Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačních technologií

Automatizace zpracování prostorových dat

Disertační práce

Autor: **Ing. Jan Pavlík**

Školitel: **doc. Ing. Pavel Šimek, Ph.D.**

Poděkování

Rád bych touto cestou poděkoval doc. RNDr. Daně Klimešové, CSc. a doc. Ing. Pavlu Šimkovi, Ph.D. za jejich odborné vedení. Dále děkuji Mgr. Janě Kholové, Ph.D. za spolupráci na vědeckých projektech, které byly řešeny v rámci praktické části práce. Poděkování patří i Mgr. Ivaně Pavlíkové za pomoc s korekturou.

Automatizace zpracování prostorových dat

Abstrakt

Práce je zaměřena na problematiku automatizace zpracování prostorových dat. Odráží aktuální trendy, jakými jsou zvyšující se nároky na zpracování dat způsobené zapojením nových technologií v oblasti IoT (Internet of Things), zapojení GIS (geografických informačních systémů) do životního cyklu prostorových dat, integraci softwarových systémů a přínosy efektivně zavedené automatizace dílčích procesů zpracování dat. Práce je prezentována formou souboru čtyř článků publikovaných v rámci výzkumu autora. Analyzuje současný stav automatizace v oblastech zemědělství, vodohospodářství a ochrany životního prostředí a identifikuje překážky, které zvyšují nutnost manuální intervence uživatelů při práci s daty. V rámci práce jsou předloženy výsledky experimentálního zavedení metodiky pro zvýšení míry automatizace v oblasti simulace zemědělské produkce.

Klíčová slova: GIS, prostorová data, automatizace, zpracování dat, big data, open data, životní cyklus dat, systémová integrace, IoT

Automation in spatial data processing

Abstract

The work is focused on the issue of automation of spatial data processing. It reflects the current trends such as increasing processing demands caused by the involvement of new technologies in the field of IoT (Internet of Things), the role of GIS (geographic information system) in the life cycle of spatial data, the integration of software systems, and the benefits of a more effectively implemented automation of partial data processing. The thesis showcases a collection of four articles published as part of the author's research. It presents the current state of data processing automation in the fields of agriculture, water management and environment protection and the identified obstacles that increase the requirements for manual intervention by users when working with data. The work includes the results of the experimental implementation of a methodology for increasing the degree of automation in the field of agricultural production simulation.

Keywords: GIS, spatial data, automation, data processing, Big Data, Open Data, data life cycle, system integration, IoT

Obsah

1 Úvod

V současné době dochází k rapidnímu rozvoji oblasti zpracování velkého objemu dat. Jedním z klíčových faktorů je výrazné zvýšení objemů prostorových dat, které je způsobeno množstvím nově aplikovaných technologií v rámci tzv. IoT (Internet of Things). Senzory umístěné v krajině přispívají do objemu dat nejenom množstvím, ale zároveň i nízkou periodicitou. Klasické zdroje prostorových dat jako např. letecké snímky, snímky z dronů nebo družicové snímky taktéž podporují trend „velkých dat“ díky kvalitnějšímu optickému vybavení, které je schopné generovat data ve vyšším rozlišení, např. VHR (Very High Resolution) satelitní snímky se pohybují okolo 30 cm na jeden pixel. Ještě před deseti lety byly tyto hodnoty v řádech jednotek metrů.

Dochází tak k celkovému zvýšení nároků na zpracování, analýzu a uložení dat. Posun v oblasti výkonu hardwaru dokáže udržet krok a z hlediska ukládání dat se čím dál více využívají cloudová řešení. Avšak zvýšené objemy dat a celkové prodloužení řetězce životního cyklu dat kladou zvýšené nároky na lidské úsilí. Tato v přeneseném smyslu manuální práce je jednou z hlavních brzd v procesech zpracování dat. Je tedy nezbytné hledat možnosti pro zlepšení míry automatizace práce s prostorovými daty, a to jak v rámci jednotlivých dílčích kroků procesu zpracování, ale zejména pak při přechodu mezi jednotlivými fázemi či softwary.

Nově získaná prostorová data mají vysoký potenciál v oblastech přírodních věd jako např. ochrana životního prostředí nebo vodohospodářství a v oblasti zemědělství. Trend tzv. chytrého zemědělství je stále na vzestupu a možnost vytěžit užitečné informace z těchto nových zdrojů může výrazně pomoci z hlediska snižování nákladů a odstranění negativních vlivů zemědělské produkce na životní prostředí. Zároveň se také jedná o jeden z klíčových prvků, jak omezit dopady budoucích klimatických změn.

2 Výzkumná mezera – Cíle práce

První částí disertační práce byla analýza dostupných literárních vědeckých zdrojů, zabývajících se problematikou automatizace zpracování prostorových dat. Na základě syntézy poznatků získaných z těchto zdrojů byly identifikovány dílčí oblasti obsahující výzkumné mezery, u nichž byl potenciál, že by výsledky této disertační práce mohly danou mezeru vhodně doplnit.

Jednou z těchto oblastí jsou zvyšující se nároky na zpracování vzhledem k vzrůstajícím objemům dat. Jedná se tedy o konkrétní aplikaci principu Big Data na oblast prostorových dat a automatizace jejich zpracování. Zvyšující se hardwarové nároky jsou dostatečně analyzovány již existujícími publikacemi, avšak zvyšující se potřeba uživatelské intervence není příliš řešena, zejména s ohledem na jednoduché datové úpravy. Jednotlivé dílčí kroky jsou často automatizovány dedikovanými programy nebo skripty. Díky narůstající délce životního cyklu dat, zvyšování počtu jednotlivých fází zpracování a nedostatečné standardizaci datových formátů, se však tyto úlohy akumulují a v souhrnu představují zásadní překážku pro celkový proces automatizace.

Druhá oblast je znovu-využitelnost vyvinutých podpůrných softwarových nástrojů. Ve většině publikací, pokud je potřeba např. napsat nový skript pro dílčí automatizaci určitého kroku zpracování, autoři zřídkakdy diskutují živostnost daného nástroje. Možnost jeho opětovného použití je tak převážně vázána na konkrétní situaci, což snižuje efektivitu vynaloženého úsilí na vytvoření takového programu. Dostupné informační zdroje se tedy zaměřují převážně na konkrétní výsledky zkoumání, nikoliv na opětovnou využitelnost použitého software.

Třetí oblast, která skýtá potenciál pro hlubší prozkoumání, je pak rychlost zpracování dat. Úlohy, které zpracovávají historická data, většinou nevyžadují urychlené zpracování a nekladou tak výrazné nároky na míru automatizace. Naopak úlohy tzv. real-time zpracování dat vyžadují téměř stoprocentní automatizaci, bez které nemohou fungovat. Existuje zde však jistá mezera pro úlohy uvnitř spektra mezi těmito dvěma extrémy. Jedná se o situace, kdy rychlost zpracování a k tomu vázaná automatizace není nezbytně nutná, avšak snížení časové prodlevy může dramaticky zvýšit využitelnost vytěžené informace. Jedná se tedy o úlohy, které sice nefungují v reálném čase, ale zpracovávají poměrně nedávná data, tedy v řádu týdnů nebo měsíců.

Hlavním cílem této práce je tedy zaměřit se na tyto dílčí oblasti a v rámci praktického experimentálního výzkumu navrhnout takový scénář, aby dosažené výsledky zasahovaly do jedné, nebo více z těchto zjištěných mezer.

Konkrétně se jedná o návrh metodiky pro zlepšení řešení modelové situace zpracování zemědělských simulací. Na základě analýzy vědecké literatury v této oblasti bude definováno typické výchozí řešení a budou identifikovány jeho nedostatky. Cílem je navrhnout a experimentálně ověřit metodikyu pro zvýšení míry automatizace celého procesu zpracování simulací nebo dílčích kroků v rámci životního cyklu použitých prostorových dat. Důležitým prvkem navrženého řešení je jeho praktické uplatnění v reálném prostředí. Výzkum zaměřený na simulace zemědělské produkce na Katedře informačních technologií v současnosti aktivně probíhá ve spolupráci se zahraničními partnery, kteří jsou na celosvětové úrovni leadeři v této výzkumné oblasti. Problematika tedy zapadá do rámce mezinárodní expertní vědecké simulační komunity. Získané výsledky budou mít potenciál významně posunout stávající metodický postup směrem k řešení s vysokou mírou automatizace, což umožní celkové rozšíření tohoto výzkumného směru.

3 Metodická poznámka

Práce kombinuje jak teoretické, tak empirické výzkumné metody. Teoretické metody byly použity primárně při poznávání zvolené problematiky. Analýza je metoda založená na rozkladu daného problému na jednotlivé složky a jejich zkoumání (Široký, 2011). Metoda analýzy byla tedy uplatněna při vymezení dílčích oblastí výzkumu a obecně při studiu vědecké literatury. Následnou syntézou získaných poznatků byl pak vytvořen ucelený přehled zkoumané problematiky a vymezeny její vnitřní zákonitosti. Metoda porovnání (komparace) umožňuje odhalit shody či rozdíly zkoumaných jevů a objektů. V případě částečné shody obsahu zkoumané problematiky je také možné využít metodu analogie. To znamená vlastnosti jednoho zkoumaného jevu uplatnit na jiný v případě jejich dostatečné shody (Široký, 2011). Pomocí metody zobecnění (generalizace) lze pak provést analýzu zkoumané problematiky jako celku, kdy jednotlivé dílčí informace, vztahující se na konkrétní implementaci, lze použít pro danou problematiku (Široký, 2011). Metody komparace, analogie a generalizace byly využity zejména za účelem specifikace výchozí modelové situace, tak aby odpovídala standartnímu řešení pro danou oblast.

Z empirických výzkumných metod bylo využito primárně měření. Jedná se kvantitativní metodu, která srovnává vlastnosti zkoumaných jevů či objektů. Platí, že pro měření je nutné, aby zkoumané vlastnosti byly konstantní za stejných podmínek (tzv. „ceteris paribus“) a aby intenzita jednotlivých vlastností byla kvantitativně vyjádřitelná pomocí porovnávacích vztahů (Široký, 2011). V případě číselného vyjádření pak existuje možnost provádění matematických operací. Konkrétně v této práci bylo těchto metod využito např. při měření hardwarové náročnosti (využití procesoru, RAM), časové náročnosti výpočtu, metrik komplexity zdrojového kódu apod. Druhou použitou empirickou metodou je experiment. Je to pokus, při kterém je realizován zvolený postup daným způsobem (Široký, 2011). Pro úspěšný experiment je také důležité pracovat v kontrolovaných a řízených podmínkách. Pro účely této práce se tedy bude jednat o experimentální implementaci navrženého metodického postupu automatizace nad zvolenou datovou sadou v prostorách specializovaných vědeckých laboratoří PEF.

3.1 Rámcový metodický postup

- Studium a analýza dostupných literárních informačních zdrojů za účelem zpracování přehledu současného stavu řešené problematiky a vymezení výzkumné mezery.
- Formulace pracovních hypotéz výzkumu, případné zúžení problematiky za účelem vymezení vhodného rozsahu výzkumu vzhledem k zjištěným nedostatkům automatizace tak, aby výsledky práce dokázaly reálně postihnout existující problematiku a byly jak vědecky relevantní, tak prakticky přínosné.
- Specifikace modelové situace pro praktické ověření. Návrh tohoto scénáře bude vycházet z provedené analýzy informačních zdrojů a zjištěné výzkumné mezery a bude vhodně zapojen do probíhajícího výzkumu katedry.
- Spolupráce v rámci vědeckého kolektivu katedry a jejich externích partnerů za účelem získání dat a přípravy experimentu.
- Návrh konkrétního řešení dílčích překážek pro efektivní automatizaci zpracování dat vymezených pro danou modelovou situaci.
- Experimentální ověření pracovních hypotéz formou praktické implementace navrženého metodického postupu. K implementaci budou využity prostory a hardwarové prostředky vědeckých laboratoří PEF.
- Prezentace dílčích částí výzkumu formou publikací ve vědeckých periodikách a na odborných konferencích.
- Celkové vyhodnocení praktického experimentálního ověření navržené metodiky pro automatizaci zpracování dat ve zvoleném scénáři, formulace případných doporučení.
- Zhodnocení výsledků práce, generalizace získaných poznatků a diskuse možností využití metodiky v rámci navazujícího výzkumu.

4 Literární přehled – současný stav výzkumu

Analýza současného stavu výzkumu byla zaměřena na tři hlavní oblasti a to: prostorová dat, automatizace zpracování dat a simulace zemědělské produkce. Při analýze byly primárně použity aktuální vědecké články a příspěvky na konferencích indexované v databázi Web of Science (případně v databázi Scopus), které tematicky zasahují do jedné ze zkoumaných oblastí.

4.1 Prostorová data

4.1.1 Prostorová data

Pro lokalizaci prostorových se používá převážně systém GPS (Global Positioning System). Při zpracování v dedikovaném GIS software je potřeba prostorové projekce na referenční elipsoid / geoid. Vhodnou volbu a nastavení takové projekce řeší např. (Gosling a Symeonakis, 2020). Celosvětově je uznávaným standardem projekce WGS 1984 (World Geodetic System). Pro lokální výzkum lze použít zobrazení, které pro dané území vykazují nižší zkreslení. V České republice je to nejčastěji tzv. křovákovo zobrazení - systém S-JTSK (systém jednotné trigonometrické sítě katastrální) (Nařízení vlády č. 430/2006 Sb.). Avšak reálně se ve výzkumné činnosti s těmito lokalizovanými systémy příliš nepracuje, jelikož korekcí GPS lze dosáhnout dostatečné přesnosti, což ukazuje např. (Chen a kol., 2017).

4.1.2 GIS

Z hlediska nejčastěji používaného GIS software převládají aplikace desktopového typu, a to zejména pro výzkum menšího rozsahu, kdy se nezpracovává velké množství dat. Ve většině případů se používá buď software ArcGIS, nebo některý z dostupných open-source nástrojů (QGIS, GRASS GIS). Jako příklad lze uvést třeba (Parent a kol., 2022) nebo (Kalbarczyk a Kalbarczyk, 2021). Existuje zde i možnost akcelerace výpočtu zapojením grafické karty, což zkoumá např. (Tischler, 2016). Jedním z nastupujících trendů je ale ústup od tohoto klasického modelu, a místo toho je funkcionalita GIS zajištěna formou cloudu (Bediroglu a Colak, 2017), (Bellman a Pupedis, 2016) nebo webové služby (Kulawiak a kol., 2019). Jedná se tedy o princip tzv. SaaS (software as a service) (Luo a kol., 2012).

4.1.3 Datové formáty, standardizace

Standardizace v oblasti prostorových dat je adekvátně zajišťována organizací OGC (Open Geospatial Consortium) (OGC, 2022). Z datových formátů se nejčastěji používá tzv. shapefile, jako příklad lze uvést (Singh a Bawa, 2016) nebo (Abreu a kol., 2020). Specifické datové formáty, např. v oblasti BIM (Building Information Modelling), je možné v případě potřeby do shapefile transformovat, což ukazuje např. (Zhu a kol., 2019). Obecně platí, že mezi běžně používanými formáty existují softwarové nástroje pro snadnou konverzi, což lze např. vidět u (Yu a Zhang, 2013) při konverzi formátu GML (Geography Markup Language).

4.1.4 Open Data

Zásadní význam pro výzkumné účely má dostupnost relevantních prostorových dat, jak zmiňuje např. (Markovinovic a kol., 2022) nebo (Amaral a Cesar Lima D'Alge, 2009). Dostupná infrastruktura otevřených dat pro každý region se odvíjí od vynaložených prostředků k získání a obnově dat v daném státě, což dokládá (Srivastava, 2018). V ČR má v tomto ohledu zásadní význam INSPIRE geoportál, který nabízí data, ovšem převážně ve formátu WMS (web map service), který není příliš vhodný pro vědecké účely, jelikož primárně umožňuje pouze vizualizaci, nikoliv práci s daty (Řezník, 2013). Problematika WMS, a obecně open data, která ve skutečnosti nejsou otevřená, je vědeckou komunitou vnímána velmi negativně (Abella a kol., 2022), (Afful-Dadzie, 2017). Konkrétně pro oblast Evropy je ale dostupnost dat v poslední době na vzestupu, a to hlavně díky tomu, že instituce Evropské Unie tlačí na zvýšení otevřenosti dat (Směrnice EU 2007/2/ES), (Směrnice EU 2019/1024).

4.1.5 IoT

Jednou z oblastí, kde se rozvoj informačních a komunikačních technologií v posledních letech výrazně projevuje, je tzv. internet věcí - IoT (Internet of Things). Jedná o zařízení připojená do internetové sítě, která se vyznačují snadnou dostupností a škálovatelností, kdy je možné síť senzorů nebo chytrých zařízení snadno rozšiřovat a nasazovat i v jinak nepříznivých podmínkách, což řeší např. (Poursafar a kol., 2017). IoT s sebou přináší ale i řadu úskalí díky rapidnímu nárůstu objemu dat a potřebě tzv. real-time zpracování.

Senzorická zařízení dávají vědcům možnost pořídit velké množství dat za rozumnou finanční pořizovací cenu. Jako příklad výzkumu zabývajícího se finančními náklady lze

uvést např. (Joshva Devadas a kol. 2019). Dochází k budování rozsáhlých platforem a IoT infrastruktury pro práci s takto získanými daty. Využití takových platforem zkoumá např. (Hejazi a kol., 2018). V případě robustních sítí, jaké se využívají ve velkých městech pro tzv. Smart Cities je pak ale potřeba síť dostatečně optimalizovat, tomu se věnuje např. (Anagnostopoulos a kol., 2015). Kromě několikanásobného navýšení celkového množství získaných dat, které popisuje třeba (Agbo a kol., 2019), pak technologie IoT umožňují i lepší propustnost dat celým řetězcem zpracování, což umožňuje vznik nových aplikací na principu real time GIS, jako příklad lze uvést (Isikdag a Pilouk, 2016) nebo (Kaippilly a kol., 2018). Pro zpracování a vizualizaci takovýchto dat lze pak využít systémy web GIS, které umožnují práci v reálném čase, což lze vidět u (Nourjou a Hashemipout, 2017).

4.2 Automatizace zpracování dat

Způsob, jakým je zabezpečena automatizace zpracování dat se liší projekt od projektu. V oblasti prostorových dat tedy převládá způsob řešení „na míru“. Každý řetězec životního cyklu dat pracuje na odlišné bázi. Liší se použitý software a hardware, zdroje, typ a formát dat, způsob zpracování i evaluace výsledků. Lze se tedy setkat s různými způsoby automatizace. Může se jednat o rozsáhlejší automatizace celého procesu, tzv. workflow, jako například u (Kliment a kol., 2015). Poměrně časté jsou situace, kdy dílčí úlohy v rámci GIS jsou automatizovány za účelem snížení manuální práce, např. (Gegelova a kol., 2014). Typicky se jedná o využití již existujícího automatizačního nástroje, který je zabudován přímo do GIS softwaru. V případě nejpoužívanějšího ArcGIS se jedná buď o automatizaci pomocí ArcGIS task designer modulu nebo o automatizaci pomocí externích skriptů v jazyce Python, který ArcGIS podporuje, jako například u (Abdella a Alfredsen, 2010) nebo (Rahmati a kol., 2018). Jiné softwary mohou podporovat obdobný způsob automatizace, například formou skriptu v jazyce R pro GRASS GIS (Grippa a kol., 2017).

Kromě automatizace zpracování dat je také důležitá automatizace přípravy a transformace dat, a to zejména u delších řetězců zpracování, kdy data přecházejí z jednoho softwaru do jiného a je potřeba měnit nebo upracovat data tak, aby byla dodržena dobrá propustnost celého systému. Neúnosnost manuálního čištění a kontroly dat zmiňuje např. (Skoogh a kol., 2010). Jako příklad automatizace datové transformace lze uvést např. (Ureche a kol., 2015) nebo (Solihin a kol., 2017).

Nároky na zvýšení míry automatizace jsou způsobené převážně již zmíněným rozsahem nově používaných dat. Zejména data z IoT mnohdy vyžadují tzv. high-throughput

zpracování, ve kterém je dobře zavedená automatizace velice efektivní (Leonard a kol., 2014). Důležitou roli zde hraje i škálovatelnost, jak ukazuje např. (Zhang a kol., 2019). Z tohoto důvodu se velice často zpracování dat přenáší taktéž do cloudového prostředí.

Při zapracování dat vlastním hardwarem se nejčastěji uplatňuje buď princip paralelizace v případě, že je hardware dostatečně výkonný, nebo princip distribuce. Distribuované zpracování je důležité hlavně v případě hůře vybaveného výzkumného týmu, kdy je možné využít větší množství slabších strojů, jak ukazuje např. (Bartoněk, 2016). Pro řízení takovéto automatizace se dají použít existující automatizační nástroje jako třeba HTCondor (Zhao a kol., 2013).

4.3 Simulace zemědělské produkce

V oblasti simulace zemědělské produkce primárně existují dva běžně používané postupy. Prvním je využití softwaru APSIM (Agricultural Production Systems Simulator), který umožnuje pro zvolený rostlinný genotyp simulovat produkci v dané lokalitě. Jako zdroj dat o počasí slouží převážně volně dostupné satelitní snímky z misí NASA (National Aeronautics and Space Administration). Software APSIM je primárně používaný jako desktopová aplikace pro jednorázové simulace ale umožňuje i automatizované spouštění z příkazové řádky, které využil např. (Vogeler a kol., 2011). Je však potřeba automaticky vygenerovat řídící soubory simulací, které jsou na bázi XML a poté program dávkově spouštět. Tento software však není příliš dobře optimalizován pro velké množství simulací (v řádů tisícovek až miliónů) a faktorizační modul, který je v APSIM implementován pro automatické generování simulací neumožňuje sám o sobě dávkové zpracování, což zmiňuje např. (Fainges, 2015). Množství souběžných simulací je tedy omezeno hardwarovou kapacitou stroje. Software APSIM se však stále rozvíjí a v budoucnu se chystají vylepšení pro zvýšení míry automatizace (Holzworth a kol., 2018). Komplikovaný proces kontroly a přípravy dat pro simulace pomocí APSIM, zejména v případě integrace datových zdrojů z různých lokalit popisuje (Ojeda a kol., 2021). V této oblasti se tedy skýtá značný prostor pro zavedení automatizace datové přípravy.

Druhou možností je pak použití SSM (simple simulation model), který může být vyvinut na míru dané rostlině nebo pro zvolenou situaci (Soltani a kol., 2021). Typicky jsou modely SSM vyvíjeny v prostředí Microsoft Excel, což vědcům umožňuje snadnou práci s daty, avšak pro potřeby zpracování velkého počtu simulací je toto naprosto nevyhovující.

Díky modulární struktuře SSM lze však celý model snadno přeprogramovat v prostředí vhodnějším pro automatické zpracování, jako např. pomocí jazyka C# nebo Python.

Z nedostatku zkoumané literatury zaměřené na IT aspekt v této oblasti je ovšem zřejmé, že vědecká komunita, která v současné době simulační nástroje APSIM a SSM využívá, se skládá převážně ze specialistů v oblasti agronomie a pěstitelství.

.

5 Výsledky

5.1 Podpora rozhodování v oblasti zemědělství

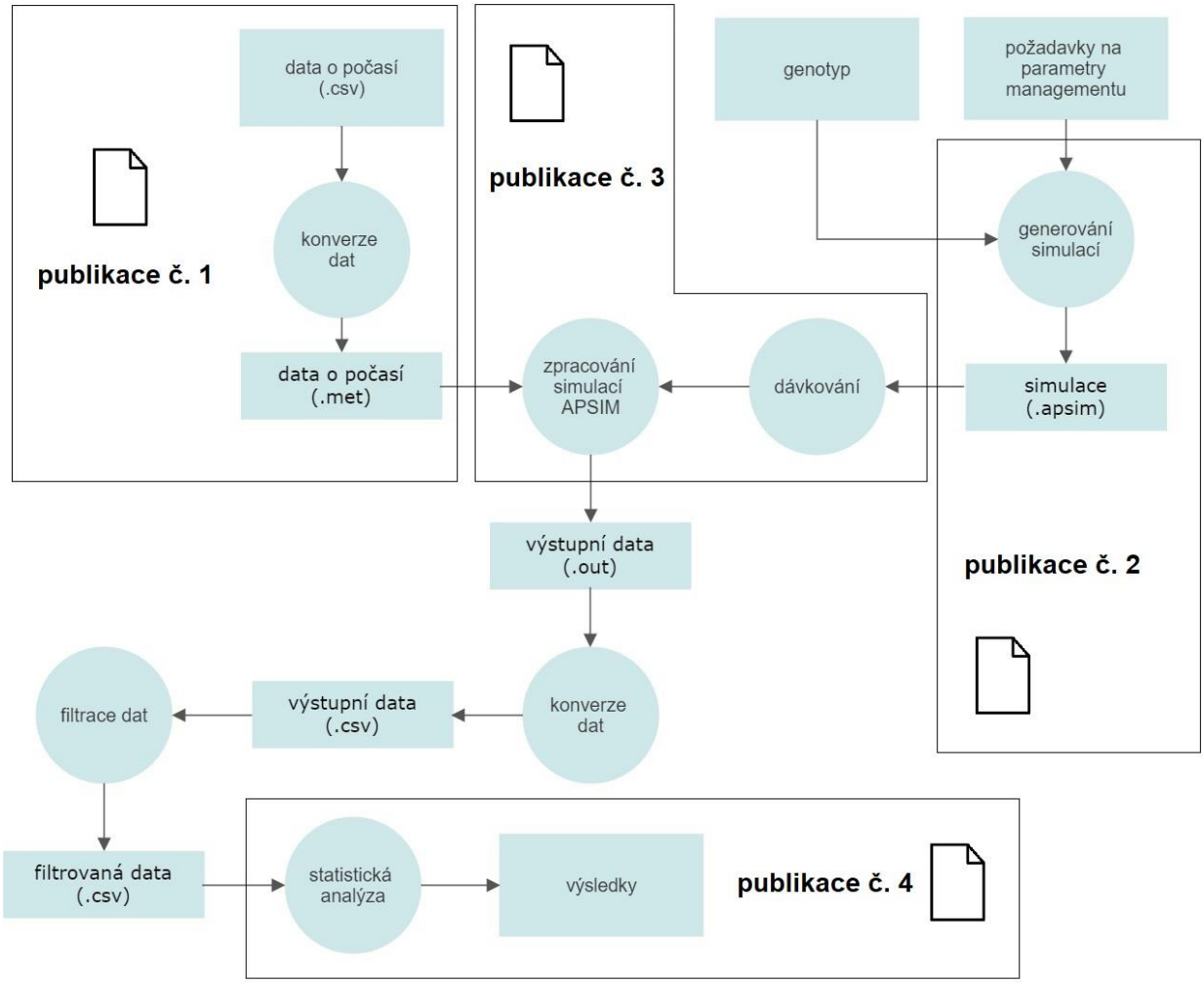
V rámci spolupráce Katedry informačních technologií s jejími zahraničními partnery je od roku 2019 prováděn výzkum zaměřený na zpracování dat pro podporu rozhodování v zemědělských a pěstitelských procesech. V praktické části disertační práce je konkrétně rozveden výzkum zabývající se zpracováním simulací zemědělské produkce. Tyto simulace fungují na principu tzv. GxExM (Genotype, Environment, Management), kdy genotypem se rozumí fyziologické vlastnosti pěstované rostliny. Data o prostředí jsou například typ půdy nebo počasí. Management se zabývá způsobem pěstování, tedy například využití hnojiv, zavlažování, hustota setí apod.

Výsledkem jedné simulace je výstup zachycující množství sklizené produkce (zrno a biomasa). Jelikož účelem je podpora rozhodování zemědělců pro optimalizaci výnosu, je nutné zjistit, jaké nastavení výchozích parametrů povede k nejlepšímu výsledku – tedy jaký genotyp rostliny použít a jakým způsobem ji pěstovat (enviromentální data jsou pro danou lokalitu konstantní). Kombinací zvolených parametrů tak vzniká sada simulací, jejíž velikost závisí na počtu proměnlivých parametrů a počtu možností nastavení pro každý z nich. Výsledky všech simulací je poté nutné filtrovat a statisticky vyhodnotit pro nalezení optimálního nastavení jednotlivých parametrů. Tyto výsledky pak umožňují zemědělcům maximálně využít potenciál pěstované rostliny.

5.2 Výchozí situace

Specifikace výchozí modelové situace je založena na analýze současného stavu vědeckého výzkumu v oblasti zpracování simulací zemědělské produkce. Jednotlivé kroky a použité postupy tedy odpovídají tomu, jakým způsobem je v běžné vědecké praxi daný problém většinou řešen.

V současné době výzkumníci většinou používají software APSIM. Na obrázku č. 1 je znázorněn diagram datového toku, který zachycuje, jakým způsobem probíhají jednotlivé části procesu zpracování simulací. Zároveň jsou zde vyznačeny oblasti, kterým se věnují jednotlivé publikace prezentované v rámci disertační práce.



Obrázek č. 1- Diagram datového toku – výchozí situace

Nastavení parametrů pro management a volba genotypu závisí na konkrétním výzkumném cíli a probíhá ve spolupráci s agronomy a pěstiteli. Data o počasí je typicky nutné převést ze zdrojového formátu .csv do formátu .met. K tomu může být použit dílčí skript vytvořený pro tento účel. Je potřeba ho správně inicializovat (nastavit konvence pojmenování souborů a jejich umístnění) a poté spustit.

Překážkou v tomto kroku může být nedostupnost dat pro danou lokalitu. Tomuto problému se věnuje publikace č. 1, která je zaměřená na oblast open data, tedy otevřených dat. Tato publikace zachycuje situaci v České republice, konkrétně to, jaká data jsou volně dostupná pro vědecké účely v této oblasti. Je prezentován přehled dostupných datových repositářů a diskutována vhodnost způsobu poskytování těchto dat.

Publikace č. 1

typ:	vědecký článek
název:	Usability of IoT and Open Data Repositories for Analyzing Water Pollution. A Case Study in the Czech Republic
autoři:	**PAVLÍK, J.**, Hrnčírová, M., Stočes, M., Masner, J. and Vaněk, J.
rok:	2020
vydáno v:	ISPRS International Journal of Geo-Information
indexováno:	Web of Science, impakt faktor 3.165 Scopus, CiteScore 5.0
odkaz:	https://doi.org/10.3390/ijgi9100591

V dalším kroku je provedeno vygenerování simulací. V případě nízkého počtu simulací lze využít faktorizační nástroj přímo zabudovaný v rámci softwaru APSIM. Pro vygenerování velkého množství simulací je možné použít např. program, který funguje na principu záměny číselných hodnot. Po načtení základní simulace projde možnosti nastavení hodnot parametrů a pro každou z nich v řídícím souboru simulace na příslušném řádku nahradí číselnou hodnotu pro daný parametr. Každá simulace pak může být uložena jednak zvlášť nebo je možné seskupení více simulací do jednoho souboru pro snadnější dávkové zpracování. Program zajišťující tuto funkcionalitu je taktéž nutné inicializovat (nastavení umístění souborů, konvence pojmenování, volba lokality, velikost vygenerované dávky).

Příprava dat před vlastním spuštěním simulace byla detailně popsána v publikaci č. 2, která se věnuje zejména procesu automatizace v této fázi. V uvedené publikaci je také diskutována problematika nastavení velikosti dávky, pomocí které lze dosáhnout lepší časové optimalizace výpočtu v rámci APSIM.

Publikace č. 2

typ:	vědecký článek
název:	Data Pre-processing for Agricultural Simulations
autoři:	Jarolímek, J., **PAVLÍK, J.**, Kholova, J. and Ronanki, S.
rok:	2019
vydáno v:	Agris on-line Papers in Economics and Informatics
indexováno:	Scopus, CiteScore 2.0
odkaz:	https://doi.org/10.7160/aol.2019.110105

Hlavním krokem je vlastní spuštění simulací v programu APSIM. Pro dávkování lze využít program, který funguje na principu fronty, nebo vhodný automatizační software (např. HTCondor, který se často využívá právě pro automatizaci APSIM). V tomto kroku je nutné spustit APSIM v režimu příkazového řádku a na vstupu dodat umístění souboru

s dávkou, která obsahuje seskupení řídících souborů simulací. Po dokončení výpočtu následuje spuštění další dávky.

Možnosti, jak tento proces automatizovat jinými způsoby, jsou detailněji zpracovány v publikaci č. 3. Konkrétně se tato publikace věnuje běžně používaným způsobům automatizace, kde převažují dvě základní linie – využití automatizačního software (většinou HTCondor) nebo vytvoření vlastního programového nástroje, který lze lépe přizpůsobit specifickým požadavkům dané výzkumné činnosti.

Publikace č. 3

typ:	příspěvek na konferenci
název:	Support tools for agricultural production simulation processing
autoři:	**PAVLÍK, J.**, Vaněk, J., Masner, J., Stočes, M., Očenášek, V.
rok:	2020
vydáno v:	sborník konference: 9th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2020)
indexováno:	Scopus, CiteScore 1.1
odkaz:	http://ceur-ws.org/Vol-2761/HAICTA_2020_paper67.pdf

Vlastní zpracování simulací je náročné z hlediska délky výpočtu. Zásadní nevýhodou v této části je hardwarová závislost softwaru APSIM, což je potřeba řešit pro každý stroj zvlášť. APSIM totiž alokuje velké množství paměti RAM vzhledem k velikosti dávky zpracovávaných simulací. Při malé velikosti dávky dochází k časovým ztrátám způsobených režií a u velké dávky APSIM havaruje v důsledku nedostatku paměti. Pro optimalizaci je tedy nutné pro každý stroj zjistit vhodně nastavení velikosti dávky tak, aby byla co možná největší, avšak nepřesáhla kritickou hranici pro havárii.

Další překážkou pro automatizaci v tomto kroku je velikost výstupních souborů. V typické situaci (a také v závislosti na počtu simulací a lokalit) kapacita běžných PC diskových úložišť neumožňuje zpracování všech simulací najednou. Je tedy nutné zpracování přerušovat a výstupní data přehrávat do velkokapacitního úložiště nebo do cloudu pro uvolnění místa. Míru automatizace v tomto kroku je tak možné výrazně zlepšit vhodným zapojením výpočetních strojů do sítě, což umožní přesun dat za běhu.

Poslední fází je převedení výstupních dat z formátu .out do .csv, jejich filtrace a následné statistické vyhodnocení. Konverzi dat lze zajistit stejným způsobem jako převedení vstupních dat. Pro filtraci a statistické vyhodnocení se dá použít existující softwarové

nástroje (např. SAS, TIBCO Statistica), případně je možné je provést přímo uvnitř tabulkového editoru Microsoft Excel s pomocí maker.

Na základě analýzy výchozího stavu řešení byly identifikovány následující oblasti s nedostatky:

- **Fragmentace jednotlivých kroků**
  Dílčí části výpočtu jsou většinou automatizovány jednoúčelovými skripty / programy, které musí být samostatně inicializovány a spuštěny.
- **Nekonzistence datových formátů**
  APSIM vyžaduje vstupní data v nestandartním formátu, výstupní data je taktéž nutné převádět nebo upravit, během procesu zpracování simulací tak dochází ke zbytečným zápisům a čtení dat.
- **Hardwarová závislost**
  V kroku tvorby dávky je optimalizace závislá na konkrétním použitém hardware a je potřeba experimentálně najít kritickou hranici, při které APSIM havaruje kvůli nedostatku paměti.
- **Nedostatečná kapacita úložného zařízení**
  V případě nedostatečného síťového propojení mezi výpočetním hardwarem a úložištěm dat musí být zpracování simulací přerušováno za účelem uvolnění místa na pevném disku.

5.3 Metodika PlaGroSim

Zjištěné překážky snižují míru automatizace procesu zpracování simulací. Za účelem zlepšení situace byly proto navrženy tři základní rámcové okruhy:

1. Integrace periferních procesů do centrálního modulárního softwaru

Tento krok je v rámci optimalizace automatizace stěžejní, jelikož vede ke zlepšení situace z více hledisek. Tím, že bude vyvinut jeden centrální modulární software, který bude schopen zajistit funkcionality okrajových procesů, dojde ke snížení „manuální“ práce potřebné pro nastavování a spouštění dílčích programů. Nastavení bude provedeno centrálně pouze jednou. Stejně tak spouštění výpočtu bude koncentrováno do jednoho místa.

2. Nahrazení softwaru APSIM modelem SSM

Software APSIM není primárně koncipován jako software pro strojové zpracování velkého množství dat. Zároveň se jedná o rigidní prvek celého procesu, kterému se ostatní

kroky musí podřizovat (např. převodem na nestandartní datové formáty). Dále pak APSIM přináší velice problematický prvek hardwarové závislosti. SSM lze snadněji integrovat v rámci celého řetězce zpracování jako jeden dílčí modul.

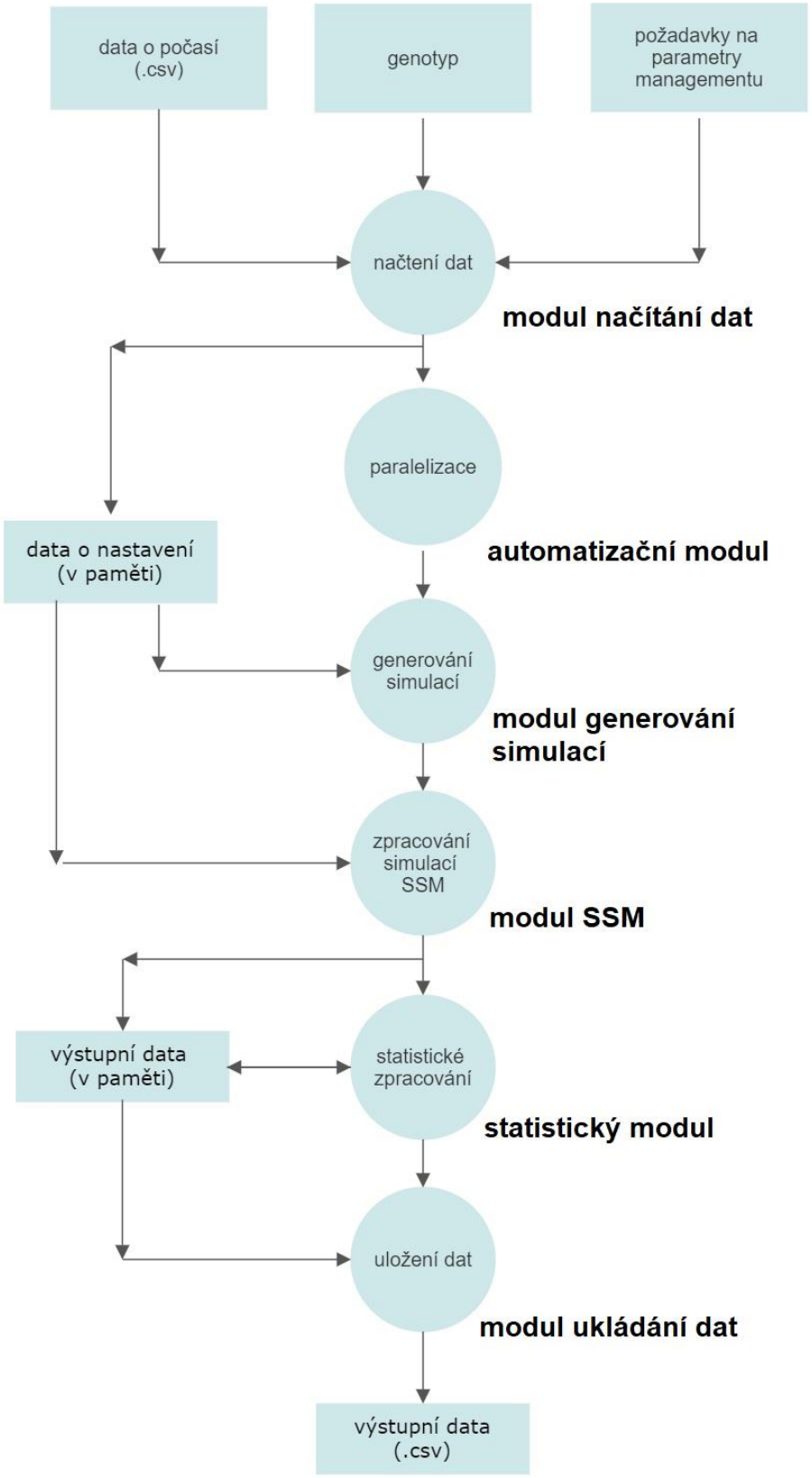
3. Kontinuální práce s daty v paměti

Vstupní data by měla být načtena pouze jednou, stejně tak výstupní data by měla být zapsána pouze jednou ve své finální podobě. Tento krok s sebou přináší omezení počtu simulací vzhledem k velikosti operační paměti. Umožňuje ale zvýšení celkové rychlosti výpočtu odstraněním přebytečných zápisů a čtení (které jsou na velkokapacitním mechanickém disku pomalé). Dále je pak možné díky tomuto postupu integrovat i statistické zpracování výsledků do centrálního software jako jeden z modulů.

Metodika PlaGroSim (Plant Growth Simulation), byla navržena tak, aby postihla tyto tři základní pilíře. Jednotlivé procesy jsou řešeny jako dílčí moduly uvnitř komplexního softwarového nástroje. Tento přístup umožňuje měnit nebo vylepšovat jednotlivé moduly, aniž by byl narušen celkový proces zpracování. To zajistí vysokou míru udržitelnosti vyvíjeného software a zároveň lze snadno přidávat další moduly pro dodatečné funkcionality. Základním principem je rozdělení na šest následujících modulů:

- **Modul načítání dat** zpracovává všechna vstupní data týkající se parametrizace GxExM a dalších inicializačních nastavení.
- **Automatizační modul** umožňuje paralelní více vláknové zpracování. Tím dochází k optimalizaci využití výpočetní kapacity.
- **Modul generování simulací** na základě načtených požadavků na parametrizaci GxExM nastavuje konkrétní hodnoty proměnných a vytváří tak kombinaci všech možných simulací.
- **Modul SSM** zajišťuje vlastní provedení simulace.
- **Statistický modul** provádí výpočty nad výstupními daty simulací, dokud jsou stále v paměti.
- **Modul ukládání dat** zapisuje výsledná data společně s jejich statistickým vyhodnocením do výstupních souborů.

Tyto moduly na sebe v průběhu zpracování simulací navazují, komunikují mezi sebou a využívají data uložená v operační paměti. Na obrázku č. 2 je navrženo zapojení jednotlivých modulů v rámci toku dat.



Obrázek č. 2 - Diagram toku dat – Metodika PlaGroSim

5.4 Ověření metodiky pomocí experimentální implementace

Experimentální implementace byla provedena během výzkumné činnosti katedry ve spolupráci se zahraničními partnery v rámci širší vědecké komunity zabývající se simulacemi zemědělské produkce. K vlastnímu zpracování bylo využito hardwarové vybavení specializovaných počítačových laboratoří PEF.

V první fázi výzkumu byly zpracovávány simulace produkce čiroku (Sorghum bicolor) na území Indie. Byl použit postup, který rámcově odpovídá výchozí modelové situaci z kapitoly 5.2. Metodika PlaGroSim byla pro tento výzkum použita pouze v dílčích bodech.

Na základě výzkumného požadavku ze strany zahraničních partnerů z instituce ICRISAT India (International Crops Research Institute for the Semi-Arid Tropics) bylo vybráno 311 lokalit na území Indie. Jednalo se o územní bloky o velikosti 50x50 kilometrů. Specifikace parametrizace GxExM zahrnovala informace o typu půdy a jejích hydrologických vlastnostech, datumu setí, hustotě setí, použití hnojiv, zavlažování, a specifických vlastnostech rostliny čiroku ovlivňující rychlost a intenzitu růstu. Kombinací těchto parametrů vzniklo na každé lokalitě 13 824 simulací.

Pro zpracování byl použit software APSIM. Pro převod datových formátů, generování simulací a dávkování byly vyvinuty dílčí programy v programovacím jazyce C#. Filtrace a statistické zpracování výsledných dat byla zajištěna přímo v Microsoft Excel pomocí makra v jazyce VBA (Visual Basic for Applications). Pro zjednodušení procesu zpracování byl spojen krok generování simulací s dávkovým spouštěním v rámci jednoho programu.

Zpracování simulací trvalo 14 dní na osmi strojích a bylo vygenerováno 14,5 TB výstupních dat. Byl použit následují hardware:

1x výkonná pracovní stanice „Typ 1“:

- dva procesory AMD EPYC 7601 32 jader / 64 vláken na CPU, 2.20/3.20 GHz
- paměť RAM 1,5 TB
- HDD SSD 2000 GB

7x výkonná pracovní stanice „Typ 2“:

- dva procesory AMD EPYC 7281 16 jader / 32 vláken na CPU, 2.10/2.70 GHz
- paměť RAM 126 GB
- HDD SSD 1000 GB

Na základě filtrace dat a jejich statistického zpracování byl pro každou územní jednotku vypočítán ideální postup pěstování pro maximalizaci zemědělské produkce. Tyto výsledky byly následně vizualizovány jako heatmapy a slouží zemědělcům v Indii jako podpora rozhodování.

Navazující výzkum se týkal pěstování podzemnice olejné (Arachis hypogaea), též na území Indie. Byly přidány některé nové parametry (například reakce rostlin na deficit tlaku par) a došlo ke zvýšení počtu možností stávajících parametrů. Zároveň byl pro tento projekt vybrán větší počet lokalit. Celkově tedy došlo k výraznému zvýšení rozsahu výzkumu.

Pro zpracování těchto simulací byla nasazena navržená metodika PlaGroSim v plné míře. Byl vytvořen centrální modulární softwarový nástroj, taktéž pojmenovaný PlaGroSim (Plant Growth Simulator). Tento nástroj v sobě integroval jednotlivé dílčí prvky zpracování v rámci dedikovaných modulů programu. Bylo tak možné řídit celý proces zpracování z jednoho místa a došlo k výraznému snížení nutnosti uživatelských činností (inicializace a spouštění jednotlivých kroků). Zároveň díky záměně softwaru APSIM za štíhlejší SSM došlo ke zrychlení výpočtu simulací. Bylo zpracováno 95 040 simulací pro každou z 1 173 lokalit. Pro výpočet byl použit pouze jeden počítač (výkonná pracovní stanice „Typ 2“) a během dvou dní bylo vygenerováno 1,7 TB výstupních dat.

Porovnání obou těchto výzkumných projektů z hlediska efektivity využití výpočetní kapacity je zobrazeno v tabulce č. 1.

	APSIM	PlaGroSim
počet lokalit	311	1173
počet simulací na lokalitu	13824	95040
počet strojů	8	1
doba výpočtu (dny)	14	2
objem dat (TB)	14,5	1,7
průměrná délka výpočtu na jednu simulaci (vteřiny)	2,25	0,002

Tabulka č. 1 - Parametry zpracování APSIM vs. PlaGroSim

Dílčí části použitého postupu zpracování simulací, a diskuse významu vypočítaných výsledků pěstování čiroku v Indii jsou zachyceny v publikaci č. 4. Jedná se o stěžejní publikační výstup ve velmi kvalitním vědeckém časopise Field Crop Reseach. Tento časopis má vysoký impakt faktor (7.234) a je v prvním decilu pro oblast agronomie.

Publikace č. 4

typ:	vědecký článek
název:	An APSIM-powered framework for post-rainy sorghum-system design in India
autoři:	Ronanki, S., **PAVLÍK, J.**, Masner, J., Jarolímek, J., Stočes, M., Subhash, D., Talwar, H., Tonapi, V., Srikanth, M., Baddam, R., Kholová, J.
rok:	2022
vydáno v:	Field Crops Research
indexováno:	Web of Science, impakt faktor 7.234 Scopus, CiteScore 10.5
odkaz:	https://doi.org/10.1016/j.fcr.2021.108422

V současné době jsou vyhodnocovány dosažené výsledky simulací získané v druhé části výzkumu, ve kterém již byla metodika PlaGroSim plně implementována pro zpracování dat, a je připravována navazující publikace. Vzniká také dodatečný modul pro vzdálenou správu a alerting, který bude umožňovat lépe kontrolovat běh výpočtu. Je plánováno stávající moduly PlaGroSim udržovat a nadále vylepšovat. Konkrétně se bude jednat např. o rozšíření modulu SSM o dodatečné funkcionality pro zpřesnění simulací a zahrnutí rostlin z jiných čeledí, než jsou leguminózy.

5.5 Usability of IoT and Open Data Repositories for Analyzing Water Pollution (A Case Study in the Czech Republic)

Pavlík, J., Hrnčírová, M., Stočes, M., Masner, J., Vaněk, J. (2020) „Usability of IoT and Open Data Repositories for Analyzing Water Pollution. A Case Study in the Czech Republic“ *ISPRS International Journal of Geo-Information*, sv. 9, č. 10. ISSN: 2220-9964. Dostupné na: https://doi.org/10.3390/ijgi9100591

International Journal of Geo-Information

MDPI

Review

Usability of IoT and Open Data Repositories for Analyzing Water Pollution. A Case Study in the Czech Republic

Jan Pavlík [1,*], Markéta Hrnčírová [2], Michal Stočes [1], Jan Masner [1] and Jiří Vaněk [1]

[1] Department of Information Technology, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague, Czech Republic; stoces@pef.czu.cz (M.S.); masner@pef.czu.cz (J.M.); vanek@pef.czu.cz (J.V.)

[2] A.R.C. spol. s r.o., Klimentská 8, Nové Město, 110 00 Prague, Czech Republic; marketa.hrncirova@arcnet.cz

* Correspondence: pavlikjan@pef.czu.cz

Received: 30 August 2020; Accepted: 30 September 2020; Published: 8 October 2020

Abstract: Recently, the process of data opening has intensified, especially thanks to the involvement of many institutions that have not yet shared their data. Some entities provided data to the public long before the trend of open data was pushed to a wider level, but many institutions have only engaged in this process recently thanks to a systemic state-level effort to make data repositories available to the public. Therefore, there are many new potential sources of data available for research, including the area of water management. This article analyses the current state of available data in the Czech Republic—their content, structure, format, availability, costs and other indicators that affect the usability of these data for independent researchers in the area of water management. The case study was conducted to ascertain the levels of accessibility and usability of data in open data repositories and the possibilities of obtaining data from IoT (Internet of Things) devices such as networked sensors where required data is either not available from existing sources, too costly, or otherwise unsuitable for the research. The goal of the underlying research was to assess the impact/ratio of various watershed factors based on monitored indicators of water pollution in a model watershed. Such information would help propose measures for reducing the volume of pollution resulting in increased security in terms of available drinking water for the capital city Prague.

Keywords: GIS; open data; IoT; diffuse water pollution; spatial data; watershed monitoring; factor analysis

1. Introduction

To conduct an effective research activity, access to suitable data is necessary. There are many datasets related to the field of water management. It can be datasets for stream, rivers, and larger water bodies themselves, data regarding weather and rainfall, or soil in the riverbed. As stated by [1], even data regarding the use of fertilizers on arable land in the areas surrounding water courses is essential for water pollution research. Therefore, the range of possible data may come from a wide variety of providers, some seemingly unrelated to water management.

The process of working with spatial data can be divided into four major steps [2]: discovery, acquisition, management and analysis. Firstly, there is the discovery process, where scientists have to learn of the existence of a possible data source and locate it—whether it be on the Internet or by getting in touch with the organization, which provides said data. In this step, sources that aggregate open data are especially useful. One of such sources for European Union member countries are national geoportals tied to the INSPIRE Directive (Infrastructure for Spatial Information in the European

Community) [3]. Another option is a nationwide catalogue of open data. These alternatives often coexists, such is also the case of the Czech Republic, where both the INSPIRE Geoportal and National Catalogue of Open Data [4] serve as primary options for data discovery.

Secondly, there is a process of data acquisition. On an intellectual property level, some data can be provided freely for anyone; sometimes there might be a registration process or license agreement requirement and a payment process in case the data is monetarily charged. The licensing requirements are generally in place to prevent sharing to third parties but can be also used in case it involves sensitive data. In some cases, datasets might be shared incompletely as a safety precaution. This means datapoints for certain locations or regions will be missing, especially when it comes to military areas, or areas that include critical country infrastructure such as nuclear powerplants, oil pipelines, etc. Then there is the actual data transfer process itself, from a technical perspective. Some data may be directly downloadable through a simple web browser; some may require advanced techniques of spatial data transferring. There is also the possibility of obtaining the data in person, in case the online publication availability is lacking for a particular dataset.

The data management step includes data storage, periodical updates if necessary and if the data was not provided in a format that is directly workable, adjustments have to be made to transform data into the format that is required. These transformations are often done utilizing tools available as part of GIS (geographical information system) software, but in cases of some niche data format, they may require dedicated software. For data where the spatial dimension is missing completely, the solution is generally a manual input of the data inside the GIS software. The last step of data analysis/processing includes the actual work with the data itself for the purposes it was obtained.

Spatial data can come in various formats. The type of format of data has important implications on both the provider of data and the final user. Different data formats utilize different types of data storage and transport. On the end user side, the variety of possible data format greatly influences the resulting usability of the data, since every user has access to different software and tools to process the data and many software solutions are the focus of a select few data formats. These interoperability issues are one of the common grievances during actual independent research, because they introduce an often-unnecessary overhead by forcing users to utilize a secondary solution to convert data into a format they can work with.

Many of the commonly used data formats are based on XML (extensible markup language). The standards and specifications for many of the various data formats are handled by OGC (Open Geospatial Consortium) which is the main authority when dealing with spatial data. The most commonly used vector data formats include [5]: GML (geography markup language), KML (keyhole markup language), GeoJSON (JavaScript object notation), ESRI Shapefile, a proprietary format of the ESRI (Environmental Systems Research Institute) company, and GPX (Global Positioning System Exchange Format). The most used raster formats are JPEG (joint photographic experts group), TIFF and GeoTIFF (tagged image file format).

When it comes to the data acquisition, there are several common approaches. One of the simplest solutions would be to provide the data files directly for download in their original form, using the common FTP (file transfer protocol). Another option to share geospatial data are the following three web services: WMS (Web Map Service), WFS (Web Feature Service) and WCS (Web Coverage Service). These services have their standards maintained by the OGC [6] and are essential in providing access and visualization to geospatial data [7]. WMS utilizes a HTTP (hypertext transfer protocol) for the exchange of geographical data in raster format. A user sends a request for data and the WMS responds with an image, which is usually a JPEG or PNG (Portable Network Graphics). One of the most important features of this service is the ease of use. It can be used in both online solutions with various geoportals but also as a source of data for end users. Most desktop software GIS can establish a WMS connection to databases using a simple link copied from a website of the data provider and import an entire layer of spatial data into the map composition. However, this service has a crucial disadvantage for researchers, which is the fact that WMS does not provide the user with the actual data behind

the image. Any user can, therefore, display the map layer as whole or not display it at all. It is not possible to process and analyze the data. Queries are not possible and the option to further process and analyze the data is almost non-existent. The WMS is, therefore, more suited for the public and not for experts in research. The WFS works on same general principles, but instead of sending a raw raster image to the data recipient it sends the actual data instead, usually in GML format [8]. The WCS is the most complex of the three services. It allows for advanced querying and can, therefore, send only a portion of the data. Unlike the WFS which only sends raw GML data, the WCS enhances the data with metadata, descriptions, and semantics to improve client-side rendering and processing of the data. Out of these three main services, WMS is the most widespread, but it is the least usable one for research and spatial planning purposes [9].

The main sources of datasets for water research purposes are usually governmental organizations since private companies do not tend to freely share their data. Depending on the level of sophistication of a given country's information technology strategy and policy, the data discovery process itself can be quite tedious, as pointed out by [10]. If a country is missing strong central authority to establish and enforce a unified open data sharing policy, the heterogeneity of approaches employed by various institutions is likely going to foster an open data landscape that is very hard to navigate by end users. An alternative to using existing open data sources is utilizing solutions based on networked internet of things (IoT) devices with appropriate sensory equipment. Research endeavors regarding water management entail measurements of precipitation, temperature, and flow rate. As shown by [11], several devices can be set up within a river basin to capture not only the final outputs at a confluence, but also the entire development of the water stream.

Our case study was conducted in Vysočina region (south-east from the Czech capital city Prague). We have focused on Bořetický stream watershed with several smaller streams leading into it. It is a small part of watershed that supplies the drinking water reservoir Švihov on Želivka river, which is the main source of drinking water for the capital city Prague. We monitored water quality levels and conducted a statistical factor analysis to determine which watershed factors contribute to monitored pollution indicators. However, in this paper, we would like to focus mainly on our usage of sensory data and open data repositories, especially what type of data were required, where we obtained them, how and if we had to transform these data. Therefore, this paper provides only a concise description of how we calculated watershed factor values that served as an input for statistical analyses to evaluate the impact on water quality indicators. This hands-on experience with a model research project will help us draw relevant conclusions about general levels of accessibility and usability of open data in the field of water management in the Czech Republic.

2. Materials and Methods

The overall process of analyzing diffuse water pollution starts by data collection and preparation. More specifically, it is necessary to gather baseline data. For in situ measurements and observations, IoT devices can be utilized. Repositories of open data, such as hydrological maps, elevation models, soil maps can be accessed if available for a given region. Based on the available data, with regards to the data format and software compatibility, a GIS software is selected to conduct the spatial data processing. The outputs are then further analyzed using statistical modelling (multi-factor regression). In this section we would like to outline this overall diffuse pollution research in detail to establish and justify why we required certain datasets. The process of data discovery and acquisition as well as review of the available open data repositories in the Czech Republic is covered further in the Results section.

2.1. Model Watershed Delineation

For the purposes of our case study, a micro-catchment area with a network of monitoring profiles was selected to easily determine the direct pollution input of a populated area (only few small point sources). A GIS software application was utilized to delineate the catchment area boundaries for each profile. We utilized two main approaches: for bigger streams (I–IV in the Strahler scale), where the

sampling point was near a confluence, we were able to use publicly available hydrological watershed boundaries from DIBAVOD (Digital Base of Water Management Data) [12]. For streams that are smaller or in cases where the sampling location was too far from a confluence, we had to delineate the boundaries manually (according to contour lines). During our research we worked with two GIS applications: ArcGIS, which is a very prominent GIS software solution; and BNHelp, a GIS application that utilizes HSLayers, a modified version of OpenLayers library. BNHelp solution also provides a web-based client for remote work.

Another possible solution for watershed delineation would be to utilize a DEM (digital elevation model) as most GIS applications have the necessary tools (either in the base version or as an add-on toolbox) to delineate a custom watershed area from a DEM for a given catchment point. In the Czech Republic, the best freely available DEM is part of the ArcČR 500 map set [13]. However, this DEM was constructed from a 50 m contour line basemap; therefore, the size of one pixel is 50 × 50 m, which is not detailed enough for small basins. Compared to manual delineation, using this raw DEM produced a higher level of inaccuracies. For better results, a higher resolution DEM would be required. There are two possible models available, the 4th and 5th generation DEMs, both from the Czech Land Surveying and Cadastre Department [14], but both are costly. Even for the small micro-catchment area that was selected for our research, the price would be at least EUR 400 for the 4th-generation model and EUR 1250 for the 5th-generation model.

An example of the manual approach is shown in Figure 1–the delineation of watershed boundaries/areas in the BNHelp GIS application for two of the selected profiles on Útěchovický stream, which is too small to be included in the DIBAVOD watershed boundaries dataset. Profile BP2 boundaries are displayed by a thick red line, the profile BP5 area is displayed with orange fill color.

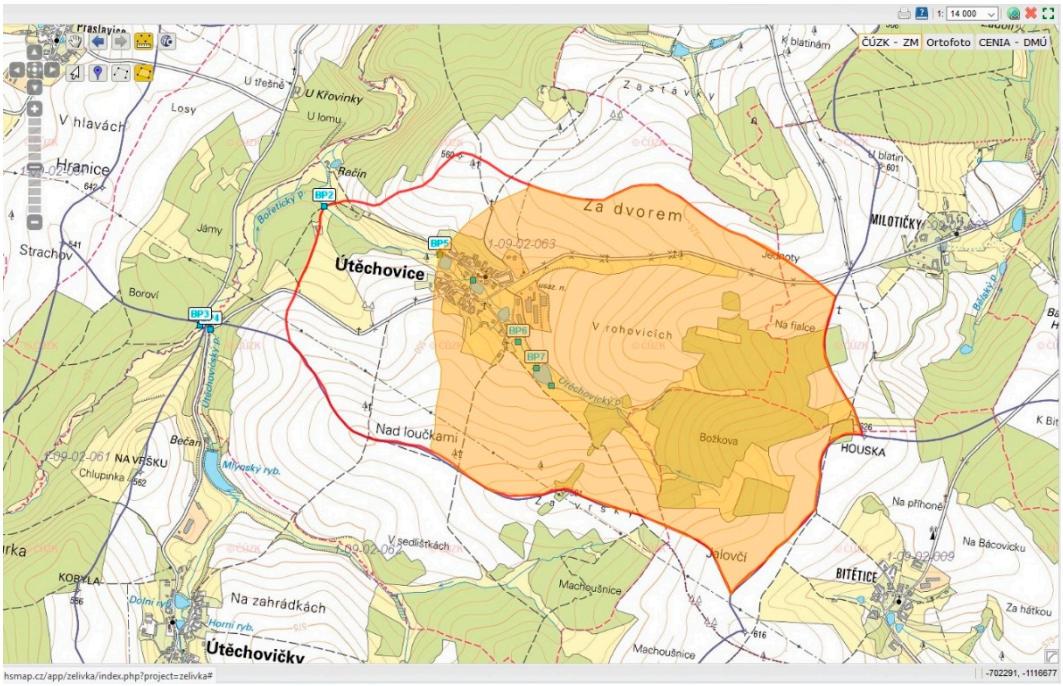


Figure 1. Boundaries of BP2 and BP5 profiles made in BNHelp geographical information system (GIS) application.

2.2. Water Monitoring

Several key pollution indicators were selected, such as ammoniacal nitrogen, nitrate nitrogen, molybdate reactive phosphorus and COD_{Cr} (chemical oxygen demand—potassium dichromate method). This selection was in accordance with the established knowledge about the main organic and nutrient water pollution sources and resulting eutrophication [15–17]. Water samples were taken

monthly on each selected water profile for three years. Samples were analyzed using available test-tube methods within conjunction with photo spectrometry methods according to [18]. Meteorological data such as daily temperatures and sum of daily rainfall, therefore, had to be recalculated into monthly averages to match the water sample periodicity. Water flows were calculated using a hydrological analogy. Direct flow measurements were conducted only during major precipitation-runoff events.

Apart from obtaining these datasets from official sources (Czech Hydrometeorological Institute), there is also a possibility to utilize networked IoT devices. For precipitation, temperature, and water flow there are fully operational devices available. Alternatively, a custom device can be built using a basekit with a conjunction of corresponding sensors. These custom devices can potentially be cheaper than solutions on the market, especially when scaling up the number of units deployed, but it requires expertise and lot of effort for the initial setup and calibration. In case of research that does not focus on the pollution factors but just on measuring the volume of pollution, a turbidity sensor can be used as proposed by [19] to create an all-in-one measuring device.

Considering our model research was very small scale, we did not require data from multiple locations. Therefore, utilizing networked sensors was for our purposes unnecessary. We decided to purchase a single-unit measuring device to obtain the weather data, specifically a wireless meteorological station Conrad Electronics RW 53, WH 5300 from the Conrad company. The cost of this device was approximately 80 EUR.

The other option would be to purchase this data from Czech Hydrometeorological Institute. It provides a selection of data for free (from measuring stations near larger cities and on selected spots), but for data at a custom location a special commission would have to be arranged. Based on data available from a publicly available contract database, we were able to ascertain that both the precipitation data and temperature data would cost at least 50 EUR each for every year.

2.3. Pollution Factors

For each catchment profile, we also compiled a database of anthropogenic factors separated into three thematic groups: human settlement indicators, land-use categories, and agricultural production.

The main human settlement indicators were population density, volume of communal wastewater production and treatment. The number of residents living inside a delimited watershed under each profile was calculated manually. A list of city and village names was compiled from the topographic baseline map for each given profile. We then used data sources available to the public to determine the number of residents for each settlement. The data sources we used were the official population data from Czech Statistical Office as well as their local branch offices in combination with PRVKÚK (Regional development plan of water supply and wastewater treatment). Luckily, both sources provided sufficiently detailed and recent data for the area that our case study dealt with. In many other Czech regions, data about small settlements is not so easily available.

A second group of factors were acreages and percentages of all land-use categories in every model watershed within the water-quality monitoring profile. The open data repository we used for this part was the LPIS (Land Parcel Identification System), which is an information system operated by the Ministry of Agriculture. It includes a database of all registered parcel blocks with agricultural land. There are distinguished categories of arable land, permanent grasslands, other arable land cultures, and pastures. It also contains several other key information datasets for each parcel such as altitude and gradient. The data provided is downloadable free of charge and is stored in .shp layers for every cadaster territory in Czech Republic. Therefore, we first had to create an intersection of our profile boundaries with the cadaster territory map (provided within the ArcČR dataset) inside our GIS software to determine which cadasters we need to download from LPIS. After inserting and merging all the LPIS layers and intersecting it with our profiles, we could calculate the ratio of all land-use categories in our selected areas. For the calculation of the ratio of forests within model areas, we used the same approach, except the source data was the freely available GIS layer CORINE (Coordination of Information on the Environment) Land Cover 2000. To calculate the ratio of built-over areas and

water surface areas, we had to delineate them manually from a topographic map. For several larger settlements, the data about built-over area was available within the Czech Statistical Office datasets (land-use categories acreage in all cadasters—data in tables are downloadable for free in .xls format).

Lastly, we conducted a monitoring of agricultural production in the area. During each agricultural production cycle, we determined which crops were being grown on arable land parcels within a model watershed on major agricultural holdings, because changes in crop variability between years can influence the resulting nutrient pollution considerably, especially for small water streams.

We analyzed all acquired data to evaluate impact of various natural and anthropogenic factors within the model watershed (population, land-use categories, agricultural production, and meteorological data) on selected indicators of water quality on monitoring profiles. Concrete results of this analysis are going to be published in a separate follow-up paper, which will be focused more on the actual values of measured data, their statistical analysis, and implications resulting from our observations/measurements.

3. Results

In the Czech Republic, most open data come from sources within the administrative bodies of the government such as ministries, research institutes under the purview of ministries, or organizations tied to various governmental institutions. The large number of data sources and the resulting fragmentation of open data in Czech Republic has a historical context; even as early as the mid-2000s there were reports [20] about the decentralized character of data providers that stressed the necessity for a more generalized catalogue of open data. It was asserted by [21] that this system has its advantages in terms of maintaining validity of the data through regular updates, since each provider is only responsible for the management of their own small section of the overall open data landscape. However, this comes at a cost of lower discoverability of data sources and forces researchers to know which data sources exist and where to look for them. This introduces an overhead to every research endeavor, because prior to research, subjects must conduct an exhaustive discovery process to locate all possible data sources. For this purpose, several user-driven open data source compilation web sites have emerged that are trying to keep an up to date lists of data locations.

The official aggregation site for open data in the Czech Republic is the National Catalogue of Open Data. It has been established only recently (during 2015) and most Czech institutions did not adopt publishing their open data in this catalogue immediately. The amount of available data is, therefore, still increasing as more organizations take part in the effort of opening data, so the situation is improving. However, there are several issues that are reducing the usability of this catalogue. Firstly, it includes both spatial and non-spatial data without the option to filter one type or the other. It includes filter by file format, so data in KML or GeoJSON can be assumed to have a spatial nature. But many spatial data files, especially those originally in .shp file format, are provided as .ZIP, making it impossible to distinguish them from non-spatial .ZIP data. However, as pointed out by [22], non-spatial data can be enhanced with spatial features, so all data within the catalogue is potentially valuable and cannot be discarded outright during the data discovery process based on the data format alone. Therefore, the fact that raw machine-readable data are published in the same manner as data files that deal with transparent administration such as public procurements, contracts, retainers, and incentives are cluttering the system and reducing the discoverability of pertinent datasets.

A second major issue with the catalogue is caused by the overall heterogeneity of the datasets as well as approaches taken by individual data providers. Some institutions share country-wide data, some datasets are separated by regions, and some are split into tens of thousands individual datasets based on municipalities (each village has its own dataset). This means that a dataset including information pertinent to the whole country can be potentially buried in the middle of thousands of other files that are focused on a small location.

Another prominent data source is the INSPIRE Geoportal since it aggregates data from many organizations within Czech Republic. It is based on the 2007/2/ES INSPIRE European directive [3],

which aims to establish a European spatial data infrastructure (SDI), that will provide high-quality standardized data to support strategic environmental policies. Most standards that are used within INSPIRE fall into the ISO (International Organization for Standardization) 19,100 series, maintained by corresponding ISO Technical Committee [23] The Czech INSPIRE is managed by CENIA (Czech Environmental Information Agency) which is part of the Ministry of Environment. In its current state, the geoportal has its own web map application for displaying maps and layers as well as WMS that provides the same data. Currently available datasets include the orthophoto map (current and historical), digital terrain model, military maps, topographic map, real estate, land cadaster map and ZABAGED (Fundamental Base of Geographic Data). The datasets are enhanced by rich metadata, which allow for filtering using a lot of parameters, such as source, time-frame, topic, location, type of service and others. The availability of metadata plays a key role during the data discovery process, which further distinguishes INSPIRE from other data sources and emphasizes its significance in the Czech open data environment.

ZABAGED is a baseline set of maps provided by the State Administration of Land Surveying and Cadaster. Among others, it includes maps regarding territorial units, buildings and roads, vegetation and land surface, terrain and geodetic points, and distribution networks and pipelines. The department runs its own geoportal with the ZABAGED maps in it, but they are also provided inside the INSPIRE Geoportal as well as in WMS and WFS forms.

LPIS is a land evidence geographical information system operated by the Ministry of Agriculture and is mainly used by farmers to input data about fertilizers, pesticides and other preparations they use for agricultural production (in order to be eligible for state-level or European subsidies). The LPIS is one the most advanced systems in terms of data sharing since it has many options how to provide data. Apart from having its own web GIS interface, the data are also available in both WMS and WFS forms as well as a direct download in .shp files. LPIS is linked to Common Agricultural Policy (CAP) payments and is, therefore, implemented in every European Union (EU) country. Apart from the required functionalities [24], each member state has full control over their LPIS, resulting in differences between countries. The Czech LPIS has extensive public access, allowing users to view and download a large portion of the included data.

DIBAVOD is a set of maps compiled by the T.G.M. Water Research Institute. This set of maps contains mostly data about watercourses and watershed areas and serves as baseline for water related layers in ZABAGED. The whole DIBAVOD set or its parts can be downloaded on the Institute's website in .shp format.

VÚMOP (Research Institute for Soil and Water Conservation) provides a set of maps regarding soil structure, density, erosion vulnerability, water capacity and others. Until recently, users had to formally request the data and received them on a CD/DVD disk, but the institute is currently moving towards adopting the WMS method of sharing for more of its datasets.

The Czech Hydrometeorological Institute provides data regarding weather, including rainfall. Data are available for download after filling in a short form and a license agreement. Most of the data are in .csv of .xls formats and it is, therefore, necessary to process the data to convert them into a suitable geospatial format. Among all the available data considered for research in the water management sector, the hydrometeorological dataset are the most suitable to be obtained utilizing IoT devices. Sensors that measure temperature, humidity and precipitation are available on the market. Implementing a custom-built network of devices is very likely to provide more precise data than the institute due to the localized nature of IoT deployment (sensors directly where they are needed instead of having to use data from closest meteo-station operated by the institute, which can be several kilometers away).

The Czech Statistical Office collects primarily data regarding trade, economic and industrial activity, demographics, and wages and so on. Many of those datasets are not directly linked to water management, but some are quite important (for instance, population density). However, the Czech Statistical Office does provide such data in a spatial form for fee. Free data are only provided in .xls

files. Additionally, a lot of statistical data are mainly summarized and presented to public in relation to local authorities, cadasters, as regional statistics etc., therefore necessitating double manual data conversion: one conversion from tables to spatial data and secondly an integration of data over several smaller cadasters, counties, or regions.

Apart from what datasets are available and the form they take, another important aspect, especially when it comes to small-scale independent research, are the associated costs. The access to INSPIRE in the Czech Republic is without charge. Any user may display available map layers in the online Geo Portal or use WMS to display raster images in their personal GIS software. Additionally, users may register on the website which gives them access to more features, such as creating and saving their own map compositions [25]. However, the actual raw data is not available. As part of the INSPIRE directives, such data are freely available to other public institutions that request it, but not to the public. This is a common theme even with the other data sources. The general rule is that governmental institutions share the data between themselves but require a fee when providing them to outside recipients. This is also the case with ZABAGED and other datasets provided by the Land Surveying and Cadaster Department, which only allows downloading their data after a payment (depending on the dataset usually payments per square kilometer or per 5×5 km square). Similarly, VÚMOP has most of their datasets locked behind a paywall as well. Data from Czech Hydrometeorological Institute have an associated fee as well, but the Institute also provides licensing options with discounts for researchers and students. There are several exceptions to the general payment schemes: LPIS allows free unlimited downloading of their source data in .shp format through a simple web interface and the DIBAVOD maps are available for download free of charge as well.

For certain datasets (in our case temperature, precipitation, and water flow) there is an alternative to the existing open data repositories and that is utilizing IoT devices for measurement. A network of sensors has the inherent advantage of low upkeep costs, higher data accuracy thanks to the localized nature, as well as customizable data velocity. These advantages are offset by higher initial costs and the required work and expertise needed to set up and calibrate the network. In the Czech Republic there are several network providers who offer services specialized to the IoT area. Sigfox is provided by SimpleCell Networks corp. LoRaWAN (long-range wide area network) is provided by the Semtech corp. in cooperation with Czech Radiocommunications. Third major technology available at the Czech market is NB-IoT (narrow-band IoT), based on the LPWA (low-power wide area), provided by mobile operator Vodafone. A common problem across these providers is that they mostly focus on implementations inside buildings, or when it comes to outdoor applications, the emphasis is on municipal areas. It is possible to find ready solutions for building monitoring, company property monitoring (for instance car park administration), factory production optimization, or smart households. Research conducted in the countryside would probably require negotiating a custom solution, especially with the regards to the potential unavailability of network coverage in the rural areas.

4. Discussion

The current situation of data accessibility for research purposes is undesirable in many aspects. The most used approach for data sharing is WMS, which is by its nature unsuitable for use in further research. Some data are available in WFS format or directly for download, but for the most part the cost of such data is too high for a typical small-scale research budget. This gives an advantage to larger, well-funded research subjects or those linked to various ministries or departments who can obtain the data for free as part of co-operative projects, state funded studies etc.

One of the major issues is also the decentralized nature of open data providers, where each institution creates datasets according to their individual know-how. There is a significant lack of aggregation of available open data, especially in the agricultural sector [26]. There is a continuous effort for technological improvements for the available datasets as part of the INSPIRE initiative—such as incorporating linked data and increasing machine readability [27]. Therefore, as the technological

and quality requirements increase on the front end (INSPIRE Geoportal), the individual back-end provider institutions are expected to eventually adopt the new standards as well, hopefully resulting in better homogeneity of available open data in the near future.

The Czech Republic is currently in a good position as far the actual data content is concerned. Most of the datasets that are important for research in water management do exist or are being actively developed. There are possible improvements in terms of non-spatial data being reworked in proper map layers on a provider level, thus reducing costs for the end users (by eliminating the need for research teams to hire a GIS specialist to manually process the data into usable formats). The availability of open databases with data in suitable formats is also a key prerequisite for development of decision and information support systems in the agricultural area.

New datasets can also be created on a custom basis, utilizing IoT devices. High volumes of data coming from a wide range of networked sensors promise an overall improvement in data coverage, periodicity, and level of detail. But there are also limitations in the overall scale by which IoT can be deployed as well as the initial costs. Other developments could come from the field of remote sensing, which might be able to provide new avenues for water-quality monitoring [28]. However, it is uncertain whether the current institutional capacities, both in hardware and manpower, are going to suffice to process such volumes of data effectively in order to provide them to the general public for reasonable costs or free of charge.

Author Contributions: Conceptualization, Jan Pavlík and Markéta Hrnčírová; Methodology, Jiří Vaněk; Software, Michal Stočes and Jan Masner; Investigation, Jan Pavlík; Formal Analysis, Markéta Hrnčírová; Resources, Michal Stočes; Data Curation, Jan Pavlík; Writing-Original Draft Preparation, Jan Pavlík; Writing-Review and Editing, Markéta Hrnčírová and Jan Masner; Visualization, Michal Stočes; Supervision, Jiří Vaněk; Project Administration, Jiří Vaněk; Funding Acquisition, Jiří Vaněk. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant number 2019B0009: "Life Sciences 4.0".

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Manuel, J. Nutrient pollution: A persistent threat to waterways. *Environ. Health Perspect.* **2014**, *122*, A304–A309. [CrossRef] [PubMed]
2. Acker, J.G.; Leptoukh, G. Online analysis enhances use of NASA Earth Science Data. *Eos Trans. Am. Geophys. Union* **2007**, *88*, 14. [CrossRef]
3. Directive 2007/2/EC of the European Parliament. Available online: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002 (accessed on 30 September 2020).
4. Czech Republic open data portal – National catalogue of open data. Available online: https://data.gov.cz (accessed on 2 October 2020).
5. Peng, Z.-R.; Zhang, C. The roles of Geography Markup Language (GML), Scalable Vector Graphics (SVG), and Web Feature Service (WFS) specifications in the development of Internet Geographic Information Systems (GIS). *J. Geogr. Syst.* **2004**, *6*, 95–116. [CrossRef]
6. OGC Standards and Resources. Available online: https://www.ogc.org/standards (accessed on 30 September 2020).
7. Wei, Y.; Santhana-Vannan, S.-K.; Cook, R.B. Discover, visualize, and deliver geospatial data through OGC standards-based WebGIS system. In Proceedings of the 17th International Conference on Geoinformatics, Fairfax, VA, USA, 12–14 August 2009.
8. Jones, J.; Kuhn, W.; Keßler, C.; Scheider, S. Making the web of data available via web feature services. *Lect. Notes Geoinf. Cartogr.* **2014**, 341–361. [CrossRef]

9. Cada, V.; Cerba, O.; Fiala, R.; Janecka, K.; Jedlicka, K.; Jezek, J.; Mildorf, T. Spatial planning—The open field for data description and web services. In Proceedings of the International Society for Photogrammetry and Remote Sensing Archive, 38; 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services, Como, Italy, 26–27 August 2010.
10. Attard, J.; Orlandi, F.; Scerri, S.; Auer, S. A systematic review of open government data initiatives. *Gov. Inf. Q.* **2015**, *32*, 399–418. [CrossRef]
11. Douinot, A.; Torre, A.D.; Martin, J.; Iffly, J.-F.; Rapin, L.; Meisch, C.; Bastian, C.; Pfister, L. Prototype of a LPWA network for real-time hydro-meteorological monitoring and flood nowcasting. *Lect. Notes Comput. Sci.* **2019**, *11803*, 566–574.
12. V.Ú.V. T.G.Masaryk—GIS Section—About DIBAVOD. Available online: https://www.dibavod.cz (accessed on 2 October 2020).
13. Arc ČR—Geographical Information Systems (GIS)—ARCDATA Prague. Available online: https://www.arcdata.cz/produkty/geograficka-data/arccr-500 (accessed on 2 October 2020).
14. ČÚZK—Introduction. Available online: https://www.cuzk.cz (accessed on 2 October 2020).
15. Carpenter, S.R.; Caraco, N.F.; Correll, D.L.; Howarth, R.W.; Sharpley, A.N.; Smith, V.H. Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Appl.* **1998**, *8*, 559–568. [CrossRef]
16. Elser, J.J.; Bracken, M.E.S.; Cleland, E.E.; Gruner, D.S.; Harpole, W.S.; Hillebrand, H.; Ngai, J.T.; Seabloom, E.W.; Shurin, J.B.; Smith, J.E. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol. Lett.* **2007**, *10*, 1135–1142. [CrossRef] [PubMed]
17. Fučík, P.; Hejduk, T.; Peterková, J. Quantifying water pollution sources in a small tile-drained agricultural watershed. *CLEAN Soil Air Water* **2014**, *43*, 698–709. [CrossRef]
18. Horakova, M. *and collective Water Analytics*; University of Chemistry and Technology: Prague, Czech Republic, 2003.
19. Daigavane, V.V.; Gaikwad, M.A. Water quality monitoring system based on IOT. *Adv. Wirel. Mob. Commun.* **2017**, *10*, 1107–1116.
20. Kubicek, P.; Horakova, B.; Horak, J. The geoinformation infrastructure in the Czech Republic. The key role of metadata. In Proceedings of the 11th EC GI & GIS WORKSHOP, ESDI: Setting the Framework, Alghero, Sardinia, 29 June–1 July 2005; pp. 82–84.
21. Řezník, T. Geographic information in the age of the INSPIRE Directive: Discovery, download and use for geographical research. *Geography* **2013**, *118*, 77–93. [CrossRef]
22. Dessì, N.; Garau, G.; Recupero, D.R.; Pes, B. Increasing open government data transparency with spatial dimension. In Proceedings of the 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, France, 13–15 June 2016; pp. 247–249.
23. ISO/TC 211. Geographic Information/Geomatics. Available online: https://www.iso.org/committee/54904/x/catalogue/ (accessed on 30 September 2020).
24. Rathonyi, G.; Varallyai, L.; Herdon, M. Best practices of GIS applications in the Hungarian agriculture. *AGRIS On-line Pap. Econ. Inform.* **2010**, *2*, 55–62.
25. Prášek, J.; Valta, J.; Hřebíček, J. National INSPIRE geoportal of the Czech Republic. *IFIP Adv. Inf. Commun. Technol.* **2013**, *413*, 425–438.
26. Vostrovsky, V.; Tyrychtr, J.; Halbich, C. Correctness of open data in the agricultural sector. In *Agrarian Perspectives XXIV*; CULS: Prague, Czech Republic, 2015; pp. 528–535.
27. Schade, S.; Lutz, M. Opportunities and challenges for using linked data in inspire. In Proceedings of the Workshop On Linked Spatiotemporal Data, Zurich, Switzerland, 14–17 September 2010.
28. Schaeffer, B.A.; Schaeffer, K.G.; Keith, D.; Lunetta, R.S.; Conmy, R.; Gould, R.W. Barriers to adopting satellite remote sensing for water quality management. *Int. J. Remote Sens.* **2013**, *34*, 7534–7544. [CrossRef]

5.6 Data Pre-processing for Agricultural Simulations

Jarolímek, J., **Pavlík, J.**, Kholova, J., Ronanki, S. (2019) „Data Pre-processing for Agricultural Simulations“ *AGRIS on-line Papers in Economics and Informatics*, roč. 11, č. 1, s. 49-53. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2019.110105

Agris on-line Papers in Economics and Informatics

Volume XI Number 1, 2019

Data Pre-processing for Agricultural Simulations

Jan Jarolímek[1], Jan Pavlík[1], Jana Kholova[2], Swarna Ronanki[3]

[1] Department of Information Technologies, Faculty of Economics and Management, University of Life Sciences Prague, Czech Republic

[2] Department of Crops Physiology, International Crops Research Institute for Semi-Arid Tropics - System

[3] Analysis for Climate Smart Agriculture, Hyderabad, IndiaDepartment of Crop Production, ICAR - Indian Institute of Millets Research, Hyderabad, India

Abstract

The process of agricultural simulation using APSIM requires input meteorological data to be prepared in a specific format and the simulation setting file to be ready before the simulation processing starts. Because of possible time savings when conducting large number of simulations at once, it is preferable to create all the input and settings files for all the simulations beforehand and process the simulations in batches as large as possible. This article specifically deals with the data acquisition, transformation and preparation process. It also outlines initial testing and computing time estimations and discusses scheduling, parallel processing and other possible simulation optimization methods..

Keywords

APSIM, big data, data processing, yield optimization, software automation, parallel processing.

Jarolímek, J., Pavlík, J., Kholova, J. and Ronanki, S. (2019) "Data Pre-processing for Agricultural Simulations", *AGRIS on-line Papers in Economics and Informatics*, Vol. 11, No. 1, pp. 49-53. ISSN 1804-1930. DOI 10.7160/aol.2019.110105.

Introduction

With increasing processing capabilities, it is becoming possible to tackle larger research endeavours. In the area of scenario simulations, this increase in hardware power allows for broader assignments in terms of variable combination. Historically, the total amount of simulations was severely limited and required either very narrow specification of simulation parameters or usage of techniques that lowered the processing requirement at the cost of less accurate results, such as downscaling (Hewitson and Crane, 1996). Nowadays, higher hardware power can be utilized to calculate more simulations extending the limits of the usual variable spectrum. However, the multi-linear nature of growth of number of simulations based on number of options for each variable still limits the simulation process in general, so some restrictions need to be upheld regardless.

One example of simulation software that was originally designed for small scale field simulations on a single computer but has seen a resurgence as a large scale (even on a global scale) tool for simulation of agricultural production is the Agricultural Production Systems sIMulator (APSIM). This software provides important insight into challenges regarding food security, climate change adaptation and carbon trading (Holzworth et.al., 2014).

By using supporting software tools for automation and scheduling it is possible to tackle large number of simulations in APSIM by splitting the computation onto multiple machines utilizing parallel processing as shown by (Zhao, et.al., 2013). Even though hardware and software scales differently during processing (Kambadur, et.al., 2013), with a proper setup and data pre-processing it is possible to make up for the increased number of simulations. Apart from increasing the range for variables, increasing the resolution of the grid will also affect the number of simulations required, however as pointed out by (Mass, et. al., 2002) when it comes to weather forecasts, reducing the grid size beyond certain limit no longer significantly improves the quality of results.

Another issue is also the period of input weather data. Due to changes in global climate, only short-term predictions are possible (Aurbacher, et.al.,

2013), so it might be necessary to recalculate simulations on a periodical basis with newest possible data sets, in order to maintain high level of usability of results. This however introduces additional layer of scaling, so in order to ensure up-to-date knowledge based on the simulation results, measures must be taken to reduce the processing requirements of each individual set of simulations (Skoogh, et al., 2010).

The requirements for data storage also scale based on the number of simulations. However, there are possibilities to cut down the storage requirements by extracting required results during the processing from output files that have been already calculated and deleting them. But considering the processing of simulations is the most time-consuming part of the research process, deleting finished output files may be ill-advised, since they are the most "expensive" to create. Therefore, a better solution would be to search for additional storage capacities. Luckily, thanks to the rise of IoT (Internet of Things) as a source of data (Stoces et.al., 2018), most research entities have bolstered their storage hardware in recent years.

Overall, the issue of large-scale simulations, their processing requirements and optimization in general is very current topic. Many researchers are looking for solutions in various areas, whether it be utilizing cloud-based capacities (Szufel, et al., 2017), exploiting existing hardware to its maximum potential (Fujimoto, 2016) or looking for new frameworks altogether (Kirby, et.al., 2018).

Materials and methods

In order to simulate agricultural production two input files are required for APSIM. Firstly, there is the .met file which contains historical meteorological data for a given field / grid square. The required parameters are daily rates of solar radiation (radn), minimum daily temperature (tmin), maximum daily temperature (tmax) and precipitation rate (rain). Apart from these daily values the .met file must also contain pre-calculated values for annual ambient average temperature (tav) and annual amplitude in mean monthly temperature (amp).

The second input file is the .apsim file that contains settings for the simulation (irrigation rates, sowing window, sowing density, fertilization etc.) as well as data related to the given grid square (such as soil properties, characteristics for given plant genotype and so on).

The meteorological data we use are from Goddard Institute for Space Studies (GISS), which is part of National Aeronautics and Space Administration (NASA). The AgMERRA Climate Forcing Datasets (https://data.giss.nasa.gov/impacts/agmipcf/agmerra/) are free to download in an .nc4 format. The datasets are split into files per year (from 1980 to 2010) and per variable. Therefore, some pre-processing will be required to transform the data, since APSIM is expecting the data split into files per grid square containing a table with all the values for all the variables and for all the years.

The .apsim files are just .xml files using the markup language to capture all the input variables for each given simulation. These files have to be prepared based on real agricultural conditions in given area. For the purposes of multi-variable simulation, each single simulation has to be reproduced so that every possible combination of variables was represented. Considering the large-scale nature due to the high number of grid squares as well as high number of variable combinations, it is unfeasible to do this task manually.

To complete the pre-processing, both the .met and .apsim file for all the simulations must be ready. The next task is to optimize the simulation computations themselves outside of the APSIM software. Possible solutions include parallel processing, utilization of cloud based resource structure, optimizations regarding scheduling and use of additional hardware resources during their downtime. We plan to publish a separate follow-up article regarding this process at a later date.

Results and discussion

The required data conversion from .nc4 files downloaded from the NASA into .met files required by the APSIM software was achieved using a MATLAB script. The calculation of (tav) and (amp) variables can be done within MATLAB as well or using R script. However, we found that the easiest way is to first convert to .xls, do the calculation in MS Excel and then convert to .prn file, which has the same required structure as the .met, and simply change the extension.

In order to create the settings for all the simulations we have written a program using C# language that loads a single .apsim file with one simulation in it and returns an .apsim file with all the possible variations of that simulation, with all the combinations of chosen variables. In our case, it was 12600 simulations per each grid square. This batch size proved to be too high for the APSIM software, so we had to adjust the program to create

several smaller files (see below). The choice to use C# was arbitrary based on experience of the programmers in our team. Any other programming language (python, java etc.) can be chosen and will work just as fine to write a similar program / script.

Our simulations used single soil settings for all the grid squares. In cases where different soil settings are required the preprocessing depends on the form and availability of data in a given country. This will provide additional layer of preprocessing, however as shown by (Kim, et.al., 2018), this step can be also automated by writing an application specific to the soil database that will fetch the data in bulk.

Overall, the pre-processing of data did not provide any challenges in terms of software / hardware requirements; even with high number of simulations (hundreds of millions) the computation time is in the range of several minutes. The majority of input in this stage was therefore programmer labour time needed to write the scripts and programs.

The simulation processing itself will be done using the command line version of APSIM. The software has a graphical user interface (GUI) provided (see Figure 1), but it does not include any functionality that would be helpful setting up computation of large number of simulations. It is designed merely as a tool to better visualize the contents of the .apsim files and to edit values when dealing with small number of simulations at a time. The computation time of both variants (command line and GUI) is similar, but the former provides easier options for automation and scheduling using third-party tools.

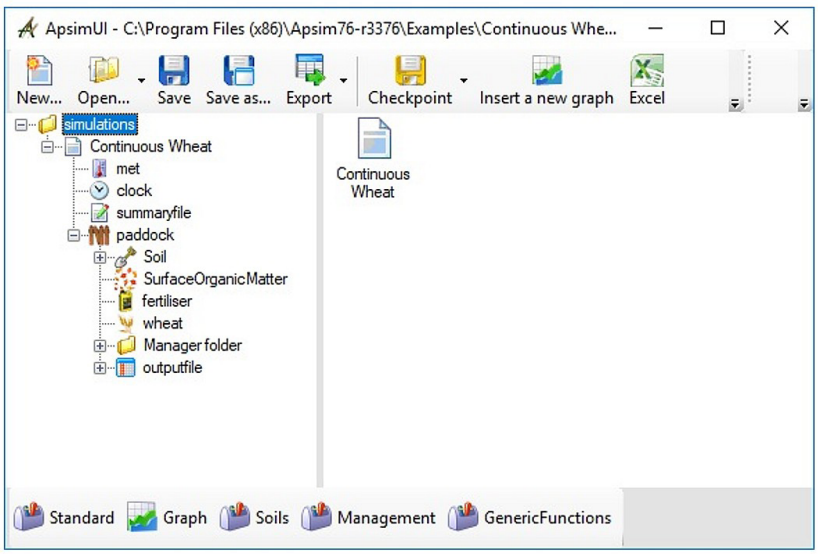
We conducted preliminary testing runs for some of the simulations on several different machines in order to estimate the overall time requirements. What we found was that the processing time doesn't scale perfectly with the amount of simulations in a single batch. Possibly due to some overhead requirements (initialization, clean-up etc.) the efficiency of simulations processed per minute goes up with the batch size (see Table 1 for approximate results).

Number of simulations	Approx. time (minutes)	Simulations per minute
100	2.5	40.0
500	11.5	43.5
1000	22.0	45.5
2000	40.0	50.0
2500	48.0	52.1

Source: own processing

Table 1: Preliminary processing efficiency for different batch sizes.

Based on these results it became clear that in order to optimize the processing, the batch size should be as large as possible. However, the APSIM software cannot handle all the simulations at once. There seem to be a limit on maximum batch size that is influenced by used hardware. Some of the stronger machines we used for testing were able to handle between 2000 and 3000 simulations at once, whereas regular desktop computers



Source: own processing

Figure 1: APSIM User Interface.

with mid-range hardware installed were not able to go over 1000 simulations in a single batch. This limit seems to be influenced by available memory capacity, but strangely during the simulation processing itself, the limiting factor was processor, not memory. This would imply that the memory capacity is mostly relevant during the initialization. We plan to conduct further testing using wider variety of hardware to reach more definite conclusion in this matter.

At this moment, the best way to optimize processing seems to be determining optimal batch size for each machine that will be involved in the computation process and use third-party scheduling software to run the simulations on every machine separately when its resources are free for use. With the way our C# program to generate simulation works at this point, that would mean creating a stockpile of simulations of varying batch sizes for each machine. Due to uneven workload of machines however, this may prove problematic, since each computer will drain its simulation stockpile at different rate. A solution to this issue might be adjusting the simulation generation so that it does not work as a static application, but rather an ongoing server application. That way the schedulers that handle processing could request batches of input files when necessary.

Conclusion

The requirements for data pre-processing when working with APSIM scale with the amount of simulations due to the lack of in-built option for variable simulations. However, this can be handled reasonably efficiently using features of MATLAB for weather data processing combined with self-written scripts to generate simulation files for all possible combinations of variables. There is little to no room for improvement or time saving when handling these necessary tasks. But when utilizing parallel processing it becomes possible to reduce computing time via optimizing the batch size for each individual machine. Having the option to select variable batch size within the simulation generation script therefore proved very advantageous.

But overall, we must conclude that the age of APSIM software really shows, especially with regards to lack of features / packages that could help with large scale research by removing or at least reducing the required pre-processing requirements. This issue is only amplified by the fact that personnel who use APSIM often do not possess enough IT knowledge and training, especially when it is required to operate additional third-party software. Similar findings regarding lack of IT expertise we pointed out by (Reinmuth and Dabbert, 2017) for instance. Some of these issues will be hopefully handled in the APSIM Next Generation as outlined by (Holzworth et. al., 2018).

Acknowledgements

This article was created with the support of the Internal Grant Agency (IGA) of FEM CULS in Prague, no. 2019A0017 „Bulk processing of large volumes of geographical data“.

Corresponding authors
Ing. Jan Pavlík
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague – Suchdol, Czech Republic
Phone: +420 224 382 356, Email: pavlikjan@pef.czu.cz

References

[1] Aurbacher, J., Parker, P. S., Sanchez, G. A. C., Steinbach, J., Reinmuth, E., Ingwersen, J. and Dabbert, S. (2013) “Influence of climate change on short term management of field crops - A modelling approach”, A*gricultural Systems*, Vol. 119, pp. 44-57. ISSN 0308-521X. DOI 10.1016/j.agsy.2013.04.005.

[2] Fujimoto, R. M. (2016) “Research Challenges in Parallel and Distributed Simulation”, *ACM Transactions On Modeling And Computer Simulation*, Vol. 26, No. 4. 10493301. DOI 10.1145/2866577.

[3] Hewitson, B. C. and Crane, R. G. (1996) “Climate downscaling: Techniques and application” *Climate Research*, Vol. 7, pp. 85-95. E-ISSN 1616-1572, ISSN 0936-577X. DOI 10.3354/cr007085.

[4] Holzworth, D., Huth, N. I., de Voil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ... Keating, B. A. (2014) "APSIM - Evolution towards a New Generation of Agricultural Systems Simulation." *Environmental Modelling & Software*, Vol. 62, pp. 327-350. ISSN 1364-8152. DOI 10.1016/j.envsoft.2014.07.009.

[5] Holzworth, D., Huth, N. I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N. I., Zheng, B. and Snow. V. (2018) "APSIM Next Generation: Overcoming Challenges in Modernising a Farming Systems Model", *Environmental Modelling & Software*, Vol. 103, pp. 43-51. ISSN 1364-8152. DOI 10.1016/j.envsoft.2018.02.002.

[6] Kambadur, M., Tang, K., Lopez, J. and Kim, M. A. (2013) "Parallel scaling properties from a basic block view", ACM SIGMETRICS *Performance Evaluation Review*, Vol. 41, pp. 365-366. ISSN 0163-5999. DOI 10.1145/2494232.2465748.

[7] Kim, K. S., Yoo, B. H., Shelia, V., Porter, C. H. and Hoogenboom, G. (2018) "START: A data preparation tool for crop simulation models using web-based soil databases", *Computers and Electronics in Agriculture*, vol. 154, pp. 256-264. ISSN 0168-1699. DOI 10.1016/j.compag.2018.08.023.

[8] Kirby, A. C., Yang, Z., Mavriplis, D. J., Duque, E. P. N. and Whitlock, B. J. (2018) "Visualization and data analytics challenges of large-scale high-fidelity numerical simulations of wind energy applications", *AIAA Aerospace Sciences Meeting*. AIAA SciTech Forum, Kissimmee, Florida. DOI 10.2514/6.2018-1171.

[9] Mass, C. F., Ovens, D., Westrick, K. and Colle, B. A. (2002) "Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific northwest", *Bulletin of the American Meteorological Society*, Vol. 83, No. 3. DOI 10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

[10] Reinmuth, E. and Dabbert, S. (2017) "Toward more efficient model development for farming systems research - An integrative review", *Computers And Electronics In Agriculture*, Vol. 138, pp. 29-38. ISSN 0168-1699. DOI 10.1016/j.compag.2017.04.007.

[11] Skoogh, A., Michaloski, J. and Bengtsson, N. (2010) "Towards continuously updated simulation models: Combining automated raw data collection and automated data processing", *Proceedings - Winter Simulation Conference*, pp. 1678-1689. ISSN 08917736. DOI 10.1109/WSC.2010.5678901.

[12] Stoces, M., Masner, J., Kanska, E. and Jarolimek J. (2018) "Processing of Big Data in Internet of Things and Precision Agriculture", *Agrarian Perspectives XXVII.: Food Safety - Food Security, Proceedings of the 27th International Scientific Conference*, pp. 353-358. ISBN 978-80-213-2890-7. ISSN 1213-7979.

[13] Szufel, P., Czupryna, M. and Kaminski, B. (2017) "Optimal execution of large scale simulations in the cloud. The case of route-To-pa sim online preference simulation", *Proceedings - Winter Simulation Conference*, pp. 3702-3703. DOI 10.1109/WSC.2016.7822408.

[14] Zhao, G., Bryan, B. A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X. and Yu, Q. (2013) "Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing", *Environmental Modelling & Software*, Vol. 41, pp. 231-238. ISSN 1364-8152. DOI 10.1016/j.envsoft.2012.08.007.

5.7 Support Tools for Agricultural Production Simulation Processing

Pavlík, J., Vaněk, J., Masner, J., Stočes, M., Očenášek, V. (2020) „Support tools for agricultural production simulation processing“ *9th International Conference on Information and* Communication *Technologies in Agriculture, Food and Environment (HAICTA 2020) 24.09.2020, Soluň, Řecko*. CEUR-WS.org, s. 468-474.

Support Tools for Agricultural Production Simulation Processing

Jan Pavlik[1], Jiri Vanek[1], Jan Masner[1], Michal Stoces[1], Vladimir Ocenasek[1]

[1]Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Abstract. The APSIM software has proven to be extremely valuable decision support tool when it comes to optimizing agricultural management practices in order to maximize yield. Conducting high amount of simulations naturally requires processing of large volumes of data, therefore the available time and hardware resources create a limit for the scale of production simulation modelling. If APSIM is to be effectively used at a local level utilizing pre-existing hardware infrastructure, an assessment of available resources must be conducted in order to optimize the scale of the simulation. Another issue is the availability of personnel with enough information technology skills and experience to conduct the processing. The focus of this paper are software automation and other assistance tools that are therefore required for production modelling to be successfully utilized by small to medium enterprises.

Keywords: APSIM; scalability; hardware requirements; data processing; automation software.

1 Introduction

Maximizing yields of agricultural production is one the critical issues for society today. Growing worldwide population exacerbates the need for sufficient food production while increasing climate anomalies such as droughts can pose a great risk for crops. One of the approaches to maintain and improve agricultural yields is to utilize information technology to optimize managerial strategies and genotype selection by conducting multi-factor analysis in form of simulations or modelling (Holzworth et al., 2014).

The development of information technology hardware provides increasingly more technological resources to conduct simulation processing on larger scales, however as shown by (Li and Li, 2014) the increase of available data, such as higher resolution of geographical data, creates hardware limitations when scaling up the processing. This is especially important when trying to utilize agricultural simulations on a local level, for instance in small to medium agricultural companies. Due to the lack of financial resources for purchasing dedicated hardware there is a need to utilize pre-existing infrastructure. Most of the computing capacity in these companies is provided by out of date machines, meaning that any large-scale data processing involving high level of

Proceedings of the 9th International Conference on Information and Communication Technologies in Agriculture, Food & Environment (HAICTA 2020), Thessaloniki, Greece, September 24-27, 2020.

parallelization such as described by (Zhao et al., 2013) is out of the question. The main approach to utilize such hardware requires individual machine optimization alongside parallelization as shown by (Bartonek, 2017). Another option would be to utilize cloud computing, but as pointed out by (Szufel, Czupryna and Kaminski, 2017) it necessary to highly optimize cloud processing in order to maintain low costs.

Second issue is the lack of qualified employees. Especially in agriculture there is a distinct lack of IT proficient workers as pointed out by (Reinmuth and Dabbert, 2017). This results in a need for easy to use support tools that would automate the simulation setup and processing. Some of these tasks are already incorporated within the APSIM software and as stated by (Holzworth et al., 2018) the ease of use and focus on automation will be one of the integral parts of next versions of APSIM.

2 Simulation Workflow

As shown in Figure 1, the main workflow of agricultural simulation processing can be divided into four main steps. Firstly, it is necessary to establish the correct scale of the processing. When adjusting the number of options for each factor, the total number of simulations that needs to be processed changes accordingly. Since the number of simulations is essentially a product of the number of varying options, restraining variability only to the most relevant settings is the main way to reduce the processing time.

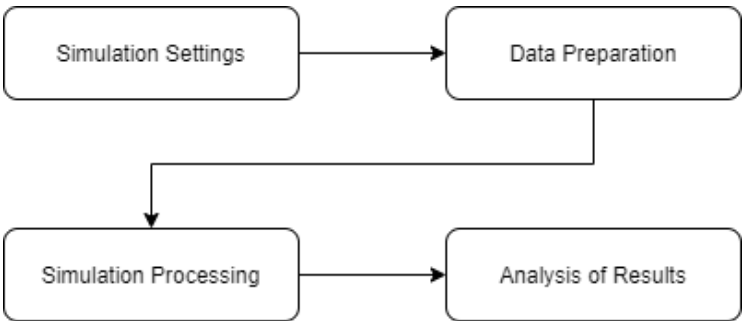


Fig 1. Basic simulation workflow

The factors and options that are determined in this first stage can be for instance different soil types, plant genotypes, different weather condition scenarios, and settings involving various managerial practices such as time of sowing, time of harvest, irrigation, fertilization, number of plants per meter square etc. (see Figure 2).

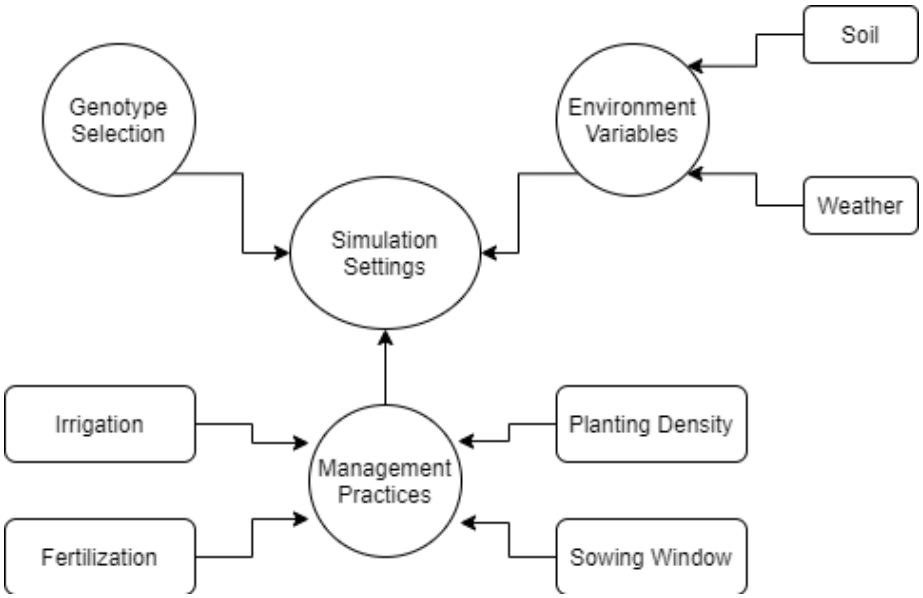


Fig 2. Basic concept of GxExM framework (genotype, environment, management)

Second step is the data preparation and preprocessing. Each simulation is stored as a single xml file. When APSIM is being deployed on dedicated hardware, is it possible to generate these files "on the fly" during the processing. However as pointed out by (Jarolimek et al., 2019) the hardware components, mainly processor and RAM, constitute a limit on to how many simulations at once can APSIM handle. Therefore, in order to maximize effectiveness, it is necessary to optimize simulation batch sizes on a per machine basis, that is why the data preparation and preprocessing step is unavoidable when trying to utilize sub-par preexisting hardware in smaller or medium enterprises.

The actual simulation processing itself should be conducted during downtimes such as nights, when the infrastructure is not required for other critical operations. This part is generally the most time consuming and therefore also the most likely to significantly improve the overall efficiency if automation and scheduling is utilized. The hardware dependence of the processing can result in bottlenecks if the previous two steps were not conducted properly.

The analysis of results is essentially a statistical data analysis and can therefore be conducted using tools like MS Excel or more specialized software such as SAS, STATISTICA etc. The most basic analysis would consist of simply taking the simulations that produced highest yields and finding commonalities in the simulation settings.

3 Automation and support tools

3.1 Simulation Settings

The number of total simulations that needs to the processed determines the scale of the processing. The calculation of the number of simulations is a simple multiplication of options for each simulation settings. The scale can be therefore decreased or increased by simply adjusting the number of options. When scaling up, it is possible to include more options or "what-if" scenarios and generally explore more combinations. When there is a need to scale down, the most likely best approach is to use hands on experience of local producers or company agronomists to narrow down the simulation option settings only to a few variations that were historically most effective or make most sense in terms of the common local managerial practices.

Therefore, the important question in step one is how many simulations we should aim for. In order to properly select the scale of simulation an estimate must be made, based on allotted time and the hardware available for the computation. And it is in this point where the absence of experienced IT employees creates first problems. Unlike bigger companies or corporations that might have dedicated IT departments, SMEs generally lack specialists that would be able to adequately estimate the capabilities of older hardware when it comes to processing large number of simulations.

The support tools for this step could be a simple web application where the user inputs basic hardware information of their machines including processing power and RAM and the application will estimate number of simulations that can be run per day or per hour.

3.2 Data Preparation

The main goal of this step is to gather all input information for the APSIM software. There is a possibility for integration between APSIM and existing agricultural software to partially automate this process, similar to that outlined by (Skoogh, Michaloski and Bengtsson, 2010). Another option is to utilize integration to existing knowledge databases such as soil or weather databases as explored by (Kim et al., 2018). But due to the lower scale and therefore limited number of option settings, this is not necessary, since manual entry of the input data is not very time consuming.

The other part of preprocessing consists of generating the APSIM simulation files beforehand and grouping them into various size batches optimized on a per machine basis. The degree of detail this task needs to be performed to depends on the sophistication of the software automation tools used in the following step. As shown by (Pavlik et al., 2019) it is possible to develop an application that can combine some of the work required in steps two and three and handle both the batch APSIM file generation and the processing scheduling and automation.

3.3 Simulation Processing

There are four basic approaches to automate the simulation processing:

1. Use built-in APSIM capabilities
2. Process simulations ex-situ – on the cloud
3. Use existing software tools for automation
4. Develop new software specifically designed to automate APSIM simulations

As explained earlier, using existing APSIM options for automation might not be possible due to the hardware limitations. Processing on the cloud is in essence similar to purchasing dedicated hardware. It might be cheaper, but in this paper, we are focusing mainly on exploiting already existing hardware and infrastructure. This leaves us with options three and four.

There are many existing software tools to automate tasks. Whether it be task schedulers that already come with the operating system, or dedicated automation software, such as HTCondor. The advantage of using such tools is that they include many useful functions such as workload monitoring, virtualization, checkpointing and can also combine serialized batch processing with parallelization. Therefore, this option is preferable when used on hardware that also performs other day to day tasks. The main disadvantage is that the previous and following steps (data preprocessing and result analysis) will require more work since they cannot be incorporated into these existing tools, not even partially.

The last approach is to design and develop brand new tools specifically to automate APSIM simulation processing. A custom-built tool could potentially encompass more that just the processing automation. It can theoretically handle all four steps of the workflow, bridging the gaps between the parts. This could be very valuable considering the lack of experienced IT employees in smaller agricultural companies. However, it will come at a cost of lower sophistication of automation and it will be harder to combine the processing with other necessary tasks the hardware needs to perform on a daily basis. This approach is therefore better suited for situations when the company can set aside one or several computers, perhaps older or currently unused machines, and fully use them towards the simulation processing.

3.4 Analysis of Results

If the analysis of the results is to be conducted in a separate statistical software, it is necessary to convert the output simulation from the basic text format into .csv or .xsl. This can be achieved with an extraction program or a script that will parse the output files and siphon only the necessary data.

However, when simulating agricultural production on a local level, the only important outputs are the simulation settings associated with the highest yields. A complex statistical analysis may therefore not be necessary. In case of developing new custom software for the previous processing steps, the extraction of yield data, selection of top simulations and any eventual visualization of simulation settings can be added to it, resulting in an overarching support tools that can automate vast majority of the workflow processes.

4 Conclusions

The paper discussed two main approaches to conducting agricultural production simulations in smaller or medium companies. The first approach is to split the process into logical parts and optimize and automate them separately, utilizing either already existing software tools, or developing a smaller programs for individual tasks, such as format conversion tools, simulation generation tools, data parsers etc. The main advantages of this approach are higher optimization and better interoperability when utilizing existing hardware that cannot be fully dedicated to the task. However, such a solution would require employee skilled in IT, prompting additional need for financial resources in order to hire or train someone.

The second approach consists of developing an overarching support tool that would incorporate all the various processing tasks. If developed as a general purpose streamlined software with focus on ease of use, it could overcome the problem with lack of experienced IT workers in small to medium companies. The disadvantages of this solution are harder incorporation with existing agricultural software running parallel on the same machines and loss of modularity options such as in-depth data analysis or conducting simulations for goals other than maximizing yields.

Acknowledgment: The results and knowledge included herein have been obtained owing to support from the following institutional grant. Internal Grant Agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 2019MEZ0005 – "Optimization of management practices in sorghum production under uncertain future weather conditions".

References

1. Bartonek, D. The Possibilities of Big GIS Data Processing on the Desktop Computers (2017). RISE OF BIG SPATIAL DATA Book Series: Lecture Notes in Geoinformation and Cartography, pp. 273-287. DOI: 10.1007/978-3-319-45123-7_20
2. Holzworth, D., Huth, N. I., de Voil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G. et al. (2014) "APSIM - Evolution towards a New Generation of Agricultural Systems Simulation." Environmental Modelling & Software, Vol. 62, pp. 327-350. ISSN 1364-8152. DOI 10.1016/j.envsoft.2014.07.009.
3. Holzworth, D., Huth, N. I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N. I., Zheng, B. and Snow. V. (2018) "APSIM Next Generation: Overcoming Challenges in Modernising a Farming Systems Model", Environmental Modelling & Software, Vol. 103, pp. 43-51. ISSN 1364-8152. DOI 10.1016/j.envsoft.2018.02.002
4. Jarolímek, J., Pavlík, J., Kholova, J. and Ronanki, S. (2019) "Data Pre-processing for Agricultural Simulations", AGRIS on-line Papers in Economics and Informatics, Vol. 11, No. 1, pp. 49-53. ISSN 1804-1930. DOI 10.7160/aol.2019.110105

5. Kim, K. S., Yoo, B. H., Shelia, V., Porter, C. H. and Hoogenboom, G. (2018) "START: A data preparation tool for crop simulation models using web-based soil databases", Computers and Electronics in Agriculture, vol. 154, pp. 256-264. ISSN 0168-1699. DOI 10.1016/j.compag.2018.08.023
6. Li, Q., Li, D. Big data GIS (2014). Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics and Information Science of Wuhan University, vol. 39, iss. 6, pp. 641-644+666. DOI: 10.13203/j.whugis20140150
7. Pavlík, J., Masner, J., Jarolímek, J., Lukáš, M. (2019) "Data Processing for Yield Optimization", Agrarian perspectives XXVIII. – Business Scale in Relation to Economics, pp. 189-193.
8. Reinmuth, E. and Dabbert, S. (2017) "Toward more efficient model development for farming systems research - An integrative review", Computers And Electronics In Agriculture, Vol. 138, pp. 29-38. ISSN 0168-1699. DOI 10.1016/j.compag.2017.04.007
9. Skoogh, A., Michaloski, J., Bengtsson, N. Towards continuously updated simulation models: Combining automated raw data collection and automated data processing (2010). Winter Simulation Conference, pp. 1678-1689. DOI: 10.1109/WSC.2010.5678901
10. Szufel, P., Czupryna, M. and Kaminski, B. (2017) "Optimal execution of large scale simulations in the cloud. The case of route-To-pa sim online preference simulation", Proceedings - Winter Simulation Conference, pp. 3702-3703. DOI 10.1109/WSC.2016.7822408.
11. Zhao, G., Bryan, B. A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X. and Yu, Q. (2013) "Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing", Environmental Modelling & Software, Vol. 41, pp. 231-238. ISSN 1364-8152. DOI 10.1016/j.envsoft.2012.08.007

5.8 An APSIM-powered framework for post-rainy sorghum-system design in India

Ronanki, S., **Pavlík, J.,** Masner, J., Jarolímek, J., Stočes, M., Subhash, D., Talwar, H., Tonapi, V., Srikanth, M., Baddam, R., Kholová, J. (2022) „An APSIM-powered framework for post-rainy sorghum-system design in India“ *Field Crops Research*, 2022, sv. 277, č. 108422. ISSN: 0378-4290. Dostupné na: https://doi.org/10.1016/j.fcr.2021.108422

Field Crops Research 277 (2022) 108422

Contents lists available at ScienceDirect

Field Crops Research

journal homepage: www.elsevier.com/locate/fcr

An APSIM-powered framework for post-rainy sorghum-system design in India

Swarna Ronanki [a,1,2], Jan Pavlík [b,1,3], Jan Masner [b,*,4], Jan Jarolímek [b,5], Michal Stočes [b,6], Degala Subhash [c], Harvinder S. Talwar [a], Vilas A. Tonapi [a], Mallayee Srikanth [c], Rekha Baddam [c], Jana Kholová [c,*,7]

[a] ICAR, Indian Institute of Millets Research, Hyderabad 500 030, Telangana, India
[b] Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, Prague 165 00, Czech Republic
[c] GEMS team, International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad 5023204, Telangana, India

ARTICLE INFO

Keywords:
APSIM
Post-rainy
Sorghum
GxExM
Agri-system design

ABSTRACT

Sorghum contributes to the livelihoods of millions of food-insecure households in semi-arid agri-systems. Annual production widely fluctuates throughout India due to erratic rainfall and suboptimal agronomic practices. Our novel approach utilizes the digital reflection of post-rainy (rabi) sorghum production systems in India to help better understand its spatio-temporal variations and enable the designing of geography-specific, climate-responsive system interventions (Genotype × Management; GxM). For this, we evaluated a range of farmer-relevant agronomic management practices across three soil types (shallow, medium, and deep vertisols) in combination with observed ranges of biological variability in sorghum cultivar characteristics. We used the crop growth simulation model Agricultural Production Systems sIMulator (APSIM) to identify GxM combinations that can support the enhancement/ stability of post-rainy sorghum production systems in India. In general, we found the post-rainy sorghum systems would benefit from early-season sowing (16th - 23rd September), short crop duration (compared to Maldandi (M35–1), commonly grown crop type), and medium fertilizer inputs (70–70 kg urea ha^{-1} as basal and top-dress application). In addition, site-specific crop management (M) and crop characters (G) optimizations would further enhance/ stabilize sorghum production. Simulations highlighted that in the poorly-endowed environmental context (i.e. shallow soils and low-rainfall areas), optimal G×M targets might involve water conservation GxM combinations, such as low plant populations and low fertilization along with low crop vigor and limited transpiration responsiveness. Details on site-specific optimum GxM are available in a web application at https://ls40.pef.czu.cz/maps/. To enable the use of the study outputs for certain applications (e.g. breeding), we separated the examined geographies based on similarities in optimum production characteristics and similarities in system response to GxM interventions into four "homogeneous system units" (HSU; i.e. geographical units within which reduced GxM interactions are expected). These HSUs intended to offer geography-specific targets to prioritize, test, and introduce distinct G×M interventions. We conclude that the APSIM-powered framework presented provides region-specific Genotype × Management options that could become a blueprint for defining quantitative breeding targets that achieve enhanced productivity/ stability of dry-season sorghum cultivation throughout India.

* Corresponding authors.
[1] These authors have contributed equally
[2] ORCID: 0000-0003-1606-5522
[3] ORCID: 0000-0002-6136-0785
[4] ORCID: 0000-0003-4593-2306
[5] ORCID: 0000-0001-7194-3055
[6] ORCID: 0000-0001-7128-1071
[7] ORCID: 0000-0001-7133-1382

https://doi.org/10.1016/j.fcr.2021.108422
Received 15 April 2021; Received in revised form 13 December 2021; Accepted 19 December 2021
Available online 10 January 2022

1. Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is a largely variable crop species adapted to cultivation across tropical to temperate climates and grown primarily for human food and animal feed as well as for the production of biofuels (Habyarimana et al., 2019). In India, sorghum is one of the few multi-purpose, resilient crops suitable for marginal lands during the post-rainy (*rabi*) season (typically September - January) and supports the livelihoods of millions. Frequent droughts caused by climatic variability combined with low input agronomic practices are the main reasons why the farmer's yields across the rabi sorghum tract fluctuate and the average grain yields stagnate at ~800 kg ha^{-1} despite yield potential being much higher (~3500 kg ha^{-1}; Ambadi et al., 2018; Dayakar Rao et al., 2009). A sensible way to bridge this yield gap is to analyze the major constraints of the production system (Kholová et al., 2013) and design the appropriate Genotype × Management (G×M) interventions to lift current yields closer to their potential (Soltani et al., 2016; Pradhan et al., 2015; Chauhan and Rachaputi, 2014; Kholová et al., 2014). Traditionally, multi-location field trials are used to evaluate cultivar, management, and environment interactions (GxExM) in-situ. However, field trials are time and resource-consuming and results are often difficult to extrapolate to other sites and seasons. In this situation, validated crop modeling set-ups in conjunction with field data can extrapolate the GxExM analyses across the spatio-temporal scales and can be used to capture the system's behavior and fluctuating GxExM interactions. By doing so, this approach can complement in-vivo field observations with in-silico predictions which could not be covered experimentally (Jones et al., 2017). For sorghum, several crop models have been implemented in simulation frameworks such as the Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al., 2003), Agricultural Production Systems sIMulator (APSIM) (Holzworth et al., 2015) or Samara (Dingkuhn et al., 2011). These models differ in the implementation of algorithms to capture the soil-crop-atmosphere interactions.

In prior studies, we found that an APSIM based set-up can reliably reflect the agronomy of post-rainy season sorghum production systems (Kholová et al., 2013, 2014). Therefore, in our present work, we aim to expand the existing post-rainy sorghum simulation set-up and deploy the structure in order to identify the optimum Genotype × Management options to improve/ stabilize sorghum production. This analysis is intended to support crop improvement program decision making on region-specific crop and management interventions that can potentially improve/ stabilize production across the rabi sorghum tract in India. This is presented in the form of an open-access interactive web-based tool to ensure stakeholders access and use.

2. Materials and methods

2.1. Overview

The majority of the Indian rabi sorghum grain is produced in Maharashtra, Karnataka, Andhra Pradesh, and Telangana (as per Kholová et al. (2013) and is the unique production system prioritized for this study. The parameters for three soil types characteristic of these major production areas were collated from available databases (National Bureau of Soil Survey and Land Use Planning, International Soil Reference and Information Centre). The gridded meteorological information was obtained from Agricultural Modern-Era Retrospective Analysis for Research and Applications (AgMERRA) – National Aeronautics and Space Administration (NASA) and evaluated as most suitable when tested against the observed meteorological information (also in Hajjarpoor et al., 2018). Crop simulations were developed using the sorghum model in APSIM (see Holzworth et al., 2015; Keating et al., 2003; Hammer et al., 2010). The rabi sorghum crop type M35–1 and its validated genotypic coefficients were used as a base for the agri-system evaluation (Hammer et al., 2010; Ravi Kumar et al., 2009; Kholová et al., 2013, 2014). This base was further expanded with system-relevant combinations of management practices and rabi-sorghum relevant cultivar parameters. The spatio-temporal information on optimum Genotype x Environment and production parameters were finally used to separate the region into clusters with higher levels of similarities in these characteristics.

2.2. APSIM sorghum module

APSIM set-ups from previous work (Ravi Kumar et al., 2009; Kholová et al., 2013, 2014) were used in this study to simulate sorghum growth and development with a range of weather and soil information, management, and genetic coefficients representing the major post-rainy sorghum production regions. Altogether, we ran 4,299,264 simulations to analyze the post-rainy sorghum production system in India. A detailed description of the APSIM model is available in Holzworth et al. (2014, 2015) and Hammer et al. (2010). In short, the APSIM sorghum module algorithms process the interactions between the daily weather (rainfall, minimum and maximum temperature, solar radiation) and soil inputs considering the crop management practices and crop genetic coefficients to arbitrate the daily status of the soil-crop-atmosphere continuum and integrates this information into comprehensive outputs on crop development, growth and the production used for further analysis in this study.

2.3. Model inputs

2.3.1. Soil information

Throughout the main rabi sorghum production tract in India (Maharashtra, Karnataka, Andhra Pradesh, and Telangana), sorghum is usually grown on vertisols (International Soil Reference and Information Centre; Kumar et al., 2017). The variation in soil depth and water holding capacity significantly influences crop production. Accordingly, for each simulation unit, the Genotype × Management options were tested in the context of the 3 vertisol composites (bulk density ~1.4 g cm-3; ~0.7% organic carbon; C:N ~14.5) with varying soil depth and plant available water (PAW); i.e. shallow (70 cm depth; 94 mm PAW); medium (105 cm depth; 132 mm PAW); deep (150 cm depth; 144 mm PAW). In all soils, the soil nitrogen content was set for 50 kg ha^{-1} NO3 and 10 kg ha^{-1} NH4. The soil conditions were automatically re-initialized before each season's simulation. These soil parameters were compiled from the reports of measured soil parameters gathered by the International Soil Reference and Information Centre (ISRIC) and the National Bureau of Soil Survey and Land Use Planning (NBSS & LUP) in Bangalore.

2.3.2. Weather information

As there is a general lack of quality weather information accessible in India, we tested several commonly used synthetic weather data (daily Tmin, Tmax, rainfall) from different sources (Agricultural Modern-Era Retrospective Analysis for Research and Applications (AgMERRA; https://data.giss.nasa.gov/impacts/agmipcf/agmerra/), NASA-POWER (https://power.larc.nasa.gov) and MarkSim (MarkSim® GCM - DSSAT weather file generator (cgiar.org)). To complement these datasets, solar radiation was estimated using an algorithm based on sunshine hours and extraterrestrial radiation (Soltani and Hoogenboom, 2003; Soltani and Sinclair, 2012). The synthetic weather data was then compared with the observed weather information according to i) their agreement with observed Tmax and Tmin and sum of rainfall and ii) the agreement between the mean simulated yields using observed weather data were compared against yields using synthetic weather data from the same locations. The distribution of meteorological stations and records used for comparison with synthetic data is described in Supplementary Fig. 1.

We used standard metrics to indicate the goodness of fit; i.e. correlation coefficient (R2), root mean square error (RMSE - Eq. 1), and index of agreement (d-index). The d-index was proposed by Willmott et al.

(1985) specifically for modeling studies. D-index value range is −1–1 (Eq. 2) and, accordingly, a D-index value closer to one indicates closer agreement between the two variables compared.

Equations are listed below:

$$RMSE(root\ mean\ square\ error) = \left[\sum_{i=0}^{n} \frac{(p_i - o_i)^2}{n}\right]^{0.5} \quad (1)$$

$$Index\ of\ agreement(d) = 1 - \left(\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}\left(|P_i'| + |O_i'|\right)^2}\right) \quad (2)$$

where n is the number of observations, Pi is the predicted observation, Oi is a measured observation, and P_i'= Pi - M and O_i' = Oi - M (M is the mean of the observed variable).

2.3.3. Crop specific coefficients used in the APSIM model

Crop coefficients need to be specified in APSIM-sorghum to reflect the growth and developmental characteristics of a crop cultivar. In our case, we used rabi sorghum – Maldandi, M35–1 (Kholová et al., 2013; Ravi Kumar et al., 2009; tt_endjuv_to_ini = 250; TPLA max = 2.8; VPD responsiveness = 0.95) crop coefficients to define a rabi sorghum "template". The synthetic cultivars were created by altering the M35–1 genotypic coefficients (Kholová et al., 2014) which had previously been found relevant for the development of improved rabi-sorghum cultivars, extensively studied, quantified, and tested in-vivo and in-silico using the APSIM platform (Kholová et al., 2013, 2014; Ronanki et al., 2018; Vadez et al., 2011). These include 1) "tt_endjuv_to_ini" which corresponds to a duration of end-of-juvenile to panicle-initiation developmental phase [thermal time units; TT] and specify the crop cycle duration; 2) coefficient of "TPLAmax" function which specifies the growth of total plant leaf area during the plant development, and 3) "VPD responsiveness" which defines the crop transpiration responsiveness to vapor pressure deficit (Hammer et al., 2010; https://www.apsim.info/documentation/model-documentation/crop-module-documentation/sorghum/). The parameters and their ranges used in this study are reported in Table 1.

Table 1
Overview of the variation in crop management (M) practices and in crop genetic (G) coefficients (representing a range of biologically relevant crop variants) tested by the crop growth model in the context of different soils. These resulted in 13,824 GxM combinations that were simulated within 311 grids to evaluate the optimum GxM supporting crops production/ resilience across post-rainy sorghum production systems in India. Here the tt_endjuv_to_ini corresponds to the duration of the end of juvenile to panicle initiation phase [thermal time units]; TPLA stands for total plant leaf area and corresponds to the power coefficient of the TPLAmax function, VPD stands for vapor pressure deficit of two crop types (VPD responsive and non-responsive crop types were created as detailed in Kholová et al., 2014).

M/ G/ soil	G/M/soil variation (APSIM coefficient/ module used)	Range of variation in G/M/soil (varied unit)
soil	Soil	Shallow soil (70 cm); Medium soil (105 cm); Deep soil (150 cm)
M	Sowing window	16th September – 23rd September; 23rd September – 30th September; 30th September – 7th October; 7th October - 14th October; 14th October – 21st October; 21st October – 28th October
M	Planting density	6; 8; 10; 12; 14; 16 plants m^{-2}
M	Nitrogen fertilization (Urea application schedule)	0–0; 20–20; 50–50; 100–100 kg ha^{-1}
G	Crop duration (tt_endjuv_to_ini; [TT])	Very Early (150); Early (200); Medium (250); Late (300)
G	Rate of canopy growth, vigor [power coefficient for TPLA max function in APSIM]	Low (2.4); Medium (2.6); High (2.8); Very high (3.0)
G	Transpiration responsiveness [Capacity of the canopy to limit transpiration in high VPD]	Low (0.95); High (0.80)

2.4. Simulations setup

APSIM is a process-based cropping systems simulation tool capable of reproducing the range of agronomic interventions and the base of several commercial applications; e.g. YieldProphet ® (Yield Prophet) (Hochman et al., 2009), WhopperCropper (The Regional Institute - J. Managing Climate Variability - Crops), CropARM (Decision support tools and modeling | Tasmanian Institute of Agriculture (utas.edu.au) (Richter et al., 2017). The APSIM sorghum module v. 7.6, with incorporated algorithms enabling simulations of crop transpiration responsiveness to atmospheric drought (details in Kholová et al., 2014), was set-up for each of the 311 gridded weather time-series (31 seasons; AgMERRA-NASA series), soils typically sown to rabi sorghum in the region (shallow, medium and deep vertisol; Trivedi, 2009; Jirali et al., 2010; https://www.millets.res.in/farmer/Recommended_packages_of_practices_Rabi_sorghum.pdf), 3 cultivar-specific parameters representing the biological variation in sorghum crops (G) and a range of management practices (M) relevant for the region. This resulted in 4, 299,264 simulations (Table 1). The baseline for the simulations was inspired by the recommended management practices for growing post-rainy season sorghum documented by Rooney et al. (2007), Trivedi (2009), and Olson (2012). The range of variation in the crop management practices (Ravi Kumar et al., 2009) relevant for the region was used as per the discussion with experts from the Indian Institute of Millets Research (IIMR), and International Crops Research Institute for Semi-Arid Tropics (ICRISAT) crop improvement teams and farmers (Table 1). Dimes and Revanuru (2004) previously tested the suitability of APSIM to reproduce these M interventions (Nitrogen), Turner and Rao (2013) looked at plant density and cultivar duration and Akinseye et al. (2020) sowing dates. The sowing within each of the specified sowing windows (Table 1) was triggered by a minimum of 9 mm of rainfall within 5 days. Upon meeting these requirements, APSIM initiated the sowing with the specified combination of inputs. The soil carbon and nitrogen were re-initialized before each sorghum season. The soil moisture profile at sowing was assumed to be fully saturated after the rainy season in all grids and, in addition, farmers often use irrigation after sowing to ensure germination (Trivedi, 2009). After setting up the simulation runs, all the Genotype × Management combinations were evaluated in-silico in all grids covering the rabi sorghum production regions.

2.5. Automation of APSIM runs with C#

The APSIM sorghum model was run using environmental data spanning 31 years with a total of 13,824 Genotype × Management combinations in 311 grids. In total 4,299,264 simulations were performed to generate the complete set of output files. This is a time-consuming task and a single commodity computer would take years to run this amount of simulations (Jarolímek et al., 2019). Additionally, the system requires huge storage capacities to hold the generated output files. Despite the fact that APSIM is designed to perform intended runs, the number of simulations exceeded the capacity of its in-built features. Therefore, we used supporting software tools for automation and scheduling of the designed factorial runs to tackle the large number of simulations. Simulations were generated by a custom software solution developed using the.NET framework and C# programming language. This resulted in a special software application that scheduled and generated simulation runs in batches as per the computational capacity of the available high-performance computing facility at the Czech University of Life Sciences Prague (128 GB RAM, 16 core AMD EPYC 7281 2.7 GHz CPUs). Therefore, the batches were run in parallel on 7

Table 2
The table shows the weightage of simulated production quantity (grain yield, stover yield) and production stability (biomass deviation, years with grain yield failure) indicators for the construction of the simulation weighting index. The value of each of the indicators was weighted (%) for the construction of a single simulation index (Eqs. 3–7). This was used to evaluate a particular GxM combination across 31 seasons of simulations within a particular grid and separately for both scenarios.

Scenario/ parameter weightage	Grain Yield, Eq. (3) (average of 31 simulated years)	Stover Yield,Eq. (3) (average of 31 simulated years)	Biomass deviation;Eq. (3) (across 31 simulated years)	Frequency of years with grain yield failure;Eq. (4) (across 31 simulated years)
Production	70%		30%	
	40%	30%	15%	15%
Stability	30%		70%	
	17.14%	12.86%	35%	35%

high-performance computers. Further technical details on this process are documented in Jarolímek et al. (2019).

2.6. Output analysis and visualization using interactive online tools and maps

Each of the simulation output files containing a particular Genotype × Management scenario generated for 31 seasons within each grid was evaluated for the main agronomically important parameters linked to the production quantity (mean of grain yield, stover yield, Eq. 3), and production stability (frequency of years with yield failure, Eqs. 4–6 and standard deviation of total biomass - grain and stover, Eq. 3). This was achieved by creating the index to weigh each of these outputs according to its anticipated importance of the farmers' demands on sorghum production in two scenarios: "production" and "stability" scenarios (Table 2, Eqs. 3–7; similarly in Thornton et al., 2018). The range of approaches to quantify agri-system production and stability were comprehensively reviewed in Zampieri et al. (2020) (used in e.g. Descamps et al., 2018, Thornton et al., 2018). In this work, we adapted some of these simple concepts to evaluate production and stability based on the available crop model outputs.

The "production scenario" intended to reflect the likely demand of the more economically secure sorghum farmers and increased the weightage of production indicators (mean of grain yield, stover yield, Eq. 3, Table 2). The "stability scenario" was designed to reflect the likely needs of economically vulnerable sorghum farmers and so the proportionally higher weightage was introduced to production stability indicators, minimizing the probability of grain yield failure and year-to-year total biomass fluctuations, (Eqs. 4–6, Eq. 3; Table 2). Accordingly, in the production scenario, the index considered the production factor weight of 70% (grain and stover yield; 40%; 30%) and 30% weightage of stability indicators that penalized production fluctuations and yield failure, i.e. frequency of years where crops failed to reach the grain filling stage with grain yield 0 and biomass deviation; 15%, 15% respectively (Table 2, Eqs. 3–7). In the "stability" scenario, the higher weightage was introduced to the stability indicators, i.e. weightage of production factors was only 30% (grain and stover yield; 17.14%; 12.86% respectively) with 70% weightage on stability indicators, i.e. frequency of years with yield failure and biomass deviation; 35%, 35%, respectively (Table 2, Eqs. 3–7).

The simulation index was calculated for each grid and combination of GxM and is an aggregated value of several features representing the simulation's time series. A simulation index consisted of grain yield, stover yield, and biomass deviation scores which were calculated as differences from the mean normalized by dividing by standard deviation, which is often referred to as "standard score" or "z-score" in statistics. The simulation average was calculated from the 31 years of simulation data for each particular GxM combination, while the overall average and standard deviation was calculated from all simulations within the same soil group in the given grid (Eq. 3).

Grain yield, stover, and biomass deviation scores were calculated as standard score (calculated as difference from the mean divided by standard deviation):

$$standard\ score\ of\ X = \frac{average\ value\ of\ X - \ average\ value\ of\ ALL}{standard\ deviation\ of\ ALL} \quad (3)$$

where X is a currently evaluated simulation and ALL are all simulations within the same soil group in that particular grid.

Production failure score was calculated to penalize simulations that contained the years with grain yield failure (Eq. 4) or simulations where the ratio of successful growth years to ALL years was below 80% (Eq. 5). The final failure score (Eq. 6) considered the higher of the two values (which increase quadratically):

$$yield\ failure\ score = (ratio\ of\ yield\ failure\ *10)^2 \quad (4)$$

$$simulation\ failure\ score\ = \ max(((0.8\ -\ ratio\ of\ successful\ simulation\ years)*10)^2,\ 0) \quad (5)$$

$$total\ failure\ score\ = max(yield\ failure\ score,\ simulation\ failure\ score) \quad (6)$$

The final simulation index (production and stability) was then calculated by multiplying the scores above (Eqs. 3, 6) by weights depending on the scenario (see Table 2). For example, the calculation of simulation weighting index for the production scenario was:

$$simulation\ index\ = \ 0.4*grain\ score\ +\ 0.3*stover\ score - \ 0.15*biomass\ score\ -\ 0.15*total\ failure\ score \quad (7)$$

For each grid, the simulation index for production and stability scenarios was generated and the 10 simulations resulting in the highest index within each scenario selected. Within each scenario, 10 simulations with the maximum simulation index score were evaluated for the occurrence of particular Genotype × Management combinations. These were then stored in a database and can be use for visualization using an interactive web application available at https://ls40.pef.czu.cz/maps/; Source code is available at https://github.com/culs-fem-dit/APSIM-maps). The frontend uses React JavaScript framework and Google Maps API and the backend Nette PHP framework. Users can choose four parameters to be shown in the map - main variable, soil type, cultivation scenario, and cultivar (optimal G combinations or M35–1 representing Maldandi crop). Results for an entire grid are visualized in the form of a discrete heatmap. Additionally, users can show a second map for the comparison of different interventions and scenarios. In this way, the user can easily visualize the maximum attainable agronomically important parameters (grain and stover yield) with optimized Genotype and Management (or optimized M for currently grown M35–1 crop type) while understanding which Genotype and Management combinations lead to this outcome within specific geographic units.

Table 3
Statistical metrics used for evaluation of agreement between the three sources of gridded meteorological characteristics (AgMERRA-NASA, NASA-POWER, MarkSIM) with observed meteorological characteristics; R2 (Pearson's correlation coefficient), RMSE (root mean squared error), D-index. The actual correlations are visualized for AgMERRA-NASA data on Fig. 1a, b and Fig. 2a, b.

Weather Source	Meteorological/ agronomic characteristic	R2	RMSE	D-index
AgMERRA-NASA	Maximum Temperature (monthly means)	0.88	1.06	0.96
	Minimum Temperature (monthly means)	0.83	2.03	0.89
	Rainfall (monthly in-season mean)	0.86	0.61	0.96
	Grain Yield (site average)	0.88	691	0.96
	Biomass (site average)	0.88	1857	0.74
NASA-POWER	Maximum Temperature (monthly means)	0.68	1.74	0.89
	Minimum Temperature (monthly means)	0.79	1.46	0.94
	Rainfall (monthly in-season mean)	0.96	0.34	0.98
	Grain Yield (site average)	0.36	1520	0.27
	Biomass (site average)	0.27	2708	0.36
MARKSIM	Maximum Temperature (monthly means)	0.48	2.36	0.81
	Minimum Temperature (monthly means)	0.71	1.85	0.90
	Rainfall (monthly in-season mean)	0.74	0.85	0.95
	Grain Yield (site average)	0.25	1988	0.19
	Biomass (site average)	0.32	3047	0.19

2.7. APSIM output file analysis

2.7.1. Identification of GxExM for optimal production and resilience scenarios

Within each grid and soil type (representing a particular E) the output files representing particular Genotype × Management combinations were evaluated using the production and stability scenario index (see Section 2.6). For each simulation grid, the obtained scenario-specific indexes were sorted and the resulting distribution evaluated using interquartile range and z-score to detect possible outliers (details in Suppl. Fig. 2). This approach revealed that the top-end of the distribution does not contain any obvious outliers and that the approach of penalizing simulations with high yield failure or simulation failure rates (Eqs. 4 and 5) was sufficient to disqualify many of the simulations that appeared on the lower tail of the distribution. Additionally, the k-means clustering method was applied to visually inspect the proportion of the data that should be utilized for further analyses (Suppl. Fig. 2). After several manual iterations, we decided that the 10 simulations attaining the highest index for each scenario would be a sufficiently large sample to provide insight on the main characteristics of the sorghum system for a particular grid (Suppl. Fig. 2; Table 4a, b); These "10 best" simulations would now include the particular combinations of Genotype × Management leading to superior agronomic performance of the system (E in the particular grid) in the "production" and "stability" scenarios. The analysis outputs within the "10 best" simulations are summarized in Suppl. Fig. 3a, b, and 4 - further grid details can be visualized and dissected using the web application.

2.8. Identification of geographically homogeneous system units

For this task, the Principal Components Analysis (PCA) was run for each combination of grid, soil, and scenario (production, stability) for the characteristics that define the 10 simulations attaining the highest scenario-related index (see above). The loadings for 3 Principal Components (explaining altogether >80% of dataset variation) of each scenario and soil have been averaged across each grid. Resulted average loadings of 311 grid items were initially separated into 3, 4 and 5 clusters (R package; https://www.r-project.org/, Table 4), visualized and the cluster-specific production and stability characteristics calculated. Considering the 3–4–5 cluster characteristics and after consultation with experts (ICRISAT and IIMR sorghum breeding teams), 4 clusters appeared the most sensible to be effectively used in crop improvement programs (discussed in 4.3). Subsequently, geographical distribution of the cluster associated with each grid item was visualized using ArcGIS software v.1.0 and the main characteristics within each cluster summarized (Fig. 3; optimal Genotype × Management and agronomic characteristics of cropping system). This approach allowed us to separate the geographies with relatively similar responses to the cultivars and management interventions (i.e. "homogeneous system units").

3. Results

3.1. Selection of an appropriate source of meteorological input

To identify the most reliable source of meteorological input, three data sources were obtained and tested (AgMERRA-NASA, NASA-POWER, MarkSIM (these are described in detail in Ruane et al., 2015; Thornton et al., 2018; Jones et al., 2002; Rienecker et al., 2011). In all three datasets, there was a good agreement with the observed monthly temperature averages (Tmin, Tmax; Table 3). The monthly in-season

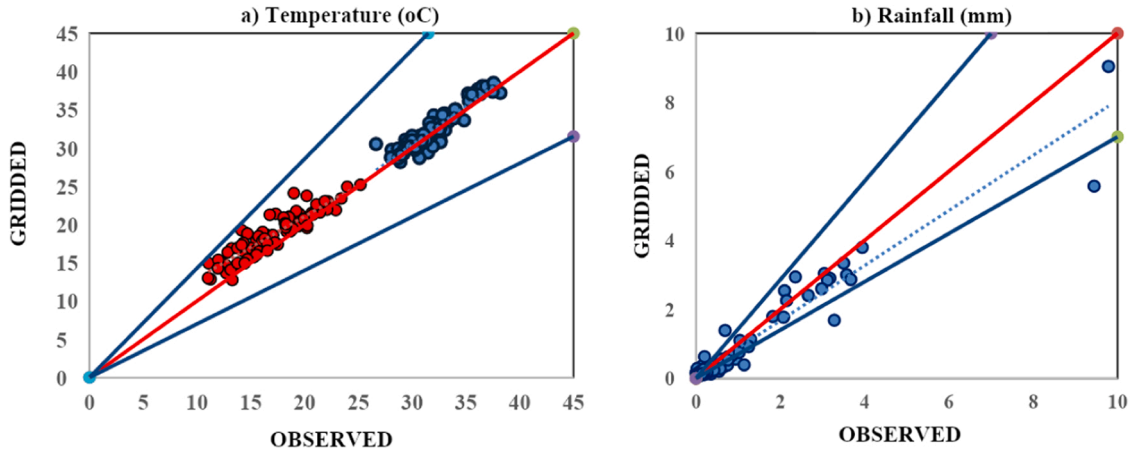


Fig. 1. a, b. Comparison of minimum (red circles) and maximum (blue circles) temperature from the gridded AgMERRA weather dataset with observed temperature (a) and comparison of in-season rainfall (October - March) from the gridded AgMERRA weather dataset with observed rainfall of the same period (b). In each graph, the middle (red) line represents 1:1 relation and the other (blue) lines represent the 30% divergence percentile of the 1:1 line. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

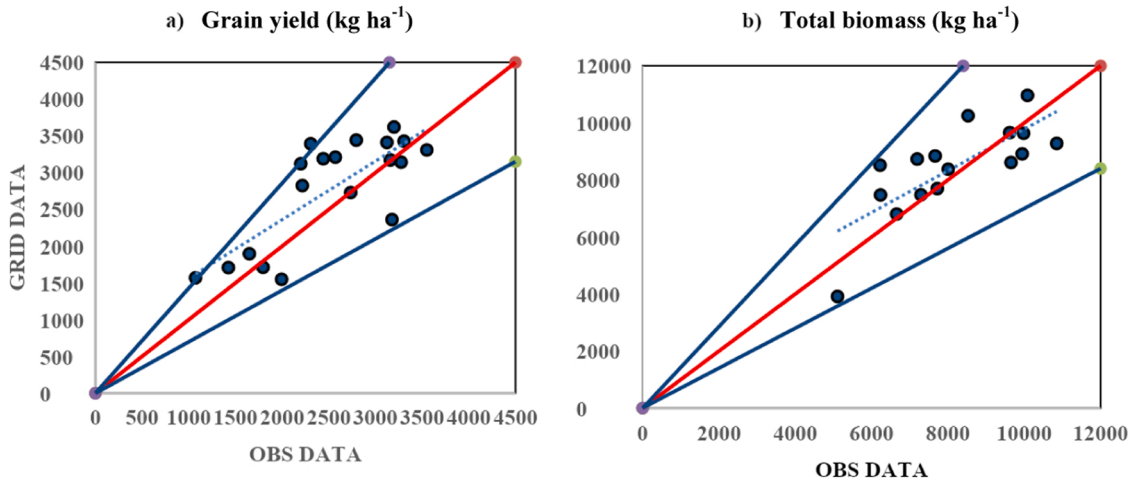


Fig. 2. a, b. Comparison of (a) grain yield and (b) total biomass simulation output from the APSIM model with observed weather data versus running APSIM with the synthetic data of AgMERRA (from Fig. 1a, b). Middle (red) lines represent a 1:1 line and the other (blue) lines denote a 30% divergence percentile of the 1:1 line. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

rainfall averages were still in good agreement for AgMERRA-NASA and NASA-POWER datasets but the statistical metrics were notably lower for the MarkSIM data (Table 3). Consequently, the statistical metrics describing the relation between agronomic parameters simulated with the observed meteorological datasets and the three tested sources of gridded meteorological information were considerably better for AgMERRA-NASA (Table 3; visualized in Fig. 1a, b, Fig. 2a, b). Based on these results, gridded AgMERRA-NASA data was used to expand the spatio-temporal dimensions of simulations. Consequently, 311 gridded meteorological records (each grid size 0.5°x 0.5° encompassing 31 years of weather records (1980–2010)) were used to cover the major rabi sorghum production tract in India.

3.2. Genotype × Management runs across the grid; main characteristics of the processes, generated data and maps

The APSIM sorghum model was run across the Indian rabi sorghum production tract (311 grid items) producing a total of 4,299,264 simulations. Computation took approximately 14 days including several downtime periods and generated 14.6 TB of output data. APSIM is natively set to generate the data in raw text format. Therefore, specifically for our study, follow-up processing was necessary to extract and parse the relevant pieces of information. This text information was transformed into.csv file format to ease the calculations required for the study (a program was written in C# language to select only the relevant data. All this data is available at DOI:10.5281/zenodo.5256068 (https://zenodo.org/record/5256068#. YSYy_S0Rpf0). Statistical

Table 4a
The overview of crop production parameters (grain, stover and total biomass yield [kg ha^{-1}]) and parameters linked to crop stability (proportion of seasons with grain yield failure and standard deviation in total biomass production) averaged across "10 best" simulations attaining highest scenario-weighting index within each scenario and for each of the 3 soils. These parameters were evaluated for site-specific optimal G combinations with optimized M practices (Table 4a) and Maldandi-specific G parameters with optimized M practices (M35–1; Table 4b).

Scenario	Soil	Grain yield [kg ha^{-1}]	Stover yield [kg ha^{-1}]	Proportion of seasons with grain yield failure [%]	Deviation in total biomass production [kg ha^{-1}]
Production	All soils	2690	4911	16	986
	Shallow	2166	4185	35	1146
	Medium	2819	5279	14	1104
	Deep	3085	5270	0	707
Stability	All soils	2455	4318	0	551
	Shallow	2032	3459	1	556
	Medium	2718	4667	0	644
	Deep	2615	4827	0	452

Table 4b
The overview of crop production parameters (grain, stover and total biomass yield [kg ha^{-1}]) and parameters linked to crop stability (proportion of seasons with grain yield failure and standard deviation in total biomass production) averaged across "10 best" simulations attaining highest scenario-weighting index within each scenario and for each of the 3 soils. These parameters were evaluated for site-specific optimal G combinations with optimized M practices (Table 4a) and Maldandi-specific G parameters with optimized M practices (M35–1; Table 4b).

Scenario	Soil	Grain yield [kg ha^{-1}]	Stover yield [kg ha^{-1}]	Proportion of seasons with grain yield failure [%]	Deviation in total biomass production [kg ha^{-1}]
Production	All soils	2298	4607	22	830
	Shallow	1753	3927	50	982
	Medium	2445	4895	14	914
	Deep	2695	5000	0	595
Stability	All soils	2136	4217	5	570
	Shallow	1649	3479	12	638
	Medium	2368	4543	2	644
	Deep	2391	4630	0	429

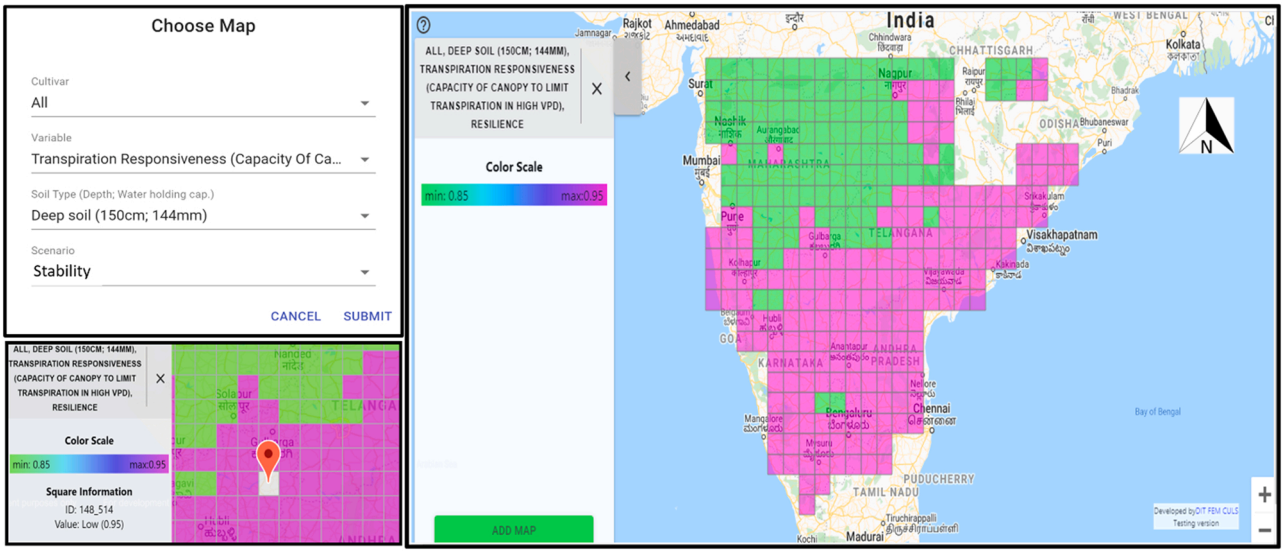


Fig. 3. The visualization of the APSIM simulations outputs via the web application (https://ls40.pef.czu.cz/maps). Each of the panels shows the target post-rainy sorghum growing region in the peninsular part of the Indian sub-continent. The coloured grids (0.5°x 0.5°) signify the geographical variation in management (M) practices and crop characters (G) expected to contribute to the improvement of post-rainy sorghum production/ stability. The user can choose to visualize the grain and stover yield (under "variable") potentially achievable for the currently grown maldandi crop type (by choosing "M 35–1″) or optimized cultivar ("All") for a particular soil type ("Deep"/"Medium"/"Shallow") and scenario ("Production"/"Stability"). Furthermore, users can visualize which level of M (density, sowing window, fertilization) and G (vigour, crop duration, transpiration responsiveness) contributes towards optimal production for the particular cultivar, soil, and scenario and in which region. The actual level of the chosen variable can be visualized by clicking on the grid of interest. The tool has a feature to visualize two maps at the time (green box "Add map") for comparison. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

analysis was done using the MS Excel environment in combination with basic tools provided by excel and specialized macros written in Visual Basic specifically for this task.

3.3. Production gains achievable by optimizing crop management and cultivar choice

Table 3a, b summarize the agronomic performance indicators of the top 10 best-simulated scenarios (i.e. attaining the highest index under production and stability scenario) for optimal cultivar (i.e. optimal combination of G-factors for each environment (grid Table 4a) and M35–1 (i.e. G-factors specific for the M35–1 maldandi crop type; Table 4b) across and within each of the tested soils and across all tested geographies (i.e. simulation units; environments represented by grids). As expected, crop production was predicted to be higher on deeper soils for optimal cultivars and M35–1. Furthermore, the simulations revealed that the site-specific optimization of the cultivar is expected to enhance the grain yields by around 10% (~350 kg ha^{-1}) and stover yields around 5% (~200 kg ha^{-1}). The optimal cultivar was further expected to minimize the proportion of years with grain yield failure compared to M35–1 across the tested conditions. On the other hand, site-specific cultivar optimization was predicted to cause more fluctuations in production across the years (higher biomass deviation indicating lower system stability) compared to M35–1 in the production scenario (compare Table 4a, b).

3.4. Identification of "homogeneous system units" based on the geographical distribution of optimal Genotype × management

The Genotype × Management interventions leading to maximum index of system production and stability are summarized in Table 4a (for optimal G combinations, i.e. optimal cultivars) and 4b (for the G combination specifying the maldandi M 35–1 cultivar). The spatial variability in the optimum G and M intervention can be found in Suppl. Fig. 3a (for optimal cultivars in certain production scenarios), Suppl. Fig. 3b (for the optimal cultivar in the "stability" scenario), and could be further geo-spatially explored using the web application (Fig. 2). Here it was apparent that the optimum Genotype × Management would be location and soil specific. Nevertheless, from Suppl. Fig. 3a, b, and the web application (Fig. 3), we could visually observe distinct geographical North-West to South-East patterns that changed with the soil depths and scenarios. Generally, there was a trend favoring interventions with high doses of N-fertilizer and earlier planting windows (except in the North-West regions, which were predicted to benefit from later sowing windows). In the most stringent scenarios (i.e. stability, shallow soils, North-West geographies), the optimal M tended to favor combinations with lower planting densities. Across all of the investigated geographies and scenarios, the optimal cultivars most frequently involved combinations of short-duration and high vigor characters. The North-Western regions would specifically benefit from the introduction of crops with tight canopy transpiration control (transpiration responsiveness 0.85, Suppl. Fig. 3a, b).

Similarly, we visualized the geographical distribution of optimum management practices for the maldandi crop type (M35–1; Suppl. Fig. 4, the web application). Output highlighted that there was an apparent North-West to South-East gradient in optimal M combinations. Most of the optimal M combinations generally favoured much lower planting densities and lower fertilizer inputs compared to the optimized crop type (i.e. compared to site-specific G combinations, compare optimal M from Suppl. Fig. 3a, b with Suppl. Fig. 4). Also, similarly to the site-specific optimized crop types analysis, the North-West part of the investigated region would benefit from later planting windows more than the rest of the production region (compare Suppl. Fig. 3a, b with Suppl. Fig. 4).

Using PCA, the outputs from each of the simulation units ("grids") were clustered into the four geographical units based on their similarities in production/ stability system characteristics as well as optimized combinations of G and M parameters. Such analysis, in principle, seperated the grids into the geographical regions with the similarities in system response to G and M interventions (Fig. 4). When visualized, these four geographical units formed the pattern of concentric layers around the "core" of the North-Western part of the sorghum production area (HSU_1; light blue, Fig. 3). The main system characteristics along

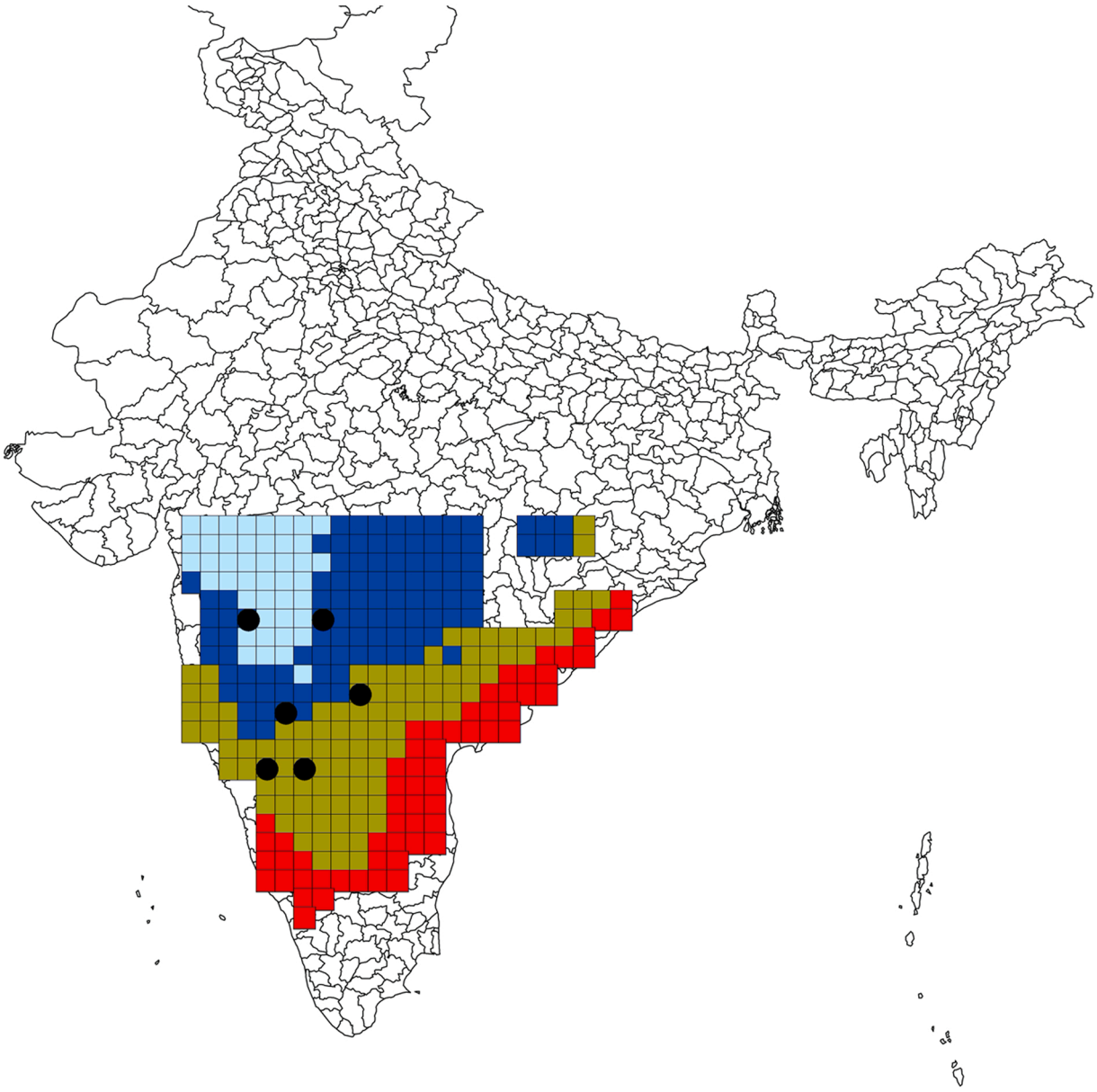


Fig. 4. The map of India over-layed with the four identified "homogeneous system units" (HSU, highlighted in different colors; summary system characteristics of the HSU clusters are in Table 5). Each of the grids (0.5°x 0.5°) is expected to respond more homogeneously to particular GxM interventions than the remaining grids within one HSU compared to the grids from a different HSU. The discrete black circles on the map highlight the current post-rainy sorghum testing sites within an All India Coordinated Research Project (AICRP; http://www.millets.res.in/aicrp.php).

with the optimal Genotype × Management combinations within these geographical units were summarized in Table 5. Generally, Table 5 illustrates that with the increasing distance from this core (HSU_1) the production potential and system stability would increase in the direction towards HSU_4 (stover and grain yield production as well as system stability indicators). This was also well-reflected in optimmalized G and M parameters within each HSU; i.e. planting density, crop duration, and plant vigor. A parameter indicating plant responsiveness to VPD also increased with increasing distance from the production core (from HSU_1 towards HSU_4). Across all HSUs, most of the optimal Genotype × Management combinations leaned towards the early sowing windows and stable fertilizer doses ~ 70–70 kg ha^{-1} (basal dose - top-dressing). Table 5.

4. Discussion

4.1. Deployment of high-performance computations for effective APSIM-runs

The technical details and challenges involved in this computational exercise were described in Jarolímek et al. (2019). In principle, we used a cluster of seven high-performance computers and tested several options to distribute the computations across the cluster effectively. This involved separation of the simulation set-ups into batches, which were consequently scheduled and run manually. This exercise enabled us to design the structure of the software tools for the computational facility used. This allows for further process automation should similar exercises be required with the cluster in the future. Alternative software resources available in the public domain, specialized for computational distribution such as HTCondor (https://research.cs.wisc.edu/htcondor/) might be also considered. Another option to be tested would include running

Table 5
The summary statistics of the system production and stability indicators resulting from optimized management (M) and genetic crop characters (G) within the identified Homogeneous System Unit (HSU) clusters as spatially defined in Fig. 4.

Cluster number (color in Figure 4)	HSU_1 (light blue)	HSU_2 (dark blue)	HSU_3 (green)	HSU_4 (red)
(M) Sowing window	16 - 23sep	23 - 30sep	16 - 23sep	16 - 23sep
(M) Plant density (plant m^{-2})	12.54	12.58	13.18	13.66
(M) Nitrogen fertilization (kg ha^{-1} Urea)	74-74	72.8-72.8	71-71	71.3-71.3
(G) Crop duration (tt_endjuv_to_ini)	181	178	197	220
(G) Rate of canopy growth, vigor (power coefficient for TPLA max)	2.88	2.882	2.926	2.96
(G) Transpiration responsiveness (Capacity of canopy to limit transpiration in high VPD)	0.9065	0.893	0.941	0.9485
Grain yield [kg·ha^{-1}]	2231	2182	2675	3308
Stover yield [kg·ha^{-1}]	4026	3982	4692	5986
Proportion of seasons with grain yield failure [%]	24	12	2	1
Deviation in total biomass production [kg ha^{-1}]	767	747	740	854

the simulations using commercial cloud services, such as Azure or Amazon Web Services.

It is important to note that the new version of APSIM is being developed (NextGen APSIM; https://apsimnextgeneration.netlify.app). NextGen APSIM modules are already capable of running multiple simulations much more efficiently compared to classic APSIM software (https://www.apsim.info). However, transiting this massive work to the NextGen APSIM-based framework would require rigorous cross-validation of the model functions that are key for the presented study as well as the model set-up. Nonetheless, the necessity of transitioning to the NextGen system must be seriously considered, especially in the context of rising demand for a similar types of analysis (e.g. in the context of CGIAR crop improvement program modernizations; https://excellenceinbreeding.org/; https://bigdata.cgiar.org/event/webinar-target-population-of-environments-tpe-beyond-helping-make-better-crop-improvement-practice/).

4.2. Cultivar x Management effects and their optimal combinations for higher and stable sorghum production

APSIM has been used to model management and genetic interventions for various cropping systems (e.g. wheat, chickpea, maize, potato; Chenu et al., 2011, 2013; Chapman et al., 2000a, 2000b, 2000c; Lobell et al., 2015, Chauhan et al., 2008; Chauhan et al., 2013; Beah et al., 2021; De Silva et al., 2021; Ojeda et al., 2020, 2021). Currently, APSIM is a base for several commercial applications used by different stakeholders such as farmers or breeders (YieldProphet ® (Yield Prophet), WhopperCropper (The Regional Institute - J.Managing Climate Variability - Crops), CropARM (Decision support tools and modeling | Tasmanian Institute of Agriculture (utas.edu.au)). In the case of sorghum, the APSIM model has been used to evaluate the production regions assessed in this study and the effect of management and crop genetic interventions (Ravi Kumar et al., 2009 - maldandi sorghum parameterization; Kholová et al., 2013 - rabi sorghum systems characterization; Kholová et al., 2014 - evaluation of genetic interventions; Dimes and Revanuru, 2004 - nitrogen application; Turner and Rao, 2013 - effect of planting density & duration of cultivars; Akinseye et al., 2020 - effect of sowing dates).

In the context of this study, where accessing observed meteorological information is problematic, we transited the entire modeling framework into the AgMERRA-NASA-based gridded framework (Ruane et al., 2015) to allow for more geographically precise and spatially balanced analysis. Out of the tested options (NASA-POWER, MarkSIM, AgMERRA-NASA), the AgMERRA-NASA-based set-up (31 years of daily weather records) was found sufficient to represent the historical weather variability across the rabi-sorghum production region. The sorghum simulation outputs were, additionally, cross-compared with the ranges of agronomic parameters reported from multi-year, multi-location agronomic trials conducted in post-rainy seasons within the Indian national sorghum evaluation network (AICRIP project; http://www.millets.res.in/aicrp.php). Information from these trials could not be closely compared with our simulations, as, for example, the tested genotypes are usually Maldandi types but not exactly M35–1 and the management practices in these trials usually involve "life-saving irrigation'' or other agronomic practices that are rarely sufficiently documented to allow strict comparison. However, the crop production ranges and responses to the management practices (fertilization, planting density) reported from AICRIP sorghum testing trials (http://www.millets.res.in/aicsip13.php) were in reasonable agreement with the model outputs. Therefore, the AICRIP trials would be an interesting data source to further emulate the presented modeling framework.

The modelling approach is one of the very few options that allows us to disentangle, quantify and optimize the effects of cultivar and crop management interventions (Jeuffroy et al., 2014; Lecomte et al., 2010; Chapman et al., 2002). This information is critical to guide any efforts for agri-system improvement and breeding. The vast amount of information generated in this simulation exercise (14.6 TB) allowed us just this - i.e. to separate and quantify the effects of particular cultivar X management interventions on the important characteristics of the post-rainy sorghum cropping system. We found that the magnitude of any GxM intervention effect depended on the soil properties. Our findings further indicated that, in general, site-specific fine-tuning the crop and crop agronomic practices within the breeder- and farmer- relevant ranges would have an important effect on crop production/ stability in

the rain-fed rabi sorghum belt of India. Despite the fact that we did not investigate the specific factors leading to production constraints, as in Kholová et al. (2013), our study emphasizes that crop products (accompanied by well-designed agronomic practices) for rain-fed systems have to be carefully tailored to the variability of production environments. We showed that the commonly occurring genetic variability in sorghum species is sufficient to enhance rabi-systems (i.e. duration, vigor and canopy conductance; Kholová et al., 2014; Bodner et al., 2015). Significant improvements can be achieved by fitting existing cultivars with suitable management to the particular context of the rabi-production system or by generating new crop products using the resources generated in this study as a guideline.

For this purpose, we provide the data generated along with the tool designed for further exploration by any stakeholders for free (e.g. post-rainy sorghum breeding programs, policy-makers or on-ground farmer advisory services). This tool is now ready to explore the spatial distributions of optimal management practices for the Maldandi crop type (M35–1) as well as the production potentially achievable with other crop types (details in Fig. 2). As the map shows, these are dependent on the location, soil type, and scenarios considered. The estimated site-specific genetic enhancement of Maldandi has the potential to increase ~10% (± 7%) grain and 5% (± 2%) stover production across all locations. The same intervention would stabilize grain production across seasons. We envision that a similar type of IT tool, could complement the recommendation packages (e.g. https://www.millets.res.in/farmer/Recommended_packages_of_practices_Rabi_sorghum.pdf; periodically released by the Indian government) in order to enable the site-specific recommendations through on-ground agencies who can operate the simple interactive map. This framework should also serve as a base upon which further enhancements required by the different users can be built, such as crop improvement teams, policy makers, and farmer advisory services, which would enable its broader deployment and impact. Similar principles have been reflected in CropArm (http://www.armonline.com.au/#/wc), an APSIM-based tool developed to support decisions in Australian farming systems. To our knowledge, the presented online tool is the first of this kind that is able to support effective system design for climate risk-prone agri-systems in developing countries. The tool's development demonstrates a diligent approach on how to condense the vast amount of data typically produced by crop modelers into digestible information for non-experts. This approach will be further expanded for other regions and crops, evolved and sensitized to the indigenous agri-system requirements and contexts.

4.3. Identification of homogeneous system units (HSUs)

The primary beneficiaries in mind while developing the study were crop breeders. Breeders typically require crop modelers to identify geographies for which a particular crop product and agronomic management can be developed and where it is best tested (e.g. BPAT review; https://plantbreedingassessment.org/bpat-project/bpatmission/) (Kholová et al., 2021). This kind of analysis required the further stratification of the information held by the generated dataset. Firstly, we reduced the data dimensionality using principal component analysis (PCA) and consequently deployed the clustering approach to form geo-spatially distinct classes - the "homogeneous system units" (HSUs). This allowed us to separate the tested geographies (grids) into four HSUs based on similarities in optimum production characteristics and system responses to GxM interventions. Such novel assessments considerably extended the previously used approaches (environmental characterization/ target population of environments) e.g. in Kholová et al. (2013), Hajjarpoor et al., (2018, 2021), Chauhan et al. (2013), Chenu et al. (2011), and Chapman et al. (2000a, 2000b). The "HSU" analysis allowed us to differentiate the geographies with maximum similarities within a geographic group and dissimilarities between the groups not only based on the modeled interactions of crop, environment, and management but also on system responsiveness to the Genotype × Management interventions. We suggest that such geospatial classification enable breeding programs to, for instance, optimize the distribution of the multi-location testing sites, improve the statistical treatment of the data generated in different geographies and precisely design and target crop product development efforts. For instance, in typical crop improvement programs, the crop is tested with very limited management options or the management is adapted only "*post-mortem*" when the genotype is already fixed. These circumstances inevitably stagnate the crop production improvement in these complex systems. To overcome this gap, we provided a unique tool that allows for the simultaneous prediction of optimal crop management along with the suitable crop cultivar, which is otherwise impossible. We conclude that the presented APSIM-powered framework enables the improvement of breeding targets, empowering breeding programs to design region-specific Genotype × Management options *ex-ante* that could significantly accelerate efforts to improve productivity/ resilience of dry-season sorghum cultivation.

4.4. Possible limitations of the study and continuous improvement of the framework

Models are reflections of our imperfect knowledge, which is why it's important to acknowledge the assumptions and other possible limitations of the acquired modeling outputs. In our case, we need to mention the use of a gridded data source (AgMERRA-NASA) instead of the actual meteorological observations that may have been preferable. Although meteorological information is becoming more available as standard across the globe, many countries, like India, are still not well covered with accessible, high-quality, and up-to-date information. Since our study required homogeneous coverage of key regions, we chose to use NASA-generated information that has supported modeling of agri-systems similar to ours (Table 3, Fig. 1a, b, 2 a, b). In the ideal case, we would have had detailed agronomic evaluations of sorghum production across locations to cross-validate the simulation set-up responsiveness to major system limitations (e.g. agronomic practices). As mentioned above, these datasets are very rare in the local context and their generation is cost- and time-intensive. While we work on such dataset generation, we do have numerous studies and even commercial products based on the APSIM sorghum module responsiveness to a range of M and G contexts (e.g. Akinseye et al., 2020; Dimes and Revanuru, 2004; Turner and Rao, 2013).

In future, we plan to use this sorghum modeling framework to support broader socio-economic modeling studies. Here we presented our attempt to demonstrate the generic approach, i.e. the scenario weighting index which is based on an educated guess founded on literature surveys (e.g. Blümmel and Rao, 2006, Tesfaye, 1998, Ravi et al., 2003, Reddy et al., 2005, Rao et al., 2017, Blümmel et al., 2015) and discussions with experts. Such estimates are to be improved as we progress in understanding and interlinking this work with socio-economic studies of the target population of stakeholders in particular regions. The understanding of community demands or particular user cases should, in principle, guide further simulation exercises and, among others, the resolution of simulations, the GxM scenarios tested, further assumptions made, tool co-creation, and design.

5. Conclusion

The presented work aims to translate current advances in crop modeling science into a quantitative understanding of crop production systems for the key pool of beneficiaries (e.g. breeding programs, farmer advisories, decision-makers, etc.) via a simple visualization tool. The presented framework simplifies the complex modeling data (~0.5 million simulations, ~14 TB) and utilizes them to understand the context-dependencies of post-rainy sorghum agricultural systems even by a community of non-experts. Although numerous on-line tools have been developed, primarily to provide advice to large-scale agricultural producers in developed countries, these might not fit the requirements

of stakeholders in countries like India. We argue that to enable an effective understanding of the diversity of small-scale agriculture systems, the tool has to be tailored: (i) to be easily accessible (possibly free of cost) (ii) simple enough and sufficiently interactive, and (iii) encompass a valid range of the farming scenarios. We have developed draft tool and analytics with the example of post-rainy sorghum production systems in India and will continue the customization and evolution of the presented tool to serve the particular needs of various end-users. A similar approach can be now adapted to other agricultural production systems, especially those that are small-scale and low input.

CRediT authorship contribution statement

Swarna Ronanki: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Jan Pavlik:** Data curation, Investigation, Resources, Software, Writing – original draft, Writing – review & editing. **Jan Masner:** Resources, Software, Writing – original draft, Writing – review & editing. **Jan Jarolimek:** Funding acquisition, Project administration, Resources, Supervision. **Michal Stoces:** Software. **Degala Subhash:** Data curation, Visualization. **Harvinder S. Talwar:** Resources. **Vilas A. Tonapi:** Funding acquisition, Resources. **Mallayee Srikanth:** Visualization. **Jana Kholová:** Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Rekha Badam:** Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The results and knowledge included in this article have been obtained withsupport from the following grants; Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2019B0009 – Life Sciences 4.0, the CGIAR Research Program on Grain Legumes and Dryland Cereals (GLDC) and a mini-grant from the CGIAR Community of Practice on Modelling (https://bigdata.cgiar.org/communities-of-practice/crop-modeling/), and the core funding of ICAR- Indian Institute of Millets Research. A Global Challenges Research Fund project - Transforming India's Green Revolution by Research and Empowerment for Sustainable Food Supplies (TIGR2ESS, BB/P02797/01). Authors are grateful to Dr. Amir Hajjarpoor (UMR DIADE, Université de Montpellier, Institut de Recherche pour le Développement) for compilation and analysis of meteorological information from several sources. Authors further acknowledge the contribution of Dr. Sunita Choudhary (International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Hyderabad) for resource mobilization as well as dissemination and promotion of scientific findings reported in this work to the key stakeholders.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fcr.2021.108422.

References

Akinseye, F.M., Ajeigbe, H.A., Traore, P.C., Agele, S.O., Zemadim, B., Whitbread, A., 2020. Improving sorghum productivity under changing climatic conditions: a modelling approach. Field Crops Res. 246, 107685.

Ambadi, A., Krishnamurty, D., Rao, S., Desai, B.K., Ravi, M.V., Shubha, S., 2018. Yield potential and economics of rabi sorghum (Sorghum bicolor L.) as influenced by different crop residues and green biomass composts. J. Appl. Nat. Sci. 10 (1), 128–132.

Beah, A., Kamara, A.Y., Jibrin, J.M., Akinseye, F.M., Tofa, A.I., Ademulegun, T.D., 2021. Simulation of the optimum planting windows for early and intermediate-maturing maize varieties in the Nigerian savannas Using the APSIM model. Front. Sustain. Food Syst. 5, 624886 https://doi.org/10.3389/fsufs.

Blümmel, M. , Rao, P.P. 2006. Economic value of sorghum stover traded as fodder for urban and peri-urban dairy production in Hyderabad, India. International Sorghum and Millets Newsletter (47):97–100.

Blümmel, M., Deshpande, S., Kholova, J., Vadez, V., 2015. Introgression of staygreen QLT's for concomitant improvement of food and fodder traits in Sorghum bicolor. Field Crops Research. Elsevier BV, pp. 228–237. https://doi.org/10.1016/j.fcr.2015.06.005.

Bodner, G., Alireza, N., Hans-Peter, Kaul, 2015. Management of crop water under drought: a review. Agron. Sustain. Dev. 35 (2), 401–442.

Chapman, S.C., Cooper, M., Butler, D.G., Henzell, R.G., 2000a. Genotype by environment interactions affecting grain sorghum. I. Characteristics that confound interpretation of hybrid yield. Aust. J. Agric. Res. 51, 197–207.

Chapman, S.C., Cooper, M., Hammer, G.L., Butler, D.G., 2000b. Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust. J. Agric. Res. 51 (2), 209–221.

Chapman, S.C., Cooper, M., Butler, D.G., Hammer, G.L., 2000c. Genotype by environment interactions affecting grain sorghum. III. Temporal sequences and spatial patterns in the target population of environments. Aust. J. Agric. Res. 51, 223–233.

Chapman, S.C., Cooper, M., Hammer, G.L., 2002. Using crop simulation to generate genotype by environment interaction effects for sorghum in water-limited environments. Aust. J. Agric. Res. 53, 379–389. https://doi.org/10.1071/AR01070.

Chauhan, Y., Wright, G., Rachaputi, N., McCosker, K., 2008. Identifying chickpea homoclimes using the APSIM chickpea model. Aust. J. Agric. Res. 59 (3), 260–269.

Chauhan, Y.S., Solomon, K.F., Rodriguez, D., 2013. Characterization of north-eastern Australian environments using APSIM for increasing rainfed maize production. Field Crops Res. 144, 245–255.

Chenu, K., Cooper, M., Hammer, G.L., Mathews, K.L., Dreccer, M.F., Chapman, S.C., 2011. Environment characterization as an aid to wheat improvement: interpreting genotype-environment interactions by modelling water-deficit patterns in North-Eastern Australia. J. Exp. Bot. 62 (6), 1743–1755.

Chenu, K., Deihimfard, R., Chapman, S.C., 2013. Largescale characterization of drought pattern: a continent-wide modelling approach applied to the Australian wheatbelt - spatial and temporal trends. New Phytol. 198, 801–820.

Dayakar Rao B., Shashidhar Reddy Ch, Nirmal Reddy K., Ratnavathi CV, Shyamprasad G. Seetharama N. , 2009. Package of Practices for Improved Rabi Sorghum Cultivation (English), National Agricultural Innovation Project (NAIP), ITC, Secunderabad, 500 003 and NRCS, Rajendranagar, 500 030, Andhra Pradesh, India. Technical Bulletin number NAIP/ITC/NRCSTECH/4/2009, 14pp.

Descamps, C., Quinet, M., Baijot, A., Jacquemart, A.L., 2018. Temperature and water stress affect plant–pollinator interactions in Borago officinalis (Boraginaceae). Ecol. Evol. 8 (6), 3443–3456.

De Silva, S.H.N.P., Takahashi, T., Okada, K., 2021. Evaluation of APSIM-wheat to simulate the response of yield and grain protein content to nitrogen application on an Andosol in Japan. Plant Prod. Sci. 1–12.

Dimes, J.P., Revanuru, S., 2004. Evaluation of APSIM to simulate plant growth response to applications of organic and inorganic N and P on an Alfisol and Vertisol in India. In *Aciar Proceedings* (pp. 118–125). ACIAR; 1998.

Dingkuhn, M., Soulié, J.C., Lafarge, T., 2011. Samara V2: A cereal crop model to study G x E x M interaction and phenotypic plasticity, and explore ideotypes. In: AgMIP Rice International Workshop, 28–30 August, Beijing, China.

Habyarimana, E., Piccard, I., Catellani, M., De Franceschi, P., Dall'Agata, M., 2019. Towards predictive modeling of sorghum biomass yields using fraction of absorbed photosynthetically active radiation derived from sentinel-2 satellite imagery and supervised machine learning techniques. Agronomy 9 (4), 203.

Hajjarpoor, A., Vadez, V., Soltani, A., Gaur, P., Whitbread, A., Babu, D.S., Gumma, M.K., Diancoumba, M., Kholová, J., 2018. Characterization of the main chickpea cropping systems in India using a yield gap analysis approach. Field Crops Res. 223, 93–104.

Hammer, G.L., van Oosterom, E., McLean, G., Chapman, S.C., Broad, I., Harland, P., Muchow, R.C., 2010. Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. J. Exp. Bot. 61 (8), 2185–2202.

Hochman, Z., Van Rees, H., Carberry, P.S., Hunt, J.R., McCown, R.L., Gartmann, A., Holzworth, D., Van Rees, S., Dalgliesh, N.P., Long, W., Peake, A.S., 2009. Re-inventing model-based decision support with Australian dryland farmers. 4. Yield Prophet® helps farmers monitor and manage crops in a variable climate. Crop Pasture Sci. 60 (11), 1057–1070.

Holzworth, D.P., Snow, V., Janssen, S., Athanasiadis, I.N., Donatelli, M., Hoogenboom, G., White, J.W., Thorburn, P., 2015. Agricultural production systems modelling and software: current status and future prospects. Environ. Model. Softw. 72, 276–286.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., Keating, B.A., 2014. APSIM - evolution

towards a new generation of agricultural systems simulation. Environ. Model. Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009.
Jarolímek, J., Pavlík, J., Kholová, J., Ronanki, S., 2019. Data pre-processing for agricultural simulations. AGRIS Line Pap. Econ. Inform. 11 (1), 49–53.
Jeuffroy, M.H., Casadebaig, P., Debaeke, P., Loyce, C., Meynard, J.M., 2014. Agronomic model uses to predict cultivar performance in various environments and cropping systems. A review. Agron. Sustain. Dev. 34 (1), 121–137.
Jirali, D.I., Biradar, B.D., Rao, S.S., 2010. Performance of Rabi sorghum genotypes under receding soil moisture conditions in different soil types. Karnataka J. Agric. Sci. 20 (3).
Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., 2017. Toward a new generation of agricultural system data, models, and knowledge products: state of agricultural systems science. Agric. Syst. 155, 269–288.
Jones, P.G., Thornton, P.K., Díaz, W., Wilkens, P.W., Jones, A.L., 2002. MarkSim: A computer tool that generates simulated weather data for crop modeling and risk Assessment: version 1 [CD-ROM].
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18 (3–4), 235–265.
Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288.
Kholová, J., McLean, G., Vadez, V., Craufurd, P., Hammer, G.L., 2013. Drought stresscharacterization of post-rainy season (*rabi*) sorghum in India. Field Crops Res. 141, 38–46.
Kholová, J., Tharanya, M., Sivasakthi, K., Srikanth, M., Rekha, B., Hammer, G.L., McLean, G., Deshpande, S., Hash, C.T., Craufurd, P., Vadez, V., 2014. Modelling the effect of plantwater use traits on yield and stay-green expression in sorghum. Funct. Plant. 41, 1019. https://doi.org/10.1071/FP13355.
Kholová, J., Urban, M.O., Cock, J., Arcos, J., Arnaud, E., Aytekin, D., Azevedo, V., Barnes, A.P., Ceccarelli, S., Chavarriaga, P., Cobb, J.N., Connor, D., Cooper, M., Craufurd, P., Debouck, D., Fungo, R., Grando, S., Hammer, G.L., Jara, C.E., Xu, Y., 2021. In pursuit of a better world: crop improvement and the CGIAR. J. Exp. Botany 72 (14), 5158–5179. https://doi.org/10.1093/jxb/erab226.
Kumar, S.R., Bhat, P., Rajappa, P.V., 2017. Management strategy to improve input use efficiency and enhance sorghum productivity per stored rain drop in vertisols during rabi season. Curr. Sci. 30, 304–307.
Lecomte, C., Prost, L., Cerf, M., Meynard, J.M., 2010. Basis for designing a tool to evaluate new cultivars. Agron. Sustain Dev. 30, 667–677. https://doi.org/10.1051/agro/2009042.
Lobell, D.B., Hammer, G.L., Chenu, K., Zheng, B., McLean, G., Zheng, B., Chapman, S.C., 2015. The shifting influence of drought and heat stress for crops in Northeast Australia. Glob. Change Biol. 21, 4115–4127.
Olson, S. , 2012. Designing an Ideal Energy Crop: The Case for Sorghum bicolor (Doctoral dissertation).
Ojeda, J.J., Rezaei, E.E., Remenyi, T.A., Webb, M.A., Webber, H.A., Kamali, B., Harris, R. M., Brown, J.N., Kidd, D.B., Mohammed, C.L., Siebert, S., 2020. Effects of soil-and climate data aggregation on simulated potato yield and irrigation water requirement. Sci. Total Environ. 710, 135589.
Ojeda, J.J., Rezaei, E.E., Remenyi, T.A., Webber, H.A., Siebert, S., Meinke, H., Webb, M. A., Kamali, B., Harris, R.M., Kidd, D.B., Mohammed, C.L., 2021. Implications of data aggregation method on crop model outputs–the case of irrigated potato systems in Tasmania, Australia. Eur. J. Agron. 126, 126276.
Pradhan, P., Fischer, G., van Velthuizen, H., Reusser, D.E., Kropp, J.P., 2015. Closing yield gaps: how sustainable can we be? PLoS One 10, e0129487.
Rao, Benhur ,Mukherjee, Deep Narayan, Devi, Y., Tonapi, Vilas, 2017. An Economic Analysis of Improved Rabi Sorghum Cultivars in Rainfed Situation of Maharashtra, India. 4. 7–15.
Ravi, D., Vishala, A.D., Nayaker, N.Y., Seetharama, N. and Blümmel, M. 2003. Grain yield and stover fodder value relations in rabi sorghum. International Sorghum and Millets Newsletter. 44: 28–31.
Ravi Kumar, S., Graeme, L., Hammer, Ian Broad, Peter, Harland, Greg, Mc.Lean, 2009. Modelling environmental effects on phenology and canopy development of diverse sorghum genotypes. Field Crops Res. 111, 157–165.
Reddy, K.G. and Michael, B. and Rao, P.P. and Reddy, B.V. S. and Ramesh, S. and Reddy, K.M. V.P. (2005) Evaluation of farmer-grown improved sorghum cultivars for stover quality traits. International Sorghum and Millets Newsletter, 46. pp. 86–89.
Richter, M.E., Phelan, D., Harrison, M., Dean, G., Pengilley, G., Hinton, S., Mohammed, C., 2017. CropARM: An agronomic support tool assisting Tasmanian farmers for rainfed and irrigated wheat production. In *" Doing More with Less", Proceedings of the 18th Australian Agronomy Conference 2017, Ballarat, Victoria, Australia, 24–28 September 2017* (pp. 1–4). Australian Society of Agronomy Inc.
Rienecker, M.M., Suarez, M.J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M.G., Schubert, S.D., Takacs, L., Kim, G.K., Bloom, S., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. J. Clim. 24 (14), 3624–3648.
Ronanki, S., Kholová, J., Talwar, H.S. , 2018. Simulation of post rainy sorghum yield response N fertilization in India. ⟨https://21centurysorghum.com/wp-content/uploads/2018/06/09h40-Ronanki-Hall-B-Thurs1.pdf⟩.
Rooney, W.L., Blumenthal, J., Bean, B., Mullet, J.E., 2007. Designing sorghum as a dedicated bioenergy feedstock. Biofuels Bioprod. Bioref. 1 (2), 147–157.
Ruane, A.C., Winter, J.M., McDermid, S.P., Hudson, N.I., 2015. AgMIP climate datasets and scenarios for integrated assessment. In: Rosenzweig, C., Hillel, D. (Eds.), Handbook of Climate Change and Agroecosystems: The Agricultural Model Intercomparison and Improvement Project (AgMIP) Integrated Crop and Economic Assessments. ICP Series on Climate Change Impacts, Adaptation, and Mitigation, Vol. 3. Imperial College Press, pp. 45–78. https://doi.org/10.1142/9781783265640_0003. Part 1.
Soltani, A., Hoogenboom, G., 2003. Minimum data requirements for parameter estimation of stochastic weather generators. Clim. Res. 25 (2), 109–119.
Soltani, A., Sinclair, T.R. 2012. Modelling physiology of crop development, growth and yield. CABI: 322.
Soltani, A., Hajjarpour, A., Vadez, V., 2016. Analysis of chickpea yield gap and water-limited potential yield in Iran. Field Crops Res. 185, 21–30.
Tesfaye A. , 1998. Economics of milk production in and around Hyderabad of Andhra Pradesh. M.Sc. thesis, Acharya NG Ranga Agricultural University, Hyderabad 500 030.
Trivedi, T.P., 2009. Handbook of Agriculture. Directorate of Information and Publications of Agriculture, Indian Council of Agricultural Research, New Delhi, India.
Thornton, P.K., Whitbread, A., Baedeker, T., Cairns, J., Claessens, L., Baethgen, W., Bunn, C., Friedmann, M., Giller, K.E., Herrero, M., Howden, M., Kilcline, K., Nangia, V., Ramirez-Villegas, J., Kumar, S., West, P.C., Keating, B., 2018. A framework for priority-setting in climate smart agriculture research. Agricultural Systems 167, 161–175. https://doi.org/10.1016/j.agsy.2018.09.009.
Turner, N.C., Rao, K.P.C., 2013. Simulation analysis of factors affecting sorghum yield at selected sites in eastern and southern Africa, with emphasis on increasing temperatures. Agric. Syst. 121, 53–62.
Vadez, V., Krishnamurthy, L., Hash, C.T., Upadhyaya, H.D., Borrell, A.K., 2011. Yield, transpiration efficiency, and water-use variations and their interrelationships in the sorghum reference collection. Crop Pasture Sci. 62 (8), 645–655.
Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation of model performance. J. Geophys. Res. 90 (C5), 8995–9005.
Zampieri, M., Weissteiner, C.J., Grizzetti, B., Toreti, A., van den Berg, M., Dentener, F., 2020. Estimating resilience of crop production systems: from theory to practice. Sci. Total Environ. 735, 139378 https://doi.org/10.1016/j.scitotenv.2020.139378.
Chauhan, Y.S., Rachaputi, R.C.N., 2014. Defining agro-ecological regions for field crops in variable target production environments: a case study on mungbean in the northern grains region of Australia. Agric. For. Meteorol. 194, 207–217. https://doi.org/10.1016/j.agrformet.2014.04.007.

6 Závěr

Výsledky dosažené v rámci autorovy vědecké a publikační činnosti dokládají existenci významných překážek automatizace při zpracování velkého objemu dat v rámci jejich životního cyklu. Disertační práce byla konkrétně zaměřena na problematiku přechodů mezi jednotlivými fázemi zpracování a na rozhraní mezi jednotlivými integrovanými softwarovými systémy.

Oblast simulací zemědělské produkce byla zvolena vzhledem k probíhajícímu výzkumu katedry, aby výsledky disertační práce prakticky přispěly a obohatily stávající vědeckou činnost. Na základě analýzy vědecké literatury byl specifikován výchozí scénář běžného používání softwarových nástrojů pro zpracování simulací a byly identifikovány konkrétní nedostatky procesu automatizace zpracování dat. Metodika PlaGroSim byla navržena tak, aby postihla kritické překážky automatizace, přispěla k publikační činnosti zaplněním informační mezery, a zároveň aby její aplikace zefektivnila výpočetní procesy při zpracování dat v současných a budoucích výzkumných projektech.

Pro konkrétní aplikaci navržené metodiky a její experimentální ověření byl vyvinut modulární softwarový nástroj PlaGroSim. Jeho použití výrazně zvýšilo efektivitu vynaloženého úsilí na proces automatizace. Tento software je aktuálně intenzivně využíván pro výzkumnou činnost v rámci Katedry informačních technologií a její spolupráce se zahraničními partnery.

V současné době vznikají další publikace, které se věnují nejnovějšímu výzkumu, pro který již byla navržená metodika plně implementována. Probíhá i vylepšování softwaru PlaGroSim zapojením nových modulů. Je plánováno vylepšit modul SSM tak, aby bylo možné vyvinuté programové nástroje použít pro širší škálu výzkumných projektů. Zvýšení míry automatizace v oblasti simulací zemědělské produkce je velmi kladně vnímáno v rámci stávající expertní vědecké komunity.

7 Seznam použitých zdrojů

Abella, A., Ortiz-de-Urbina-Criado, M., De-Pablos-Heredero, C. (2022) "Criteria for the identification of ineffective open data portals: pretender open data portals" *Profesional de la información*, sv. 31, č. 1. Dostupné na: https://doi.org/10.3145/epi.2022.ene.11

Abdella, Y., Alfredsen, K. (2010) "A GIS toolset for automated processing and analysis of radar precipitation data" *Computers & Geosciences*, sv. 36, č. 4, s. 422-429. Dostupné na: https://doi.org/10.1016/j.cageo.2009.08.008

Abreu, M., Reis, A., Moura, P., Fernando, A.L., Luís, A., Quental, L., Patinha, P. and Gírio, F. (2020) "Evaluation of the Potential of Biomass to Energy in Portugal—Conclusions from the CONVERTE Project," *Energies*, sv. 13, č. 4. Dostupné na: https://doi.org/10.3390/en13040937

Afful-Dadzie, E., Afful-Dadzie, A. (2017) "Open Government Data in Africa: A preference elicitation analysis of media practitioners" *Government Information Quarterly*, sv. 34, č. 2, s. 244-255. Dostupné na: https://doi.org/10.1016/j.giq.2017.02.005

Agbo, B., Qin, Y., Hill, R. (2019) "Research Directions on Big IoT Data Processing using Distributed Ledger Technology: A Position Paper" *4th International Conference on Internet of Things, Big Data and Security, SCITEPRESS - Science and Technology Publications*, s. 385-391. Dostupné na: https://doi.org/10.5220/0007751203850391

Amaral, S., Cesar Lima D'Alge, J. (2009) "Spatial data availability and its implications for sustainable development of the Brazilian Amazon" *Earth Science Informatics*, sv. 2, č. 4, s. 193-203. Dostupné na: https://doi.org/10.1007/s12145-009-0032-9

Anagnostopoulos, T., Zaslavsky, A., Medvedev, A. (2015) "Robust waste collection exploiting cost efficiency of IoT potentiality in Smart Cities" *2015 International Conference on Recent Advances in Internet of Things (RIoT), Singapur, Singapur, 07.04.2015*. Dostupné na: https://doi.org/10.1109/riot.2015.7104901

Bartoněk, D. (2016) "The Possibilities of Big GIS Data Processing on the Desktop Computers" *Lecture Notes in Geoinformation and Cartography*, s. 273-287. Dostupné na: https://doi.org/10.1007/978-3-319-45123-7_20

Bediroğlu, G., Colak, H.E. (2017) "Cloud Gis Based Watershed Management" *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, sv. 42/4, č. W6, s. 31-33. Dostupné na: https://doi.org/10.5194/isprs-archives-xlii-4-w6-31-2017

Bellman, C.J., Pupedis, G. (2016) "Lost in The Cloud - New Challenges For Teaching GIS" *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, sv. 41, č. B6, s. 25-29. Dostupné na: https://doi.org/10.5194/isprsarchives-xli-b6-25-2016

Chen, X.Y., Guo, J., Geng, J.H., Li, C.H. (2017) "Performance of Linearly Interpolated GPS Clock Corrections" *2017 Forum on Cooperative Positioning And Service (CPGPS), 19.05.2017, Harbin, Čína,* s. 198-201.

Fainges, J.L. (2015) "Using APSIM, C# and R to Create and Analyse Large Datasets" *21st International Congress On Modelling And Simulation (Modsim2015), 29.11.2015, Gold Coast, Austrálie*, s. 333-339.

Gergelova, M., Kuzevicova, Z., Kovanic, L., Kuzevic, S. (2014) "Automation of Spacial Model Creation in GIS Environment" *Inzynieria Mineralna-Journal Of The Polish Mineral Engineering Society*, č 1., s. 15-22. ISSN: 1640-4920

Gosling, P.C., Symeonakis, E. (2020) "Automated map projection selection for GIS" *Cartography and Geographic Information Science*, sv. 47, č. 3, s. 261-276. Dostupné na: https://doi.org/10.1080/15230406.2020.1717379

Grippa, T., Lennert, M., Beaumont, B., Vanhuysse, S., Stephenne, N. and Wolff, E. (2017) "An Open-Source Semi-Automated Processing Chain for Urban Object-Based Classification" *Remote Sensing*, sv. 9, č. 4. Dostupné na: https://doi.org/10.3390/rs9040358

Hejazi, H., Rajab, H., Cinkler, T., Lengyel, L. (2018) "Survey of Platforms for Massive IoT" *2018 IEEE Internation Conference on Future IoT Technologies, Eger, Maďarsko, 16.01.2018.*

Holzworth, D., Huth, N.I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N.I., Zheng, B. and Snow, V. (2018) "APSIM Next Generation: Overcoming

challenges in modernising a farming systems model" *Environmental Modelling & Software*, sv. 103, s. 43-51. Dostupné na: https://doi.org/10.1016/j.envsoft.2018.02.002

Isikdag, U., Pilouk, M. (2016) "Integration of Geo-Sensor Feeds and Event Consumer Services for Real-Time Representation of IoT Nodes" *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, sv. 41, č. B4, s. 267-274. Dostupné na: https://doi.org/10.5194/isprsarchives-xli-b4-267-2016

Joshva Devadas, T., Thayammal, S. and Ramprakash, A. (2019) "IoT Data Management, Data Aggregation and Dissemination" *Intelligent Systems Reference Library*, sv. 174, s. 385-411. ISSN 18684394. Dostupné na: https://doi.org/10.1007/978-3-030-33596-0_16

Kaippilly Radhakrishnan, K., Moirangthem, J., Panda, S.K., Amaratunga, G. (2018) "GIS Integrated Automation of a Near Real-Time Power-Flow Service for Electrical Grids" *2016 IEEE International Conference on Sustainable Energy Technologies (ICSET)*, s. 48-53. Dostupné na: https://doi.org/10.1109/tia.2018.2855645

Kalbarczyk, R., Kalbarczyk, E. (2021) "Precipitation variability, trends and regions in Poland: Temporal and spatial distribution in the years 1951–2018" Acta geographica Slovenica, sv. 61, č. 2, s. 41-71 Dostupné na: https://doi.org/10.3986/ags.8846

Kliment, T., Bordogna, G., Figerio, L., Crema, A., Boschetti, M., Brivio, P.A., Sterlacchini, S. (2015) "Image Data And Metadata Workflows Automation in Geospatial Data Infrastructure Deployed for Agricultural Sector" *IEEE International Symposium on Geoscience and Remote Sensing IGARSS*, s. 146-149.

Kulawiak, M., Dawidowicz, A., Pacholczyk, M.E. (2019) "Analysis of server-side and client-side Web-GIS data processing methods on the example of JTS and JSTS using open data from OSM and geoportal" *Computers & Geosciences*, sv. 129, s. 26-37. Dostupné na: https://doi.org/10.1016/j.cageo.2019.04.011

Leonard, P.B., Baldwin, R.F., Duffy, E.B., Lipscomb, D.J., Rose, A.M. (2014) "High-throughput computing provides substantial time savings for landscape and conservation planning" *Landscape and Urban Planning*, sv. 125, s. 156-165. Dostupné na: https://doi.org/10.1016/j.landurbplan.2014.02.016

Luo, J., Zu, X., Zhang, C., Wu, X. (2012) "The Origin of Building GIS Software Development Model" *IERI Procedia*, sv. 2, s. 914-920. Dostupné na: https://doi.org/10.1016/j.ieri.2012.06.191

Markovinovic, D., Cetl, V., Samanovic, S. and Bjelotomic Orsulic, O. (2022) "Availability and Accessibility of Hydrography and Hydrogeology Spatial Data in Europe through INSPIRE" *Water*, sv. 14, č. 9. Dostupné na: https://doi.org/10.3390/w14091499

Nařízení vlády č. 430/2006 Sb., o stanovení geodetických referenčních systémů a státních mapových děl závazných na území státu a zásadách jejich používání. (2006) Dostupné na: https://www.psp.cz/sqw/sbirka.sqw?cz=430&r=2006

Nourjou, R. and Hashemipour, M. (2017) "Smart Energy Utilities based on Real-Time GIS Web Services and Internet of Things" *Procedia Computer Science*, sv. 110, s. 8-15. Dostupné na: https://doi.org/10.1016/j.procs.2017.06.070

OGC Standards and Resources. [citováno 2022] Dostupné na: https://www.ogc.org/standards

Ojeda, J.J., Huth, N., Holzworth, D., Raymundo, R., Zyskowski, R.F., Sinton, S.M., Michel, A.J., Brown, H.E. (2021) "Assessing errors during simulation configuration in crop models – A global case study using APSIM-Potato" *Ecological Modelling*, sv. 458. Dostupné na: https://doi.org/10.1016/j.ecolmodel.2021.109703

Parent, J.R., Witharana, C. and Bradley, M. (2022) "Classifying and Georeferencing Indoor Point Clouds With ArcGIS" Photogrammetric Engineering & Remote Sensing, sv. 88, č. 6, s. 383-389. Dostupné na: https://doi.org/10.14358/pers.21-00048r2

Poursafar, N., Alahi, M.E.E., Mukhopadhyay, S (2017) "Long-range Wireless Technologies for IoT Applications: A Review" *2017 Eleventh International Conference on Sensing Technology (ICST), 04.12.2017, Sydney, Austrálie,* s. 304-309.

Rahmati, O., Kornejady, A., Samadi, M., Nobre, A.D., Melesse, A.M. (2018) "Development of an automated GIS tool for reproducing the HAND terrain model" Environmental *Modelling & Software*, sv. 102, s. 1-12. Dostupné na: https://doi.org/10.1016/j.envsoft.2018.01.004

Řezník, T. (2013) “Geographic information in the age of the INSPIRE Directive: discovery, download and use for geographical research” *Geografie*, sv. 118, č. 1, s. 77-93. Dostupné na: https://doi.org/10.37040/geografie2013118010077

Singh, H., Bawa, S. (2016) “Spatial Data Analysis with ArcGIS and MapReduce“ 2016 Ieee International Conference on Computing, Communication and Automation (Iccca), 29.04.2016, Noida, Indie, s. 45-49.

Skoogh, A., Michaloski, J., Bengtsson, N. (2010) “Towards continuously updated simulation models: combining automated raw data collection and automated data processing” *Winter Simulation Conference. 2010 Winter Simulation Conference - (WSC 2010), IEEE*, s. 1678-1689. Dostupné na: https://doi.org/10.1109/wsc.2010.5678901

Směrnice Evropského parlamentu a Rady 2007/2/ES o zřízení Infrastruktury pro prostorové informace v Evropském společenství (INSPIRE) (2007) Dostupné na: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002

Směrnice Evropského parlamentu a Rady 2019/1024 o otevřených datech a opakovaném použití informací veřejného sektoru. (2019) Dostupné na: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L1024

Solihin, W., Eastman, C., Lee, Y.-C., Yang, D.-H. (2017) “A simplified relational database schema for transformation of BIM data into a query-efficient and spatially enabled database” *Automation in Construction*, sv. 84, s. 367-383. Dostupné na: https://doi.org/10.1016/j.autcon.2017.10.002

Soltani, A., Alimagham, S.M., Nehbandani, A., Torabi, B., Zeinali, E., Dadrasi, A., Zand, E., Ghassemi, S., Pourshirazi, S., Alasti, O., Hosseini, R.S., Zahed, M., Arabameri, R., Mohammadzadeh, Z., Rahban, S., Kamari, H., Fayazi, H., Mohammadi, S., Keramat, S., Vadez, V., van Ittersum, M.K., Sinclair, T.R. (2020) “SSM-iCrop2: A simple model for diverse crop species over large areas” *Agricultural Systems*, sv. 182, č. 102855. Dostupné na: https://doi.org/10.1016/j.agsy.2020.102855

Srivastava, N.N. (2018) “Emerging Trends in Open Source Geographic Information Systems“, ISBN10: 1522550399

Široký, J. a kolektiv. (2011) "Tvoříme a publikujeme odborné texty" *Computer Press*. ISBN: 978-80-251-3510-5

Tischler, M.A. (2016) "Accelerating Geospatial Modeling in ArcGIS with Graphical Processor Units" *International Journal of Applied Geospatial Research*, sv. 7, č. 4, s. 41-52. Dostupné na: https://doi.org/10.4018/ijagr.2016100104

Ureche, V., Biboudis, A., Smaragdakis, Y., Odersky, M. (2015) "Automating ad hoc data representation transformations" *ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, sv. 50, č. 10, s.801-820. Dostupné na: https://doi.org/10.1145/2814270.2814271

Vogeler, I., Cichota, R., Snow, V., Jolly, B. (2011) "Development and desktop-assessment of a concept to forecast and mitigate N leaching from dairy farms" *19th International Congress on Modelling and Simulation (Modsim2011), 12.12.2011, Perth, Austrálie*, s. 891-8

Young, N.E., Jarnevich, C.S., Sofaer, H.R., Pearse, I., Sullivan, J., Engelstad, P., Stohlgren, T.J. (2020) "A modeling workflow that balances automation and human intervention to inform invasive plant management decisions at multiple spatial scales" *PLOS ONE*, sv. 15, č. 3. Dostupné na: https://doi.org/10.1371/journal.pone.0229253

Yu, X., Zhang, T. (2013) "The Application of GML in Spatial Data Conversion" 2013 *International Conference on Computer Sciences and Applications, 14.12.2013, Wuhan, Čína,* s. 788-791. Dostupné na: https://doi.org/10.1109/csa.2013.188

Zhang, J., Xu, L., Zhang, Y., Liu, G., Zhao, L., Wang, Y. (2019) "An On-Demand Scalable Model for Geographic Information System (GIS) Data Processing in a Cloud GIS" *ISPRS International Journal of Geo-Information. MDPI AG*, sv. 8, č. 9. Dostupné na: https://doi.org/10.3390/ijgi8090392

Zhao, G., Bryan, B.A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X., Yu, Q. (2013) "Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing" *Environmental Modelling & Software*, sv. 41, s. 231-238. Dostupné na: https://doi.org/10.1016/j.envsoft.2012.08.007

Zhu, J., Wang, X., Chen, M., Wu, P., Kim, M.J. (2019) “Integration of BIM and GIS: IFC geometry transformation to shapefile using enhanced open-source approach” *Automation in Construction*, sv. 106. Dostupné na: https://doi.org/10.1016/j.autcon.2019.102859

8 Příloha - Přehled publikací autora

8.1 Článek impaktovaný

Šimek, P., Vaněk, J., Stočes, M., Jarolímek, J., **Pavlík, J.** (2017) "Mobile accessibility expense analysis of the agrarian WWW portal" *Agricultural Economics (Zemědělská ekonomika)*, roč. 63, č. 5, s. 197-203. ISSN: 0139-570X. Dostupné na: https://doi.org/10.17221/313/2015-AGRICECON

Benda, P., **Pavlík, J.**, Masner, J. (2019) "Practical education of adults with intellectual disabilities using a web course" *Problems of Education in the 21st Century*, sv. 77, č. 4, s. 463-477. ISSN: 1822-7864. Dostupné na: http://dx.doi.org/10.33225/pec/19.77.463

Pavlík, J., Hrnčírová, M., Stočes, M., Masner, J., Vaněk, J. (2020) "Usability of IoT and Open Data Repositories for Analyzing Water Pollution. A Case Study in the Czech Republic" *ISPRS International Journal of Geo-Information*, sv. 9, č. 10. ISSN: 2220-9964. Dostupné na: https://doi.org/10.3390/ijgi9100591

Ronanki, S., **Pavlík, J.**, Masner, J., Jarolímek, J., Stočes, M., Subhash, D., Talwar, H., Tonapi, V., Srikanth, M., Baddam, R., Kholová, J. (2022) "An APSIM-powered framework for post-rainy sorghum-system design in India" *Field Crops Research*, 2022, sv. 277, č. 108422. ISSN: 0378-4290. Dostupné na: https://doi.org/10.1016/j.fcr.2021.108422

8.2 Článek Scopus

Pavlík, J., Vaněk, J., Stočes, M. (2015) "Software tools for movement visualization in agrarian sector" *AGRIS on-line Papers in Economics and Informatics*, roč. 7, č. 2, s. 68-76. ISSN: 1804-1930. Dostupné na: https://online.agris.cz/archive/2015/02/07

Šimek, P., Vaněk, J., **Pavlík, J**. (2015) "Usability of UX Methods in Agrarian Sector – Verification" *AGRIS on-line Papers in Economics and Informatics*, roč. 7, č. 3, s. 49-56. ISSN: 1804-1930. Dostupné na: https://online.agris.cz/archive/2015/03/05

Stočes, M., Vaněk, J., Masner, J., **Pavlík, J**. (2016) "Internet of Things (IoT) in Agriculture - Selected Aspects" *AGRIS on-line Papers in Economics and Informatics*, roč. 8, č. 1, s. 83-88. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2016.080108

Habibi, A., Ulman, M., Vaněk, J., **Pavlík, J**. (2016) "Measurement and Analysis of Quality of Service of Mobile Networks in Afghanistan – End User Perspective" *AGRIS on-line Papers in Economics and Informatics*, roč. 8, č. 4, s. 73-84. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2016.080407

Stočes, M., Šimek, P., **Pavlík, J**. (2017) "Metadata Formats for Data Sharing in Science Support Systems" *AGRIS on-line Papers in Economics and Informatics*, roč. 9, č. 3, s. 61-69. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2017.090306

Jarolímek, J., Stočes, M., Masner, J., Vaněk, J., Šimek, P., **Pavlík, J.**, Rajtr, J. (2017) "User-Technological Index of Precision Agriculture" *AGRIS on-line Papers in Economics and Informatics*, roč. 9, č. 1, s. 69-75. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2017.090106

Pavlíček, J., Jarolímek, J., Jarolímek, J., Pavlíčková, P., Dvořák, S., **Pavlík, J.**, Hanzlík, P. (2018) "Automated Wildlife Recognition" *AGRIS on-line Papers in Economics and Informatics*, roč. 10, č. 1, s. 51-60. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2018.100105

Rajtr, J., Šimek, P., **Pavlík, J.** (2018) "Proposing of Single Entity Design Pattern in Big Agricultural Positioned Data Sets (ADS)" *AGRIS on-line Papers in Economics and Informatics*, roč. 10, č. 4, s. 65-69. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2018.100407

Jarolímek, J., **Pavlík, J.**, Kholova, J., Ronanki, S. (2019) "Data Pre-processing for Agricultural Simulations" *AGRIS on-line Papers in Economics and Informatics*, roč. 11, č. 1, s. 49-53. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2019.110105

Novák, V., Stočes, M., Kánská, E., **Pavlík, J.** and Jarolímek, J. (2019) "Monitoring of Movement on the Farm Using WiFi Technology" *AGRIS on-line Papers in Economics and Informatics*, roč. 11, č. 4, s. 85-92. ISSN 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2019.110408

Novák, V., **Pavlík, J.**, Stočes, M., Vaněk, J., Jarolímek, J. (2020) "Welfare with IoT Technology Using Fuzzy Logic" *AGRIS on-line Papers in Economics and Informatics*, roč. 12, č. 2, s. 111-118. ISSN: 1804-1930. Dostupné na: https://doi.org/10.7160/aol.2020.120210

8.3 Stať ve sborníku

Šimek, P., **Pavlík, J.**, Jarolímek, J. (2015) "Increase in work efficiency with information sources in areas of agriculture and rural development using UX" *Agrarian Perspectives XXIV. – Global Agribusiness and Rural Economy, 16.09.2015, Praha, Česká republika,* s. 433-439.

Šimek, P., **Pavlík, J.** (2016) "Analysis of software licensing options in agricultural companies" *Agrarian perspectives XXV. – Global and European Challenges for Food Production, Agribusiness and the Rural Economy, 14.09.2016, Praha, Česká republika,* s. 365-370.

Pavlík, J., Masner, J., Rajtr, J. (2017) "Data processing methods for information system testing in agrarian sector" *10th IADIS International Conference on Information Systems 2017, IS 2017 10.04.2017, Budapešť, Maďarsko,* s. 207-210.

Havránek, M., Šmejkalová, M., Masner, J., **Pavlík, J.** (2017) "Mobile applications for animal loss mitigation caused by haymaking" *IADIS International Conference e-Society 2017, ES 2017 10.4.2017, Budapešť, Maďarsko,* s. 57-64.

Očenášek, V., Masner, J., Vaněk, J., Šilerová, E., **Pavlík, J.** (2017) "Evaluation of Farmer's E-shops" *8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2017) 21.09.2017, Chania, Řecko*. CEUR-WS, s. 224-231.

Stočes, M., **Pavlík, J.**, Masner, J., Jungwirth, M., Havlíček, Z. (2017) "Monitoring visualization of the animal movement in buildings" *Agrarian Perspectives XXVI. Competitiveness of European Agriculture and Food Sectors, 13.09.2017, Praha, Česká republika*, s. 369-374.

Šimek, P., **Pavlík, J.**, Jarolímek, J., Očenášek, V., Stočes, M. (2017) "Use of Unmanned Aerial Vehicles for Wildlife Monitoring" *8th International Conference on Information and Communication Technologies in Agriculture, Food and Environment, HAICTA 2017 21.09.2017, Chania, Řecko,* CEUR-WS, 2017. s. 795-804.

Stočes, M., Jarolímek, J., Šimek, P., Charvát, K., Masner, J., **Pavlík, J.**, Vaněk, J. (2017) "User-technological index of precision agriculture" *7th Asian-Australasian Conference on Precision Agriculture 16.10.2017, Hamilton, Nový Zéland*.

Vaněk, J., Jarolímek, J., **Pavlík, J.**, Masner, J., Stočes, M. (2018) "Practical Applications Of User-Technological Index of Precision Agriculture" *Agriculture & Food 19.06.2018, Ellenite Bulharsko,* s. 256-263.

Pavlík, J., Benda, P., Šilerová, E., Jungwirth, M. (2018) "Scalability of GIS applications with regards to IoT" *Agrarian perspectives XXVII., Food Safety – Food Security 19.09.2018, Praha, Česká republika*, s. 210-214.

Pavlík, J., Masner, J., Jarolímek, J., Lukáš, M. (2019) "Data Processing for Yield Optimization" *Agrarian perspectives XXVIII. Business Scale in Relation to Economics 18.09.2019, Praha, Česká republika,* s. 189-193.

Šimek, P., Jarolímek, J., Kánská, E., Stočes, M., Vaněk, J., **Pavlík, J.**, Vasilenko, A. (2019) "Earth Observation Data and Spatial Data Sets Analysis" *EFITA 2019, 27.06.2019, Rhodos, Řecko*.

Masner, J., Vaněk, J., **Pavlík, J.**, Kánská, E., Šilerová, E. (2019) "Options for Automatic Identification of User Activities in Usability Testing" *EFITA-HAICTA-WCCA CONGRESS 27.01.2020, Rhodos, Řecko*.

Sabou, J., Ulman, M., **Pavlík, J.** (2019) "Survey of Social Media Usage in the Czech Agricultural Sector" *Agrarian Perspectives XXVIII. Business Scale in Relation to Economics 18.09.2019, Praha, Česká republika,* s. 257-264.

Masner, J., Šimek, P., Jarolímek, J., Očenášek, V., **Pavlík, J.** (2020) "Analysis of CSS organization styles and expensive properties in regard to rendering performance" *Agrarian Perspectives XXIX. Trends and Challenges of Agrarian Sector 16.09.2020, Praha, Česká republika,* s. 215-222.

Pavlík, J., Vaněk, J., Masner, J., Stočes, M., Očenášek, V. (2020) "Support tools for agricultural production simulation processing" *9th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2020) 24.09.2020, Soluň, Řecko*. CEUR-WS.org, s. 468-474.

Kánská, E., Očenášek, V., Jarolímek, J., Vaněk, J., **Pavlík, J.** (2021) "Survey Methodology – Survey 2021" *Agrarian Perspectives XXX. Sources of Competitiveness under Pandemic and Environmental Shocks. 15.09.2021, Praha, Česká republika,* s. 117-128.