

**VÝZKUMNÝ ÚSTAV
VODOHOSPODÁŘSKÝ
T.G. MASARYKA**

veřejná výzkumná instituce

Vztahy mezi indikátory sucha

**Projekt Návrh koncepce řešení krizové situace vyvolané výskytem sucha
a nedostatkem vody na území ČR**

VG20102014038

Havlíček V. a kol.

Zadavatel: Ministerstvo vnitra ČR

Praha, prosinec 2014

Název organizace:

Výzkumný ústav vodohospodářský T. G. Masaryka, v.v.i.

Podbabská 30, 160 00 Praha 6

Ředitel:

Mgr. Mark Rieder

Zadavatel:

Ministerstvo vnitra České republiky

Milady Horákové 5/133, 160 00 Praha 6

Zástupce zadavatele:

Ing. Lydie Kotajná, M.A.

Projekt:

Návrh koncepce řešení krizové situace vyvolané výskytem sucha a nedostatkem vody na území ČR VG20102014038

Zahájení a ukončení projektu:

10/2010–12/2014

Místo uložení zprávy:

SVTI VÚV TGM, v.v.i.

Náměstek ředitele pro výzkumnou a odbornou činnost:

Ing. Petr Bouška, Ph.D.

Vedoucí odboru:

Ing. Anna Hrabánková

Vedoucí řešitel:

Ing. Radek Vlnas

Řešitelé VÚV:

Ing. Adam Beran, Ing. Martin Hanel, Ph.D, Ing. Anna Hrabánková, RNDr. Tomáš Hrdinka, Ph.D., Ing. Ladislav Kašpárek, CSc., Mgr. Marta Martínková, Ing. Martina Peláková, Mgr. Pavel Tremel, Ing. Adam Vizina

Řešitelé ČZÚ:

Ing. Petr Bašta, Ing. Lukáš Jačka, Ing. Petr Máca, Ph.D., Ing. Jiří Pavlásek, Ph.D., prof. Ing. Pavel Pech, CSc.

GENETICKÉ PROGRAMOVÁNÍ	1
Základní stavební prvky a operace v GP	1
Metodika	2
Rozdělení modelů v závislosti na typu vstupů	3
Nastavení programu SORD! a ANN pro účely modelování indikátorů sucha.....	3
VÝSLEDKY A DISKUZE	5

GENETICKÉ PROGRAMOVÁNÍ

Genetické programování patří mezi evoluční výpočetní techniky, přesněji do kategorie evolučních algoritmů (EA). Evoluční algoritmy jsou populárními metodami pro řešení mnoha obtížných problémů a nabízejí výkonnou alternativu ke klasickým matematickým metodám. Časté použití EA je v oblasti optimalizačních úloh, data miningu, klasifikací, automatického programování, strojového učení a v bioinformatice.

V současné době se rozlišují 4 hlavní kategorie EA – evoluční programování (EP), evoluční strategie (ES), genetické algoritmy (GA) a genetické programování (GP). Hlavní rozdíly v klasických kategoriích EA jsou v reprezentaci jedinců, použití variačních operátorů a v selekčních a reprodukčních mechanismech (Bäck, 1997).

Základními prvky EA jsou jedinci reprezentující řešení, selekce a variace. Soubor jedinců tvoří populaci. Počáteční generace jedinců je vytvořena nejčastěji pomocí náhodných inicializačních metod. Hlavním znakem EA je aplikace darwinovské evoluce – lepší jedinci přežívají a množí se. Kvalitu (zdatnost) jedince definuje fitness funkce.

Selekční operátory simulují interakce jedinců s prostředím – představují tlak prostředí na jedince, jehož kvalita je vyjádřena pomocí hodnoty fitness. Jedinci v populaci se vyvíjejí a interagují mezi sebou, což je zajištěno variačními operátory (genetické operátory). Tyto operátory jsou vždy stochastické. Mezi variační operátory se řadí zejména rekombinace a mutace (Eiben, 2002).

Genetické programování je nejmladší široce uznávaná samostatná kategorie EA. Metoda GP je odvozena z GA. Hlavní rozdíl mezi GA a GP je v reprezentaci jedince. V původním GP jsou reprezentací jedinců syntaktické stromy. V GA jsou reprezentací jedinců binární řetězce. GP je zobecněním GA (Langdon, 2002).

GP je obecná optimalizační metoda, která slouží k automatickému vyvíjení programů s využitím simulovaného evolučního procesu. GP umožňuje přímo najít vyhovující optimální funkční vztah (program) pro daný problém. GP umožňuje manipulaci nejen s hodnotami argumentů, ale i se samotnými funkcemi či soubory funkcí.

V roce 1992 vydal J. Koza první monografii věnovanou GP (Koza, 1992). Rok 1992 je obecně přijímán za čas vzniku GP jako samostatné evoluční techniky.

Základní stavební prvky a operace v GP

V GP se pracuje s terminály a funkcemi. Jedná se o primitiva, z nichž je pomocí GP sestavován výsledný program. Volba prvků množiny funkcí a terminálů je první úkolem při návrhu algoritmu GP. Do množiny terminálů se řadí externí hodnoty (vstupní data) a konstanty generované algoritmem. Množina funkcí zahrnuje procesy prováděné s terminálovými hodnotami.

Terminály a funkce se sestavují nejčastěji do podoby stromových grafů (syntaktických stromů), což je hlavní reprezentace GP, ale jsou možné i jiné reprezentace.

Průběh GP optimalizace začíná inicializací nových jedinců, nebo-li nových náhodných stromových struktur, které představují první generaci programů v evolučním procesu GP.

Evaluační metrikou v GP je fitness funkce. Správná volba fitness funkce je závislá na řešeném problému. Účel fitness funkce je poskytnout GP algoritmu zpětnou vazbu nesoucí informaci o jedincích, kteří jsou vhodnými kandidáty pro přežití a tvorbu nové generace. Mechanismem rozhodujícím o přežití jedince na základě hodnoty fitness je selekce.

Operátor selekce nevytváří nová řešení, ale z již existujících řešení vybírá ta, která jsou dostatečně dobrá. Selektce je nezávislá na reprezentaci, a tak je možné v GP používat jakékoliv selekční algoritmy pro evoluční algoritmy. Nejvíce využívanou metodou selekce v GP je turnajová selekce. Obliba této metody vyplývá z

jednoduché implementovatelnosti a automatického přeškálování fitness. Výhodou je také snadná regulace selekčního tlaku. Jedinci vybraní na základě selekce buď postupují do další generace beze změny (replikace) nebo jsou využiti ve variačních operátorech.

Variační operátory slouží v GP stejně jako u jiných EA k změně jednotlivých jedinců a umožňuje evoluci jedinců v populaci. Ve většině GP systémů je dominantním variačním operátorem rekombinace (Koza, 1992). Obvyklou formou rekombinace v GP je křížení podstromů. Mutace přináší do populace diverzitu. Jedná se o variační operátor, který při aplikaci pracuje jen s jedním jedincem.

Jedna generace běhu GP je souborem postupně provedených operací:

- selekce
- aplikace variačních operátorů
- stanovení fitness
- test zastavovacího kritéria

Pro zastavení běhu GP se používá definované zastavovací kritérium. Tím bývá nejčastěji definovaný maximální počet generací, čas nebo daná hodnota fitness.

Metodika

Pro účely modelování indikátorů sucha pomocí GP byl použit software SORD!. Program SORD! je vytvořen jako nástroj pro modelování pomocí GP v prostředí programovacího jazyka R (R Development Core Team, 2011). SORD! je z hlediska použitého typu GP systém pracující se stromovou reprezentací jedinců, která je použita v linearizované formě. Kromě standardních součástí GP disponuje program SORD! možnostmi volby speciálních funkcí, možnostmi více běhového režimu a následného statistického vyhodnocení, možnostmi ukládání textových i grafických výstupů. V programu SORD! Jsou rovněž implementovány vybrané metody vážení dat (nelineární vážení, vážení pomocí k-nejbližších sousedů, SAW – Stepwise Adaptation Weights, (Eggermont, 2001; Vladislavleva, 2007)). Program SORD! je freeware <http://www.kvhem.cz/vyzkum/software/>.

Hlavní parametry programu SORD! jsou následující: počet jedinců v populaci, počet generací, počet nezávislých běhů programu, typ fitness, maximální hloubka stromů v počáteční generaci a v průběhu evoluce, pravděpodobnosti variačních operátorů, typ selekce, množina terminálů a funkcí.

V programu SORD! jsou implementovány křížení podstromů tři typy mutací: mutace podstromů, konstant a separace.

Při selekci je možné využít čtyři typy selekcí: jednoduchý a dvojitý turnaj, dvojitý kontrolovaný turnaj a elitismus.

Volitelné funkce pro vytvoření množiny funkcí a terminálů jsou uvedeny v **Tab. 1**. Horní index p značí chráněnou funkci. Ochrana funkcí je nutná z důvodu zabránění pádu programu v důsledku nevhodného vstupu do funkce. SMA značí funkci jednoduchého klouzavého průměru, CSUM klouzavý součet, RES značí lineární nádrž a DLY je funkce pro zpoždění vstupu.

Kritéria použitelná jako fitness funkce jsou střední absolutní chyba (MAE), střední kvadratická chyba (MSE), $\sqrt{\text{MSE}}$ (RMSE), upravený Nashův-Sutcliffův koeficient (NS0) a index persistence.

Tab. 1 Dostupné funkce a terminály programu SORD!

Funkce	$+$, $-$, $*$, $/^p$, \sqrt^p , \exp^p , \sin , \cos , \tan , \tanh , \ln^p , \log^p , x^y^p , sigmoid^p , \max , \min , hstep , sign , less , greater , sawtooth^p , SMA^p , RES^p , DLY^p , CSUM^p
Terminály	vstupní proměnné, celočíselné nebo reálné konstanty

Rozdělení modelů v závislosti na typu vstupů

Odhad vztahů pro stanovení indikátoru sucha – závislá proměnná DMRI (Drought Magnitude Runoff Index) – bylo možné provést na základě tří nezávislých proměnných: DMPI (Drought Magnitude Precipitation Index), DMPEI (Drought Magnitude Precipitation Evaporation Index) a SEP (Standardized Effective Precipitation). V závislosti kombinaci použitých nezávislých proměnných byly stanoveny následující čtyři varianty modelů:

- $DMRI = f(DMPI, DMPEI, SEP)$
- $DMRI = f(DMPI)$
- $DMRI = f(DMPEI)$
- $DMRI = f(SEP)$

Dále jsou jednotlivé modely značeny jako M-ALL, M-DMPI, M-DMPEI a M-SEP. Rovnice pro tyto čtyři varianty modelů byly hledány na datech 11 povodí. Datové soubory pro indikátory sucha DMRI, DMPI, DMPEI a SEP pro tuto povodí byly dostupné pro jednotné období 1961–2010. Na těchto datech bylo prováděno trénování. Validace prováděna nebyla, záměrem studie bylo zjistit vztahy mezi jednotlivými proměnnými, zjistit strukturu výsledných modelů a identifikovat nezávislé proměnné (DMRI, DMPI, DMPEI a SEP), které nejvíce ovlivňují hodnotu DMRI. Pro tyto účely není nutné provádět validaci. V případě použití GP pro vlastní operativní řízení a předpovědi sucha na daných povodích by však bylo nutné validaci provést.

V rámci stanovených modelů budou porovnány výsledky jednotlivých modelů a tím i identifikace nezávislých proměnných, které mají největší vliv na stanovení DMRI. Dále bude provedeno porovnání kvality simulace provedené pomocí vztahů odvozených pomocí GP a kvality simulace provedené modelem umělé neuronové sítě (ANN). Výpočet pomocí ANN a následné porovnání bude provedeno pouze pro variantu M-ALL.

Nastavení programu SORD! a ANN pro účely modelování indikátorů sucha

Pro nalezení vztahů simulujících nejlépe závislosti mezi DMRI a ostatními nezávislými proměnnými bylo zvoleno následující výchozí nastavení programu SORD!. Hodnoty většiny parametrů odpovídají obecně doporučovaným hodnotám, které jsou uváděny v literatuře (Koza, 1992).

Parametr hloubky stromů byl oproti doporučovaným hodnotám zvolen výrazně nižší a to z důvodu snížení složitosti výsledné rovnice modelu a tím i snazší interpretovatelnosti, vyšší robustnosti modelu a lepší porovnatelnosti jednotlivých rovnic. Nicméně porovnání kvality simulace mezi modely odvozenými na základě větší hloubky stromů a modely s malou hloubkou bylo provedeno a výsledný rozdíl byl v průměru 8 % v hodnotě kritéria NS ve prospěch složitějších modelů. Tento rozdíl je zanedbatelný vzhledem k cílům této studie, kde získání jednodušších modelů bylo prioritou.

Parametr množiny používaných funkcí je závislý na řešeném problému a byl tedy pro účely této studie sestaven na základě zkušebních výpočtů, při kterých se postupuje od rozsáhlejší množiny funkcí k méně početné množině funkcí často zastoupených ve výsledných vztazích.

Tab. 2 Nastavení programu SORD!

Parametr	Hodnota
počet nezávislých běhů	50
počet jedinců v populaci	1000
počet generací	100
hloubka stromů (počáteční generace/průběh evoluce)	1/2
fitness	RMSE
pravděpodobnosti křížení	0,7
pravděpodobnost mutace (podstromů/separace/konstant)	0,5/0,3/0,5
typ selekce	turnaj (velikost 4)
použité funkce	+, -, *, /, CSUM, SMA, x^y
použité terminály	vstupní proměnné, reálné konstanty

Referenční metodou pro porovnání výsledků GP byl zvolen model umělé neuronové sítě (ANN). Použit byl standardní ANN algoritmus vícevrstvého perceptronu se zpětnou propagací chyby. Nastavení hlavních parametrů neuronové sítě bylo následující: počet nezávislých běhů: 50, počet epoch: 1000, 2 vnitřní vrstvy s 5 respektive 3 neurony, aktivační funkce: hyperbolická tangenta, bias: ano, učící konstanta 0,5, momentum 0,8, testována historie vstupů 1–12.

VÝSLEDKY A DISKUZE

Výsledné hodnoty Nashova-Sutcliffeova kritéria pro simulace DMRI na jednotlivých povodí pomocí všech variant modelů jsou uvedeny v **Tab. 3**. Hodnoty pro model se všemi možnými typy vstupů se pohybují v rozmezí od 0,108 po maximum 0,449, které bylo dosaženo na povodí 0590. Nejhorších výsledků bylo dosaženo na povodí 0665 (NS=0,108). Medián pro variantu modelu se všemi typy vstupů je 0,321. V porovnání s ostatními variantami modelů poskytovala varianta M-ALL obecně nejvyšší hodnoty NS kritéria. Při porovnání variant modelů pouze s jedním z možných vstupů je patrná dominance modelu M-SEP, jehož simulace měly ve většině případů nejvyšší hodnotu NS. Pouze u povodí 0590, 0665 a 1901 měl lepší simulaci model M-DMPEI. Hodnota mediánu pro M-SEP je srovnatelná s mediánem pro M-ALL a je možné usuzovat, že proměnná SEP má z jednotlivých proměnných největší vliv na výslednou hodnotu DMRI.

Rovnice pro jednotlivé modely odvozené pomocí programu SORD! pro jednotlivá jsou v obecném tvaru uvedeny v **Tab. 4**, rovnice včetně hodnot konstant jsou uvedeny v **Tab. 5**. V těchto tabulkách je uvedena struktura modelu a reálné konstanty jsou uvedeny v podobě parametrů modelu *a* a *b*.

Z porovnání modelů je zřejmé, že ve variantě M-ALL je v jednotlivých rovnicích zastoupena proměnná SEP spolu s proměnnou DMPEI a to celkově v 5 případech z 11. Na základě tohoto nezanedbatelného výskytu obou proměnných je možné konstatovat, že pro simulaci DMRI je nejefektivnější kombinace vstupů DMPEI a SEP. Naproti tomu se proměnná DMPI v rovnicích pro variantu M-ALL nikde nevyskytuje a v porovnání s ostatními variantami má M-DMPI nejnižší hodnoty NS.

Z rovnic modelů je dále patrné, že nejčastěji se na jejich stavbě podílí funkce SMA (funkce jednoduchého klouzavého průměru). SMA je dominantní zejména v modelech M-DMPI a M-DMPEI. Dále je zejména u modelu M-SEP využívána funkce CSUM (klouzavý součet). U modelu M-ALL je zastoupení výše zmíněných funkcí přibližně stejné. V rámci různých variant modelů je možné identifikovat určité podobnosti v rovnicích napříč povodími. Nejvíce podobných rovnic je u modelu M-DMPEI, kde jednoduchá funkce klouzavého průměru s konstantou upraveným vstupem je nejčastější. Dost podobné jsou i rovnice pro M-SEP, které jsou tvořeny klouzavým součtem, která však má složitější argument počtu prvků pro sumu – argumentem je zde samotná upravená proměnná. Naopak největší rozdíly ve struktuře rovnic pro jednotlivá povodí jsou u varianty M-ALL.

Tab. 6 uvádí porovnání simulací pro variantu M-ALL pomocí vztahů z GP a výsledků simulace naučenou umělou neuronovou sítí. Výsledky GP jsou ve všech případech lepší než výsledky ANN, i když v některých případech jsou hodnoty NS velmi blízko. Model ANN často slouží jako ověření výsledků GP i jiných datově orientovaných modelů. Vzhledem k tomu, že výsledky GP jsou lepší nebo podobné jako výsledky ANN je možné konstatovat, že v rámci řešené úlohy provedlo GP při zvoleném nastavení dostatečné prohledání prostoru řešení.

Na **Obr. 1 a 2** jsou ukázány simulace indikátorů DMRI pro M-ALL s nejvyšší, respektive nejnižší hodnotou NS kritéria pro 1000 záznamů z simulované řady.

Tab. 3 Hodnoty NS kritéria pro jednotlivé modely

Povodí	M-ALL	M-DMPI	M-DMPEI	M-SEP
0148	0.340	0.179	0.157	0.340
0180	0.231	0.076	0.145	0.229
0240	0.321	0.197	0.233	0.267
0250	0.237	0.183	0.198	0.237
0490	0.404	0.189	0.326	0.372
0520	0.282	0.151	0.228	0.282
0590	0.449	0.161	0.395	0.328
0665	0.108	0.060	0.108	0.090
1900	0.268	0.121	0.086	0.256
1901	0.378	0.059	0.157	0.037
2175	0.332	0.102	0.016	0.332
Max.	0.449	0.197	0.395	0.372
Medián	0.321	0.151	0.157	0.267
Min.	0.108	0.059	0.016	0.037

Tab. 4 Rovnice jednotlivých modelů v obecném tvaru

Povodí	M-ALL: DMRI=	M-DMPI: DMRI=	M-DMPEI: DMRI=	M-SEP: DMRI=
0148	$SMA(SEP/a)+b^hSEP$	$SMA(DMPI/a/(DMPI+b))$	$SMA(DMPEI/a/b)$	$SMA(SEP/a)+b^hSEP$
0180	$CSUM(SEP/a.(DMPEI+b))$	$SMA(DMPI/a.b)$	$SMA((DMPEI+a).b)$	$CSUM(SEP/a.(SEP+b))$
0240	$(SEP+a)/(DMPEI+b)$	$CSUM(DMPI/a.(DMPI+b))$	$CSUM(DMPEI/a.(DMPEI+b))$	$CSUM(SEP/a.(SEP+b))$
0250	$CSUM(SEP/a.(SEP+b))$	$CSUM(DMPI/a.(DMPI+b))$	$SMA(SMA(DMPEI.a).(DMPEI)^a)$	$CSUM(SEP/a.(SEP+b))$
0490	$SMA(DMPEI.a)+SEP/b$	$CSUM(DMPI/a.(DMPI+b))$	$SMA(DMPEI.a)+DMPEI/b$	$CSUM(SEP/a.SMA(SEP.b))$
0520	$CSUM(SEP/a.CSUM(SEP.b))$	$SMA((DMPI+a).b)$	$SMA(DMPEI/a.b)$	$CSUM(SEP/a.CSUM(SEP.b))$
0590	$SMA(DMPEI.a)+SEP/b$	$SMA((DMPI+a).b)$	$SMA(DMPEI*a.b)$	$CSUM(SEP/a.CSUM(SEP.b))$
0665	$SMA(DMPEI/a.b)$	$SMA((DMPI+a).b)$	$SMA(DMPEI*a.b)$	$SMA(SEP*a.b)$
1900	$SMA(SEP-DMPEI.SEP*a)$	$SMA(DMPI/a.b)$	$SMA((DMPEI+a).b)$	$SMA((SEP+a).CSUM(SEP.b))$
1901	$SMA(a/SEP.b)$	$SMA(-(DMPI+a).b)$	$SMA(a/DMPEI.b)$	$(SEP+a)*b^hSEP$
2175	$CSUM(SEP/a.CSUM(SEP.b))$	$SMA((DMPI+a).b)$	$SMA(SMA(DMPEI.a).b)$	$CSUM(SEP/a.CSUM(SEP.b))$

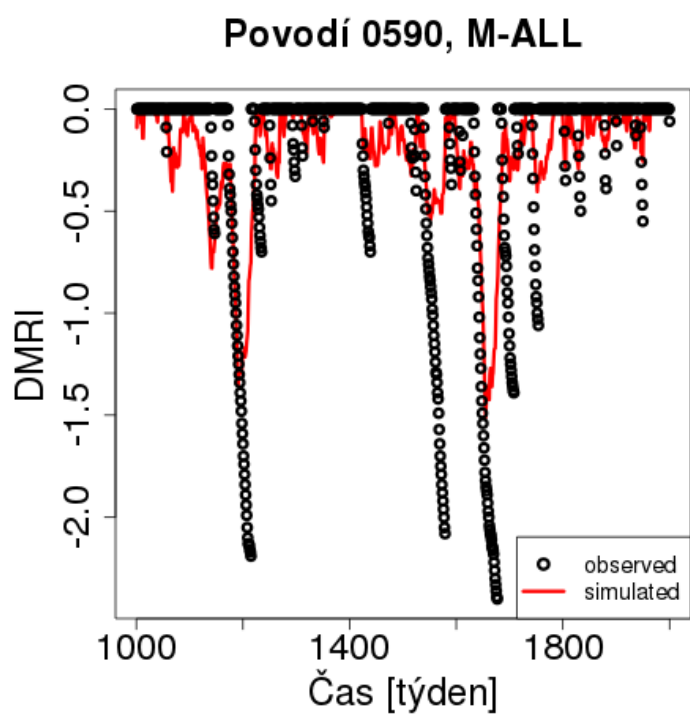
Tab. 5 Rovnice jednotlivých modelů s hodnotami konstant

Povodí	M-ALL: DMRI=	M-DMPI: DMRI=	M-DMPEI: DMRI=	M-SEP: DMRI=
0148	$SMA(SEP,24)+6.452^{*}SEP$	$SMA(DMPI/1.127,(DMPI+28.412))$	$SMA(DMPEI*1.392,84)$	$SMA(SEP,24)+6.452^{*}SEP$
0180	$CSUM(SEP/15.731,(DMPEI+6.59))$	$SMA(DMPI/1.292,30)$	$SMA((DMPEI-0.041),32)$	$CSUM(SEP/16.799,(SEP+8.269))$
0240	$(SEP-0.306)/(DMPEI+3.573)$	$CSUM(DMPI/15.862,(DMPI-8.474))$	$CSUM(DMPEI/12.811,(DMPEI+-8.49))$	$CSUM(SEP/6.997,(SEP-1.132))$
0250	$CSUM(SEP/15.133,(SEP-5.723))$	$CSUM(DMPI/26.057,(DMPI-18.346))$	$SMA(SMA(DMPEI,17),(DMPEI^2))$	$CSUM(SEP/15.134,(SEP-5.724))$
0490	$SMA(DMPEI,20)+SEP/5.177$	$CSUM(DMPI/26.855,(DMPI-18.44))$	$SMA(DMPEI,22)+DMPEI/6.162$	$CSUM(SEP/3.818,SMA(SEP,18))$
0520	$CSUM(SEP/47.162,CSUM(SEP,23))$	$SMA((DMPI+0.035),122)$	$SMA(DMPEI/0.805,54)$	$CSUM(SEP/47.162,CSUM(SEP,23))$
0590	$SMA(DMPEI,32)+SEP/7.361$	$SMA((DMPI+0.069),105)$	$SMA(DMPEI*1.151,32)$	$CSUM(SEP/29.056,CSUM(SEP,10))$
0665	$SMA(DMPEI/0.809,77)$	$SMA((DMPI+0.041),154)$	$SMA(DMPEI*1.235,77)$	$SMA(SEP*0.879,155)$
1900	$SMA(SEP-DMPEI,SEP*-6.422)$	$SMA(DMPI/0.949,51)$	$SMA((DMPEI-0.1),38)$	$SMA((SEP+0.321),CSUM(SEP,9))$
1901	$SMA(0.06/SEP,56)$	$SMA(-(DMPI+0.379),800)$	$SMA(0.597/DMPEI,34)$	$(SEP-0.124)*6.438^{*}SEP$
2175	$CSUM(SEP/63.901,CSUM(SEP,47))$	$SMA((DMPI+0.033),93)$	$SMA(SMA(DMPEI,27),522)$	$CSUM(SEP/63.897,CSUM(SEP,47))$

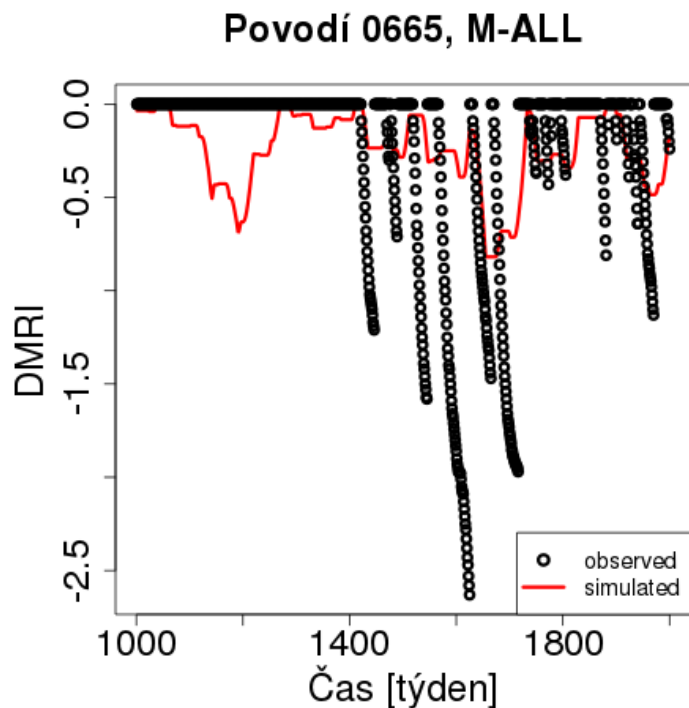
Tab. 6 Výsledky simulací pomocí GP a ANN pro M-ALL

Povodí	NS GP	NS ANN
0148	0.340	0.014
0180	0.231	0.265
0240	0.321	0.343
0250	0.237	0.279
0490	0.404	0.366
0520	0.282	0.208

0665	0.108	0.031
1900	0.268	0.173
1901	0.378	0.093
2175	0.332	-0.131
Max.	0.449	0.366
Medián	0.321	0.208
Min.	0.108	-0.131



Obr. 1: Průběh simulace pro povodí s nevyšším NS pro model M-ALL.



Obr. 2: Průběh simulace pro povodí s nejnižším NS pro model M-ALL.

LITERATURA

Bäck, T., Fogel, D., Michalewicz, Z. (2000) *Evolutionary Computation 1: Basic Algorithms and Operators*. Institute of Physics Publishing, Bristol.

Eiben, A., Schoenauer, M. (2002) Evolutionary computing. *Information Processing Letters.*, 82, 1, 1–6.

Langdon, W. B., Poli, R. (2002) *Foundations of Genetic Programming*. Springer-Verlag, 274 s.

Koza, J. R.: *Genetic Programming (1992) On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA : MIT Press.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2011). URL <http://www.R-project.org>. ISBN 3-900051-07-0.

Vladislavleva, E., Smits, G., den Hertog, D. (2010) On the importance of data balancing for symbolic regression. *IEEE Transactions on Evolutionary Computation* 14(2), 252–277.

Eggermont, J., Hemert, J.I.v. (2001) Adaptive genetic programming applied to new and existing simple regression problems. In: *Proceedings of the 4th European Conference on Genetic Programming, EuroGP '01*, s. 23–35. Springer-Verlag.