しておき，物理的化学的性状のプリセットデータライブラリを構築しておくことが解決策となるだろう。特に水質汚濁性試験のような水中残留試験データを PCPF-1R モデルで解析する場合，農薬の土壌中濃度は実測データには含まれないため，パラメータ化の段階で信頼性の高い分解性指標を取得しておくことが，より質の高いデータ解析結果につながると期待される。Fenner *et al*.（2016）は，残留性（persistence）の評価のため，OECD308 の水–底質中での移行性試験および OECD309 の表層水中での生分解性試験の信頼性評価を行い，最終的に 13 種の医薬品と 14 種の農薬について，解析データとそのメタデータを抽出し，残留性指標の抽出を実施した。筆者もこのような先行事例を参考に，現在，農薬抄録および農薬の審査報告書から OECD 307 に基づく土壌中動態試験成績の解析を行っており，PCPF-1R モデルと構造互換性のある数理モデル（Kondo *et al*. 2020）を用いた土壌の分解情報抽出も含めデータ整備を進めている。

(2) 機械学習モデル利用の可能性

対象農薬の物理的化学的性状に関わる生データが入手できない，またはデータの品質が著しく低い場合には，定量的構造活性相関（Quantitative Structure-Activity Relationship: QSAR）のような機械学習モデルによる予測値によって代替する手法がある。現在，比較的利用しやすいフリーソフトウェアとして，US EPA の EPI SUITE（US EPA 2012）や欧州のプロジェクトで開発された VEGA（Benfenati *et al*. 2013）がある。いずれのモデルも対象農薬の SMILES（Simplified Molecular Input Line Entry System）を入力することで目的のデータを得ることができる。表 -4 には PCPF-1R モデルで必要となる物理的化学的性状データについて，EPI SUITE および VEGA の予測値を取りまとめた。QSAR では予測値の精度も重要であるが，入力した農薬の情報が，予測モデルの適用範囲内（applicability domain: AD） であるか否かも同じように重要である。VEGA では複数の規定に基づく総合指標として AD 内か否かを判定してくれるが，EPI SUITE ではユーザー自身が判断する必要がある。さらに，QSAR がどのようなデータで訓練されたかについても確認が必要である。少なくとも土壌中での分解モデルでは，土壌が湛水状態で得られたデータではないため，VEGA による予測値を代用したとしても解析結果の信頼性は乏しい。こうした観点からも，前項のデータライブラリを活用した予測モデルの構築についても今後並行して進めていく必要がある。

(3) 文芸的プログラミングの活用

近年，査読前原稿を保存・公開可能な arXiv の出現や，一定のアクセス不可期間（エンバーゴ）を経た査読済み最終稿を各種リポジトリにセルフ・アーカイブするグリーン OA 化等，最新の学術情報の共有がより身近になっている。こうした動きはオープンネスや透明性の観点からオープンサイエンスの促進において大変重要である。しかし，数理モデルをはじめ，多くのデータ解析分野において，論文中に掲載されている方法を再現しようとする際に，どのようにしてコード化するかという問題にしばしば直面する。こうした問題は再現可能性に大きく影響する。この対応として，近年では，第 2 項(1)で示した GitHub のようなリポジトリを原稿中にリンクすることでプログラムのサンプルコードが入手可能となるよう推奨するジャーナルも少なくない（Bernard 2017)。より応用的な対応として，文芸的プログラミング（literate programming）による解析レポートや電子書籍の作成が挙げられる。文芸的プログラムは，図 -8 に示すように，コードチャンク／セグ

Rマークダウンによる文芸的プログラムの例

近藤 圭

2022-10-06

テキストチャンクの見出し

この部分がテキストチャンクである。以下のコードチャンクは，例としてグローバル感度解析で得られたSRRCを再出力する。knit して出力すると，実施した操作，出力結果および図が表示される。

```
# この部分がコードチャンクである。
X    <- CRL1[,1:8]
y    <- as.vector(CRL1[,9])
SRRC <- src(X, y, rank = TRUE)
print(SRRC)
```

```
##
## Call:
## src(X = X, y = y, rank = TRUE)
##
## Standardized Rank Regression Coefficients (SRRC):
##           original
## alp   -0.006893145
## Kd     0.206075587
## ksorp  0.036407545
## f      0.117025877
## kvol  -0.024194051
## kdiff -0.007151567
## kbulk -0.898434244
## kbios -0.121900301
```

```
plot(SRRC)
abline(h=0,col="red")
```

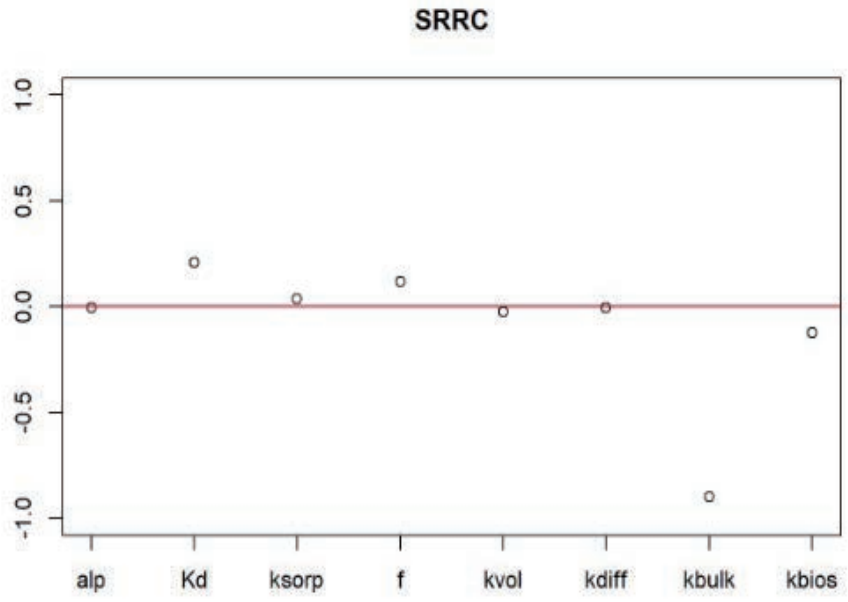

図 -8　R マークダウンを用いた文芸的プログラムの例

メントとテキストチャンク／セグメントが混在し，コンピュータではなく人間が読めるように書かれ，フォーマットされ，構成されている文書である（Zheng 2021）。これにより読者は，文章や数式による説明，それらを表現したコード，そしてコードから生成された図表が一体となった形式で解析をトレースすることができる。第 1 項でも紹介したが，R および RStudio を用いた解析では，”rmarkdown” および ”knitr” パッケージによって制御される R マークダウンという機能を利用することにより，文芸的プログラムを作成することができる。作成した R マークダウンファイル (*.Rmd) は，”pandoc” と呼ばれるドキュメント・コンバーターにより HTML，PDF，Word 等，出力形式を選択することが可能である。R マークダウンの応用として，”rticles” パッケージを使用して R マークダウンファイルを学術雑誌用フォーマットに合わせ，論文原稿を文芸的プログラムで作成することもできる（Holmes *et al*. 2021）。また ”bookdown” パッケージ（Xie 2016）を利用すればオープンフォーマットの電子書籍ファイルフォーマット規格である ePub 形式で書籍を作成することができ，モデルのユーザーマニュアル作成に有用である。このように，文芸的プログラムは，作成者の側から見ても，データ解析からレポートの作成までの一連の作業を統一した環境下で実行できるため，解析データの散逸やファイルの互換性を気にする必要もなくなり，結果として再現性を高めることに大きく寄与することが期待される。

おわりに

本稿では，数理モデル PCPF-1R および PCPF-BR による水田用除草剤の水中残留データの解析および流域中での河川水中濃度予測への応用の模擬実演を通じ，オープンソフトウェア R およびその IDE である RStudio を用いた再現性の高い解析方法についての解説を行った。さらに，より再現性を高めるための諸課題と，それらに対する取り組みについても述べた。R はスクリプトベースでの解析になるため，これまで MS Excel® のようなスプレッドシートソフトウェアでの解析を行ってきたユーザーにとって，（筆者がそうであったように）その操作を苦痛に感じる部分もあるかもしれない。しかしこの欠点を補って余りあるほどの利便性が享受できることは既に述べた通りである。さらに，R はオンライン上の集合知が充実していることも大きな利点の一つであり，高額な書籍やセミナーがなくとも一通りの知識や技術を習得することができる。こうした環境が充実している理由は，R が多くの R ユーザーにとって商用ではなく，オープンサイエンスのためのツールとして位置づけられ，利用されてきたからに他ならない。本稿を通じ，読者の方々が少しでも R に興味を持ち，使用するきっかけとなれば幸いである。また筆者が公開したコードがユーザーの研究や新たなイノベーションの支援となれば望外の喜びである。

謝辞

本研究は一般財団法人残留農薬研究所・公益目的支出計画（企業コード：A012055）の一環として実施しました。関係者の皆様に御礼申し上げます。

参考文献

Alarid-Escudero, F. *et al*., 2019. A Need for Change! A Coding Framework for Improving Transparency in Decision Modeling. Pharmacoeconomics. 37(11), 1329-1339.

青木学聡 2021. オープンサイエンスと研究データ管理の動向 . 情報処理 62(5).

Baio, G. and A. Heath 2017. When Simple Becomes Complicated: Why Excel Should Lose its Place at the Top Table. Global & Regional Health Technology Assessment 4(1), grhta.5000247.

Benfenati, E. *et al*. 2013. VEGA-QSAR: AI Inside a Platform for Predictive Toxicology, Proceedings of the Workshop Popularize Artificial Intelligence co-located with the 13th Conference of the Italian Association for Artificial Intelligence（AIxIA 2013), Turin, Italy.

Bernard, C. 2017. Editorial: Code Case - Investigating Transparency and Reproducibility. eneuro, 4(4): ENEURO.0233-17.2017.

BES, Cooper, N., 2017. A Guide to Reproducible Code in Ecology and Evolution. British Ecological Society.

Beulke, S. *et al*. 2006. User subjectivity in Monte Carlo modeling of pesticide exposure. Environ. Toxicol. Chem. 25(8), 2227-2236.

Boesten, J.J.T.I. 2000. Modeller subjectivity in estimating pesticide parameters for leaching models using the same laboratory data set. Agric. Water Manag. 44(1), 389-409.

Boulange, J. *et al*. 2012. Analysis of parameter uncertainty and sensitivity in PCPF-1 modeling for predicting concentrations of rice herbicides. J. Pestic. Sci. 37(4), 323-332.

Choi, Y.-D. *et al*. 2021. Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model Application Programming Interfaces. Environ. Modell Softw 135, 104888.

Eddy, D.M. *et al*. 2012. Model Transparency and Validation:A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7. Med. Decis. Making 32(5), 733-743.

Fenner, K. *et al*. 2016. Suitability of laboratory simulation tests for the identification of persistence in surface waters (Tech. Rep. FKZ 3715 65 415 3), Dessau, Germany. German Environment Agency (Umweltbundesamt).

FOCUS 2006. Guidance Document on Estimating Persistence and Degradation Kinetics from Environmental Fate Studies on Pesticides in EU Registration. Report of the FOCUS Work Group on Degradation Kinetics, EC Document Reference Sanco/10058/2005 version 2.0.

Haario, H. *et al*. 2006. DRAM: Efficient adaptive MCMC. Stat. Comput. 16(4), 339-354.

Holmes, D.T. *et al*., 2021. Reproducible manuscript preparation with RMarkdown application to JMSACL and other Elsevier Journals. Journal of Mass Spectrometry and Advances in the Clinical Lab. 22, 8-16.

Holvoet, K.M.A. *et al*. 2007. Monitoring and modeling pesticide fate in surface waters at the catchment scale. Ecol. Modell. 209(1), 53-64.

Inao, K. and Y. Kitamura 1999. Pesticide paddy field model (PADDY) for predicting pesticide concentrations in water and soil in paddy fields. Pestic. Sci. 55(1), 38-46.

Iooss, B. *et al*. sensitivity: Global Sensitivity Analysis of Model Outputs, R package version 1.15.2 https://CRAN.R-project.org/package=sensitivity [accessed 24 October 2018].

Ippolito, A. and G. Fait 2019. Pesticides in surface waters: from edge-of-field to global modelling. Current Opinion in Environmental Sustainability 36, 78-84.

Kondo, K. *et al*. 2012. Probabilistic assessment of herbicide runoff from Japanese rice paddies: The effects of local meteorological conditions and site-specific water management. J. Pestic. Sci. 37(4), 312-322.

Kondo, K. *et al*. 2019. Inverse analysis to estimate site-specific parameters of a mathematical model for simulating pesticide dissipations in paddy test systems. Pest. Manag. Sci. 75(6), 1594-1605.

Kondo, K. *et al*. 2020. Inverse modeling of laboratory experiment to assess parameter transferability of pesticide environmental fate into outdoor experiments under paddy test systems. Pest Manag Sci. 76(8), 2768-2780.

Kondo, K. 2022. Use of mathematical modeling and its inverse analysis for precise assessment of pesticide dissipation in a paddy environment. J. Pestic. Sci. 47(3), 146-153.

高橋康介 2018. 再現可能性のすゝめ : RStudio によるデータ解析とレポート作成 . Wonderful R / 市川太祐 [ほか] 編 , 3. 共立出版, 164 pp.

National Academies of Sciences, Engineering, Medicine 2019. Reproducibility and Replicability in Science. The National Academies Press, Washington, DC. 256 pp.

Phong, T.K. *et al*. 2011. Exposure risk assessment and evaluation of the best management practice for controlling pesticide runoff from paddy fields. Part 2: Model simulation for the herbicide pretilachlor. Pest. Manag. Sci. 67(1), 70-76.

Price, W.L. 1977. A controlled random search procedure for global optimisation. Comput J. 20(4), 367-370.

Soetaert, K. and P.M.J. Herman 2009. A Practical Guide to Ecological Modelling: Using R as a Simulation Platform. Springer Netherlands, Dordrecht.

Soetaert, K. and T. Petzoldt 2010. Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME. J. Stat. Softw. 33(3), 1-28.

US EPA, 2012. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11 or insert version used, United States Environmental Protection Agency, Washington, DC, USA.

Watanabe, H. *et al*. 2006. Simulation of mefenacet concentrations in paddy fields by an improved PCPF-1 model. Pest. Manag. Sci. 62(1), 20-29.

Wilkinson, M.D. *et al*. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 160018.

Xie, Y. 2016. Bookdown: Authoring books and technical documents with R markdown. Online version.

Zambrano-Bigiarini, M. 2017. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series, R package version 0.3-10 https://CRAN.R-project.org/package=hydroGOF [accessed 24 October 2018].

Zheng, Z. 2021. Reasons, challenges, and some tools for doing reproducible transportation research. Communications in Transportation Research 1, 100004.