نشریه آبیاری و زهکشی ایران
شماره3، جلد 11، مرداد - شهریور 1396، ص. 366-356

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 11, Agu.-Sept. 2017, p. 356-366

مقایسه مدل‌های KNN و درخت تصمیم M5 در پیش‌بینی تبخیر و مقایسه آن با مدل‌های تجربی (مطالعه موردی بیرجند)

آتنا خلیلی نفت‌چالی[1*]، عباس خاشعی سیوکی[2]، علی شهیدی[3]

تاریخ دریافت: 1395/4/15 تاریخ پذیرش: 1395/11/24

چکیده

در نواحی خشک و نیمه‌خشک تبخیر از سطح خاک بخش مهمی از بیلان آب در خاک است. در این تحقیق برای برآورد تبخیر از سطح آزاد آب از معادلات تجربی هفنر، شاهتین، مارسیانو، تیچومیروف، ایوانف، سازمان عمران اراضی آمریکا و مایر استفاده شد و این معادلات در منطقه مورد مطالعه اصلاح گردید. از مدل KNN و درخت تصمیم M5 نیز به‌عنوان یکی از شیوه‌های داده‌کاوی برای برآورد تبخیر از سطح آزاد آب بهره گرفته شد. بدین منظور از داده‌های هواشناسی ایستگاه همدیدی بیرجند برای یافتن بهترین داده‌های ورودی اثرگذار بر تبخیر استفاده گردید. همچنین آزمون گاما برای یافتن بهترین ترکیب پارامترهای ورودی برای برآورد تبخیر اجرا گشت. برای مقایسه نتایج آماره‌ها ریشه متوسط خطای مربعات (RMSE)، ضریب همبستگی (R^2) و متوسط قدر مطلق خطا (MAE) مورد تحلیل قرار گرفت. نتایج نشان داد که در اکثر موارد مدل درخت تصمیم M5 نتایج مطلوب‌تری را نسبت به مدل KNN و معادلات تجربی حاصل می‌کند اما معادله‌ی مایر اصلاح‌شده برای منطقه‌ی بیرجند به دلیل داشتن ضریب همبستگی 0/81 و خطای 2/06 همواره برآورد بهتری از تبخیر سطح آزاد آب ارایه می‌دهد.

واژه‌های کلیدی: تبخیر، درخت تصمیم M5، مدل KNN، معادلات تجربی

مقدمه

تبخیر یک پدیده‌ی غیرخطی و پیچیده است، زیرا اول: بستگی به عوامل اقلیمی مختلفی دارد و دوم: این عوامل بر روی یکدیگر تاثیر می‌گذارند. بنابراین تهیه‌ی یک شبیه ریاضی برای آن با در نظر گرفتن تمام عوامل اقلیمی موثر در آن، کاری دشوار بوده و در صورت امکان، با خطاهای قابل‌توجهی روبرو است، یا نیاز به اطلاعات زیادی دارد که اندازه‌گیری آن‌ها مشکل و زمان‌بر است (Jain et al., 1999). اهمیت تبخیر-تعرق در چرخه‌ی هیدرولوژی ازآن‌جا مشخص می‌شود که در مقیاس جهانی حدود 57 درصد آبی که روی خشکی‌ها به‌صورت نزولات جوی فرو می‌ریزد مستقیما تبخیر می‌شود (علیزاده، 1383). برای پیش‌بینی میزان تبخیر بر اساس شاخص‌های آب و هوایی مدل‌های مختلفی وجود دارد و پژوهشگران شیوه‌های گوناگونی را به کار گرفته تا نتایج قابل قبولی را برای نقاط مختلف دنیا به دست آورند (رحمتی و همکاران، 1394). قمرنیا و همکاران (1391) به ارزیابی و واسنجی مدل‌های تبخیر و تعرق گیاه مرجع با توجه به اثر دوره محاسباتی برای اقلیم نیمه‌خشک سرد پرداختند. نتایج نشان داد مدل پنمن اصلاح‌شده در تمامی دوره‌های محاسباتی، مقدار تبخیر و تعرق گیاه مرجع را با دقت بالاتری پیش‌بینی می‌کند و واسنجی این معادله نیز تاثیری برافزایش دقت آن نداشت. روش‌های داده‌کاوی، روش‌های مدل کردن رابطه‌ی نهفته در داده‌ها هستند که به‌صورت خودکار به دسته‌بندی مجموعه داده‌ها (معمولا مجموعه‌های بزرگ) و کشف ارتباط نهفته در بین آن‌ها به‌منظور قابل فهم شدن و درنتیجه سودمند شدن آن‌ها می‌پردازند (Hand et al., 2001). پرکاربردترین این مدل‌ها عبارت‌اند از: شبکه‌های عصبی مصنوعی، درخت تصمیم، سیستم‌های استنتاج فازی و ماشین‌های بردار پشتیبان. درختان تصمیم ابزاری است که قابلیت پاسخ‌گویی به مسایل پیچیده و غیرخطی را دارد (طالبی، 1392). در ساختار درخت تصمیم، پیش‌بینی به‌دست‌آمده در قالب یک سری قواعد توضیح داده خواهد شد (مشکانی و ناظمی، 1388). شریفان و قربانی (1393) برای بهبود برآورد تبخیر - تعرق پتانسیل با استفاده از ضریب اصلاحی از مدل درخت تصمیم M5 بهره بردند. سامتی و همکاران (1390) نتایج برآورد تبخیر - تعرق در ایستگاه هواشناسی شیراز را به کمک مدل درختی M5 و هارگریوز سامانی با نتایج روش پنمن- مانتیث مقایسه کردند و نتیجه گرفتند که مدل درختی M5 نسبت به روش هارگریوز سامانی تطابق بهتری با روش پنمن-مانتیث دارد. ستاری و همکاران (1393) از روش تصمیم‌گیری درختی M5 و شبکه عصبی مصنوعی جهت پیش‌بینی بارش ماهانه ایستگاه اهر استفاده نمودند.

1- دانشجوی دکتری مهندسی علوم آب، دانشگاه بیرجند
2- دانشیار گروه مهندسی علوم آب، دانشگاه بیرجند
3- دانشیار گروه مهندسی علوم آب، دانشگاه بیرجند
(* - نویسنده مسئول: Email: Atenakhalili_2014@yahoo.com)

استفاده از تکنیک‌های غیر پارامتریک کارایی چشم‌گیری در بهبود تخمین‌های صورت گرفته خواهد داشت (Twarakavi et al., 2009). از روش‌های غیر پارامتری برای تولید داده‌های آب و هوایی، روش خودگردان ساز KNN است که اخیرا برای تولید مصنوعی داده‌های آب و هوایی استفاده شده است. رویکرد k–نزدیک‌ترین همسایه، یکی از مهم‌ترین و توسعه یافته‌ترین رویکردهای غیر پارامتریک می‌باشد که در بسیاری از پژوهش‌های نوین جهت تشخیص الگو و کلاسه‌بندی‌های آماری به کار گرفته شده‌است (شریف آذری، 1392). سینگ و همکاران از مدل نزدیک‌ترین همسایه برای پیش‌بینی وضعیت آب و هوایی در یک ایستگاه در شمال هند در روزهای برفی و غیربرفی استفاده نمودند و نتیجه گرفتند که پیش‌بینی روزانه ریزش برف دارای دقت بالایی است(Singh and et al., 2005). در شرایطی که هیچ دانشی از معادلاتی که رفتار سامانه را مشخص سازد در دسترس نباشد، آزمون گاما می‌تواند به‌عنوان ابزاری به ساخت مدلی هموار از رفتار سیستمی آن پدیده بر اساس سری داده‌های اندازه‌گیری شده به‌کار رود. آزمون گاما یک شبیه‌سازی غیرخطی و ابزاری جهت آنالیز بوده و اجازه تجزیه و تحلیل می‌دهد تا رابطه‌ی بین ورودی‌ها و خروجی‌ها در یک مجموعه، داده‌های عددی مورد امتحان قرارگیرد (Agalbjorn and et al., 1997). مقدم نیا و همکاران در برآورد تبخیر، ریمسان و همکاران، احمدی و همکاران، مقدم نیا و همکاران در برآورد تابش خورشیدی و قبائی سوق و همکاران در برآورد تبخیر و تعرق از این آزمون استفاده نموده‌اند (Remesan and et al., 2008); (Ahmadi and et al., 2009); (Moghaddamnia and et al., 2009); (Ghabaei sough and et al., 2010).

در این مطالعه ابتدا عملکرد دو مدل درخت تصمیم M5 و مدل KNN در پیش‌بینی تبخیر از سطح آزاد آب تحت سناریوهای مختلف هواشناسی بررسی گردید. سپس نتایج این دو مدل با نتایج حاصل از معادلات تجربی با استفاده از معیارهای آماری مناسب مقایسه شد و در نهایت با استفاده از آزمون گاما و مدل درخت تصمیم و KNN بهترین سناریو برای پیش‌بینی تبخیر پیشنهاد گردید.

مواد و روش‌ها

ساختار درخت تصمیم M5: روش درخت تصمیم یک روش سلسه مراتبی یا چند مرحله‌ای است که در آن به‌صورت بازگشتی مجموعه داده‌ها به روش دودویی به تقسیمات فرعی و کوچک‌تر تقسیم‌بندی می‌شوند تا زمانی که تقسیمات فرعی نهایی نتوانند بیش‌تر از آن تجزیه شوند (طالبی، 1392). درختان تصمیم مجموعه‌ای از داده‌های معلوم را می‌گیرند و یک درخت تصمیم را از آن استنتاج می‌کنند. سپس درخت می‌تواند به‌صورت مجموعه قوانینی برای پیش‌بینی ویژگی‌های معلوم استفاده شود. مجموعه داده‌های اولیه که درخت به‌وسیله آن‌ها ایجاد می‌شود به‌عنوان داده‌های آموزشی شناخته می‌شوند. درخت تصمیم از بالا به پایین ترسیم می‌شود. در بالا ویژگی اول و مقدارش قرار می‌گیرد و از آن به بعد شاخه یا منجر به یک ویژگی و یا منجر به یک نتیجه می‌شود (یوسفی و همکاران، 1393).

یک درخت تصمیم معمولا از چهار بخش ریشه، شاخه، گره‌ها و برگ‌ها تشکیل‌شده است. گره اول در درخت تصمیم به‌عنوان ریشه‌ی درخت در نظر گرفته می‌شود. هر گره مربوط به یک خصوصیت معین است و شاخه‌ها به معنای بازه‌ای از مقادیر هستند. این بازه‌های مقادیر باید بخش‌های مختلف مجموعه‌ی مقادیر معلوم را برای خصوصیت‌ها به دست دهند. عمل انشعاب توسط یکی از متغیرهای پیش‌بینی کننده انجام می‌پذیرد (Alberg et al., 2012).

معیار تقسیم برای الگوریتم مدل M5 ارزیابی انحراف معیار مقادیر کلاسی است که به‌عنوان کمیتی از خطا به یک گره می‌رسد و کاهش مورد انتظار در این خطا را به‌عنوان نتیجه‌ی آزمون هر صفت در آن گره محاسبه می‌نماید (شریفان، 1393). کاهش انحراف معیار (SDR) [1] از رابطه 1 به دست می‌آید (Quinlan., 1992).

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \qquad (1)$$

که T بیانگر یک سری نمونه‌هایی است که به گره می‌رسد، Ti بیانگر نمونه‌هایی است که i امین خروجی سری پتانسیلی را دارند و sd بیانگر انحراف معیار است. گره‌های فرزند به علت انحراف معیار کم‌تر دارای دقت بیش‌تری نسبت به گره‌های والد هستند.

الگوریتم k-نزدیک‌ترین همسایه (KNN): این روش یک الگوریتم یادگیری بر اساس مشاهدات و نمونه‌ها می‌باشد (شیرزاد، 1387). در این روش تابع توزیع مقادیر پیش‌بینی با استفاده از توزیع ناپارامتری تابع کرنل به‌دست می‌آید. مفهوم مورد استفاده در این روش به این شرح است که با مشاهده متغیرهای مستقل در زمان واقعی، مدل به جستجوی الگوهای مشابه شرایط فعلی در سری تاریخی می‌پردازد. وقایعی که در سری تاریخی در این الگوها پیش‌آمده‌اند می‌توانند به‌عنوان گزینه‌های محتمل در شرایط فعلی در نظر گرفته شوند. احتمال وقوع هر یک از این حالت‌ها در شرایط حاضر، بستگی به شباهت بردار متغیرهای مستقل فعلی با بردار متغیرهای مستقل مشاهداتی در سری تاریخی دارد. دو عامل مهم در به‌کارگیری روش KNN، تعداد همسایه‌ها (k) و وزن پیش‌بینی کننده‌ها (wj) هستند (Tarboton etal., 1993). براساس توضیحات ارایه شده، KNN یک روش تشخیص الگوی آماری بدون متغیر است که برای

1- standard deviation reduction

الگوی هیدرولوژیکی موجود، k الگوی مشابه به نام نزدیک‌ترین همسایه‌ها را می‌یابد. برای انتخاب k نمونه مشابه از فاصله‌ی اقلیدسی استفاده می‌شود، بدین‌صورت که هر نمونه از بانک داده که با نمونه هدف (مجهول) کم‌ترین فاصله یا به‌عبارت‌دیگر بیش‌ترین تشابه از نظر ویژگی را داشته‌باشد انتخاب و وزن دهی می‌شود. جاگتاپ و همکاران رابطه‌ی فیثاغورث را در محاسبه‌ی فاصله‌ی اقلیدسی توصیه کردند(Jagtap and et al., 2004).

$$D(X,Y)=\sqrt{\sum_{i-1}^{nf}(X_i-Y_i)^2} \quad (2)$$

که در آن X نماینده نمونه‌ای از داده‌ها با چند پارامتر مشخص (x_1 تا x_n) در بانک مرجع و Y نمونه داده هدف با همان تعداد پارامتر (y_1 تا y_n) می‌باشد.

$$X=(x_1,x_2,x_3,...,x_n) \quad (3)$$

$$Y=(y_1,y_2,y_3,...,y_n) \quad (4)$$

پس از تعیین k دسته از نزدیک‌ترین همسایگی‌ها، مدل به راحتی قادر خواهدبود ترکیب وزنی آن همسایگی‌ها را به عنوان پیش‌بینی نمونه مورد بررسی انتخاب کند (شیرزاد، 1387).

معادلات تجربی: کوشش‌های زیادی به‌عمل‌آمده است تا فرمول‌ها و معادله‌های عملی و ساده‌ای برای تخمین تبخیر از سطح آزاد آب ارایه شود. ازجمله فرمول‌های ارایه شده که در کارهای هیدرولوژی از آن‌ها استفاده می‌شوند عبارت‌اند از فرمول‌های مایر، هفنر، شاهتین، مارسیانو، ایوانف، تیچومیروف و USBR که به ترتیب در ادامه آورده شده است (روابط 12-5):

$$E=(1+\frac{U}{16}).C.(e_s-e_a) \quad (5)$$

$$E=0.028U(e_s-e_a) \quad (6)$$

$$E=(0.116+0.017U)(e_s-e_a) \quad (7)$$

$$E=0.03U(e_s-e_a) \quad (8)$$

کمبود فشار بخار با در دسترس داشتن متوسط دمای روزانه (T) و متوسط رطوبت نسبی (RH) از معادله‌ی 9 قابل محاسبه است.

(9)

$$e_s-e_d=[\exp(\frac{16.78T-116.9}{T+237.3})](1-RH/100)$$

که در این فرمول e_s: فشار بخار اشباع (میلی‌متر جیوه)، e_a: فشار بخار واقعی (میلی‌متر جیوه)، C: ضریب برای دریاچه عمیق 0/36 و برای دریاچه کم‌عمق 0/5، U: سرعت باد (کیلومتر در ساعت در ارتفاع 2 متر)، T: متوسط دمای روزانه (درجه سلسیوس)، RH: متوسط رطوبت نسبی (درصد) و E: تبخیر (میلی‌متر در روز) می‌باشد.

$$E=0.0018.(T+25)^2*(100-RH) \quad (10)$$

$$E=(e_s-e_a)*(15+3U_{10}) \quad (11)$$

$$E=0.833*(4.57T+43.3) \quad (12)$$

که در این روابط U: سرعت باد (متر در ثانیه در ارتفاع 10متر)، T: متوسط دمای ماهانه (درجه سلسیوس)، RH : متوسط رطوبت نسبی (درصد) و E: تبخیر (میلی‌متر در ماه) می‌باشد (علیزاده، 1383).

آزمون گاما: آزمون گاما (G.T) یک ابزار مدل‌سازی غیرخطی است که به کمک آن می‌توان حداقل میانگین مربعات خطایی را که توسط مدل هموار می‌تواند برآورد گردد، محاسبه نمود. همچنین با استفاده از آن می‌توان ترکیب مناسب از پارامترهای ورودی برای مدل‌سازی داده‌های خروجی و تعداد داده‌های لازم برای ایجاد یک مدل هموار را تعیین نمود (Moghaddamnia and et al., 2009). در واقع با استفاده از آزمون گاما می‌توان بهینه‌ترین ترکیب ورودی‌ها و مناسب‌ترین تعداد داده‌ها که منجر به کم‌ترین میانگین مربعات خطا در هرگونه مدل‌سازی غیرخطی پیوسته می‌شود را تعیین نمود (بسالت پور، 1392).

منطقه مورد مطالعه: در این تحقیق از اطلاعات ایستگاه هواشناسی همدیدی بیرجند که در سال 1334 خورشیدی (1955 میلادی) راه‌اندازی گردید، استفاده شد. این ایستگاه در نیمه‌ی جنوبی استان خراسان در ارتفاع 1491 متری قرار گرفته است و دارای طول جغرافیایی 59 درجه و 12 دقیقه و عرض جغرافیایی 32 درجه و 52 دقیقه می‌باشد.

جمع‌آوری داده‌ها و تجزیه و تحلیل داده‌ها: در این مطالعه از داده‌های اقلیمی جهت بیشینه باد (درجه)، بیشینه سرعت باد (متر بر ثانیه)، متوسط سرعت باد (متر بر ثانیه)، ساعت آفتابی (ساعت)، جمع بارندگی (میلی‌متر)، متوسط رطوبت (درصد)، متوسط رطوبت نسبی (درصد)، درجه حرارت بیشینه (درجه سانتی‌گراد)، درجه حرارت کمینه (درجه سانتی‌گراد) و متوسط دمای ماهانه مربوط به ایستگاه (درجه سانتی‌گراد)، برای پیش‌بینی تبخیر (میلی‌متر) استفاده شد. به این منظور از داده‌های روزانه ایستگاه هواشناسی بیرجند از مهر ماه سال 1387 تا مهر سال 1392 جمع‌آوری گردید. برای استفاده از داده‌های مذکور در معادلات تجربی، مدل KNN و درخت تصمیم M5، ابتدا داده‌ها به دو دسته تقسیم شدند. 75 درصد داده‌ها برای آموزش و 25 درصد باقیمانده برای آزمون بکار گرفته شد. خصوصیات آماری پارامترهای مورد استفاده در این تحقیق در جدول 1 آمده است.

جدول 1- خصوصیات آماری پارامترهای مورد استفاده

پارامتر	کمینه	بیشینه	میانگین	انحراف معیار	ضریب تغییرات
جهت بیشینه باد (درجه)	10	360	201	110/78	0/55
بیشینه سرعت باد (متر بر ثانیه)	2	17	7/33	1/62	0/22
متوسط سرعت باد (متر بر ثانیه)	0/38	7/88	3/24	1/04	0/32
ساعت آفتابی (ساعت)	0	13/5	9/74	2/15	0/22
جمع بارندگی (میلی‌متر)	0	27/3	0/32	0/59	1/85
متوسط رطوبت (درصد)	8	93	32/48	12/74	0/39
متوسط رطوبت نسبی (درصد)	5	95	29/01	12/55	0/43
درجه حرارت کمینه (درجه سانتی‌گراد)	-11/80	27/8	11/05	6/15	0/56
درجه حرارت بیشینه(درجه سانتی‌گراد)	2/6	42/6	27/65	6/81	0/25
متوسط دمای ماهانه (درجه سانتی‌گراد)	3/27	29/71	18/41	6/56	0/36
تبخیر (میلی‌متر)	0	21/8	8/98	4/34	0/48

قبل از استفاده از داده‌ها به‌منظور افزایش کارایی مدل‌ها، همه‌ی داده‌ها به شکل نرمال بین دو عدد 0/1 و 0/9 طبق فرمول شماره 13 استاندارد شدند (Rahimi Khoob., 2008) و آن‌ها بعد از استفاده در شبیه‌سازی به مقادیر اولیه برگشتند.

$$x_i = 0.8\left(\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) + 0.1 \qquad (13)$$

در این رابطه x_i مقدار استانداردشده، x مقدار واقعی و x_{min} و x_{max} به ترتیب مقادیر کمینه و بیشینه داده‌ها می‌باشند.

مقادیر تبخیر از تشت تبخیر به‌صورت روزانه برداشت و برای سنجش دقت سایر روش‌ها استفاده شد. تشت تبخیر استفاده شده در تحقیق حاضر کلاس A است. به‌منظور پردازش و اعمال کلیه روش‌های ریاضی، از امکانات و توابع موجود در نرم‌افزار اکسل استفاده گردید. در مطالعه‌ی حاضر از 7 روش تجربی تخمین تبخیر ذکر شده در قسمت قبل برای ارزیابی نتایج تشت تبخیر استفاده شد. همان‌طور که در معادلات تجربی مشاهده می‌شود تمام روش‌ها به‌جز روش سازمان عمران اراضی آمریکا از 2 یا 3 پارامتر استفاده کردند، درحالی‌که روش سازمان عمران اراضی آمریکا فقط از یک پارامتر برای محاسبه تبخیر از سطح آزاد بهره گرفته است. از این‌رو به‌منظور مقایسه نتایج حاصل از معادلات تجربی، الگوریتم KNN و مدل درختی M5 در پیش‌بینی میزان تبخیر، سه سناریو با توجه به پارامترهای مورداستفاده در معادلات تجربی تعریف شد. در جدول 2 پارامترهای ورودی در هر سناریو ارایه گردید. برای بهره‌گیری از الگوریتم KNN و مدل درختی M5 از نرم‌افزار Weka 3.7 استفاده شده است.

انتخاب ورودی‌های بهینه با استفاده از آزمون گاما:

آزمون گاما مقدار خطای برآورد شده (واریانس خطا) را با توجه به داده‌های مستقیم نشان می‌دهد (بسالت پور، 1392). وقتی‌که مقدار گاما صفر باشد هیچ محدودیتی برای ساخت یک مدل خوب وجود نخواهد داشت، بنابراین از نقاطی که دارای مقدار گامای کم‌تر در مقایسه با دیگر متغیرها هستند برای مدل‌سازی استفاده خواهد شد (Moghaddamnia and et al., 2009). برای استفاده از آزمون گاما 10سناریو تعریف شد تا بهترین ترکیب ورودی از پارامترها برای پیش‌بینی تبخیر از تشت تبخیر انتخاب گردد. جدول 3 ترکیب پارامترهای ورودی به آزمون گاما، نشان می‌دهد.

جدول2- پارامترهای ورودی در سناریوهای مختلف

سناریو	پارامترهای ورودی
a	U2, T, RH, E
b	T, RH, E
c	T,E

U2: سرعت باد در ارتفاع 2 متری از سطح زمین (کیلومتر بر ساعت) T: متوسط دمای روزانه RH: درصد رطوبت نسبی E: تبخیر از تشت تبخیر (میلی‌متر)

معیارهای ارزیابی مدل: عملکرد معادلات تجربی، الگوریتم KNN و درخت تصمیم M5 توسط آماره‌های ریشه متوسط خطای مربعات (RMSE)، ضریب همبستگی (R2) و متوسط قدرمطلق‌خطا (MAE) ارزیابی و از طریق رتبه‌بندی آماره‌ها و نزولی کردن آن‌ها در نرم‌افزار اکسل، بهترین رتبه انتخاب گردید.

$$RMSE = \sqrt{\frac{(\sum (E_{si} - E_{oi})^2)}{n-1}} \qquad (14)$$

$$R^2 = \frac{(\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}))^2}{\frac{1}{N}\sum_{i=1}^{n} (E_{si} - E_{oi})} \qquad (15)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n} (E_{si} - E_{oi}) \qquad (16)$$

N برابر با تعداد کل داده‌ها ، E_{si} سطح آب تخمین زده‌شده، E_{oi} داده مشاهده‌ای با $\overline{X}$ و $\overline{Y}$ متوسط مقادیر X_i و Y_i هستند.

جدول3- ترکیب پارامترهای ورودی به آزمون گاما

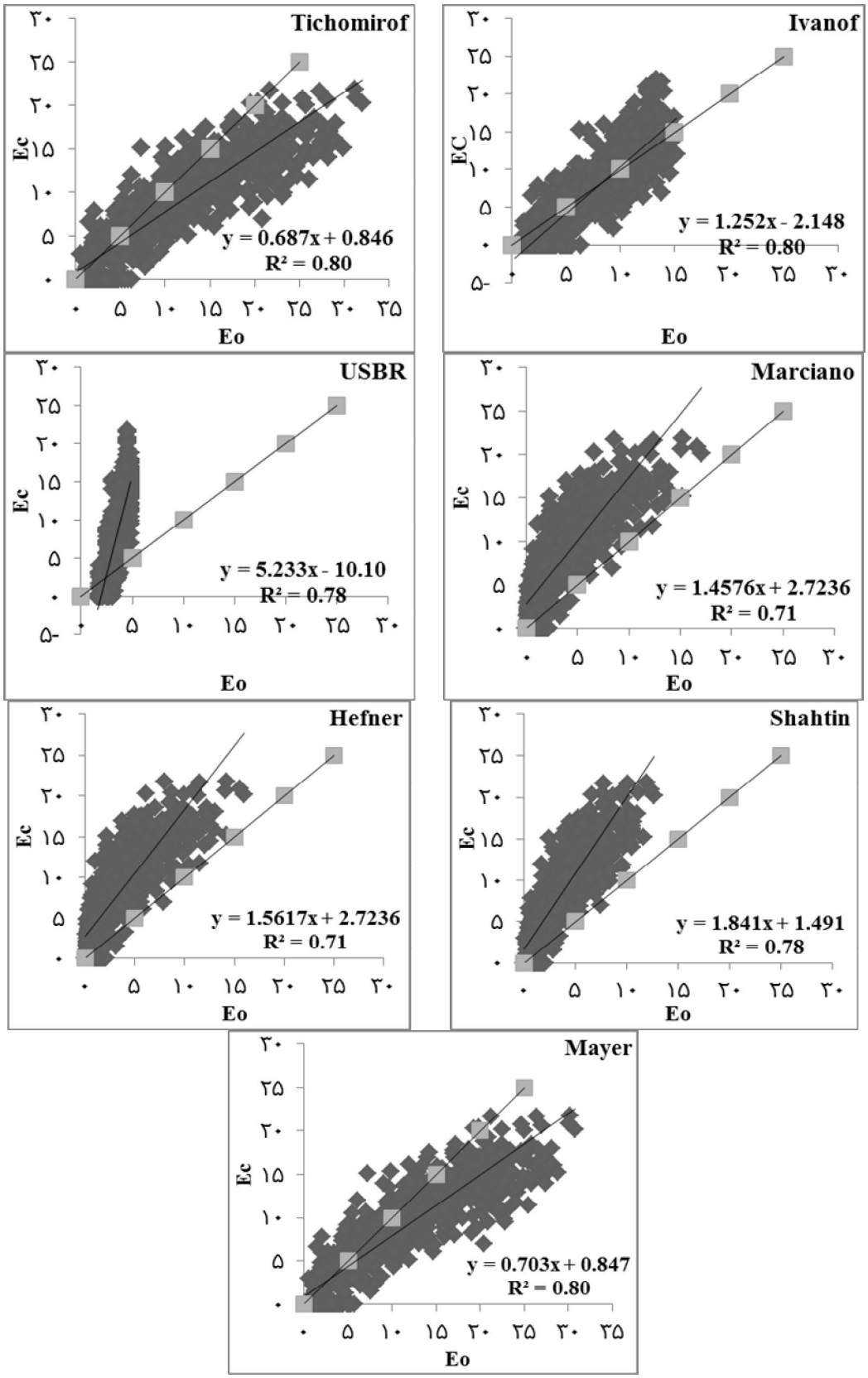
سناریو	پارامترهای ورودی
1	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
2	بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
3	متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
4	ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
5	جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
6	متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
7	متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
8	درجه حرارت کمینه، درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
9	درجه حرارت بیشینه، متوسط دمای ماهانه، تبخیر
10	متوسط دمای ماهانه، تبخیر
11	جهت بیشینه باد، تبخیر
12	جهت بیشینه باد، بیشینه سرعت باد، تبخیر
13	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، تبخیر
14	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، تبخیر
15	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، تبخیر
16	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، تبخیر
17	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، تبخیر
18	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، تبخیر
19	جهت بیشینه باد، بیشینه سرعت باد، متوسط سرعت باد، ساعت آفتابی، جمع بارندگی، متوسط رطوبت، متوسط رطوبت نسبی، درجه حرارت کمینه، درجه حرارت بیشینه، تبخیر

نتایج و بحث

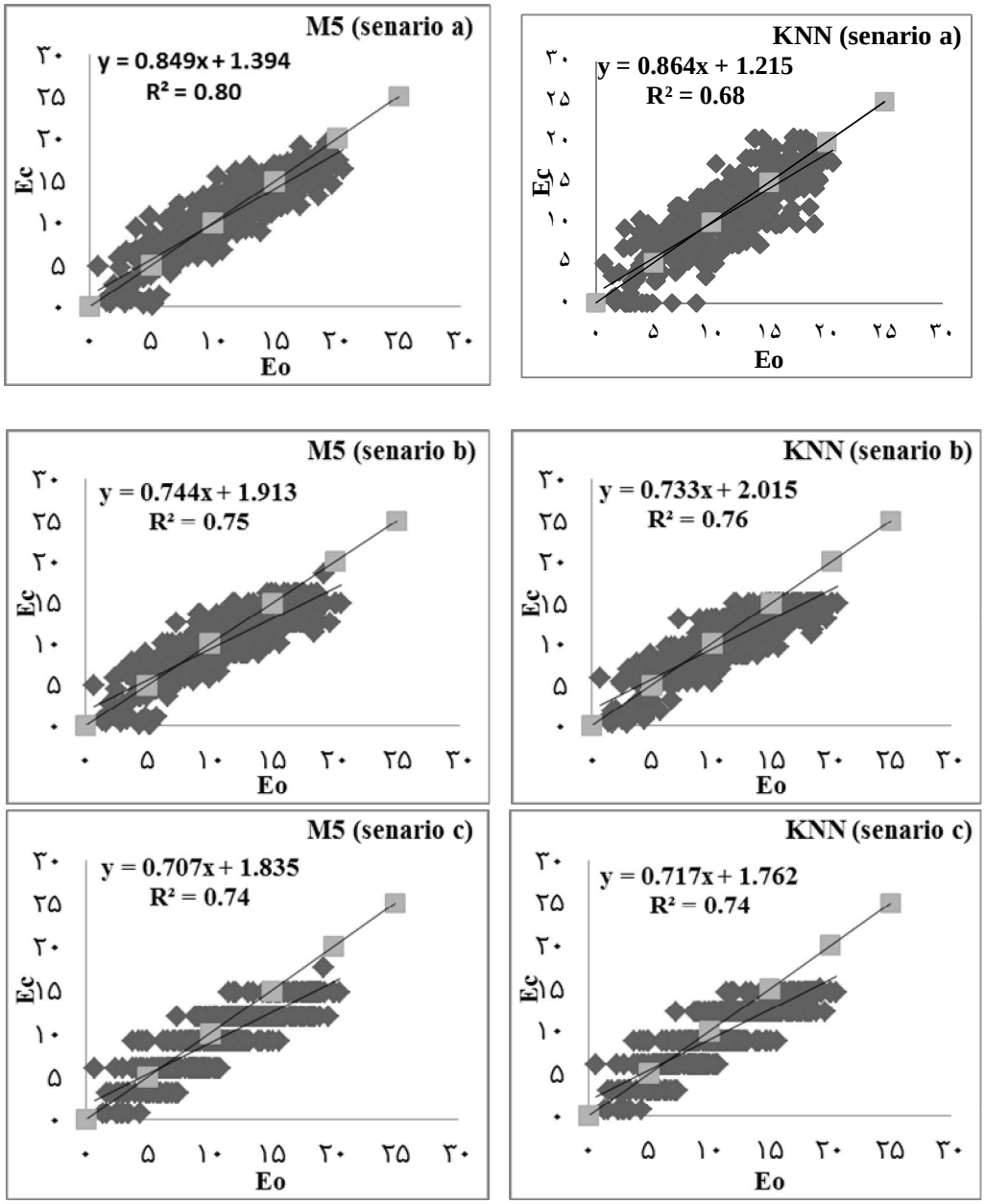
در این قسمت ابتدا فرمول‌های تجربی محاسبه تبخیر در منطقه مورد مطالعه بررسی شد و سپس به مقایسه نتایج الگوریتم KNN و درخت تصمیم M5 با معادلات اصلاح‌شده پرداخته شد.

مقادیر تبخیر هر یک از معادلات تجربی، با استفاده از داده‌های مرحله‌ی آموزش محاسبه گردید و بین مقادیر تبخیر محاسبه شده از معادلات تجربی و تشت تبخیر معادله رگرسیونی برازش داده شد و فرمول‌های تجربی برای منطقه بیرجند اصلاح گردید. خط برازش داده‌شده بین مقادیر تبخیر مشاهده شده از تشت تبخیر (Eo) و تبخیر محاسبه شده از معادلات تجربی (Ec) و فرمول‌های اصلاح‌شده در شکل 1 آمده است. خطوط و معادلات مربوط به آن‌ها نشان‌دهنده رابطه نسبتا مناسبی می‌باشند. رابطه USBR دارای شیب زیاد و فرمول مایر دارای شیب کم‌تری نسبت به خط 1:1 می‌باشند. از طرفی فرمول مایر دارای همبستگی بیش‌تری هست که این بیانگر این است که فرمول مایر در برآورد تبخیر از سطح آزاد در منطقه بیرجند مناسب‌تر می‌باشد.

با استفاده از این معادلات، مقدار تبخیر برای داده‌های مرحله آموزش به دست آورده شد و آماره‌های همبستگی و خطا محاسبه گردید. نتایج در جدول 4 آمده است. همان‌طور که در قسمت قبل نیز بیان شد، در بین معادلات تجربی فرمول مایر به دلیل داشتن R^2 بیش‌تر و RMSE و MAE کم‌تر برای پیش‌بینی تبخیر از سطح آزاد آب بهتر می‌باشد. الگوریتم KNN و درخت تصمیم M5 تحت 3 سناریو مورد بحث در معادلات تجربی در نرم‌افزار Weka اجرا گردید و شاخص‌های آماری به‌دست آمده از دو مدل مذکور در جدول 5 و 6 نمایش داده شد. همان‌طور که در جداول مشاهده می‌شود سناریوهای b و c در مدلKNN دارای ضریب همبستگی و خطای مطلوب‌تری نسبت به مدل درختی M5 می‌باشند که این نشان می‌دهد کارایی مدل KNN با توجه به ورودی تبخیر و رطوبت نسبی به‌دلیل داشتن R^2 برابر 0/76 و RMSE برابر با 2/35 بهتر می‌باشد.



شکل1- فرمول‌های تجربی اصلاح‌شده در منطقه بیرجند



شکل2- مقادیر تبخیر پیش‌بینی‌شده تحت تاثیر تبخیر مشاهده‌شده

اما با در نظر گرفتن پارامتر سرعت باد توانایی مدل درختی M5 در پیش‌بینی تبخیر بیش‌تر می‌شود. با توجه به این‌که معادلات هفنر، شاهتین، مارسیانو، تیچومیروف و مایر تحت تاثیر سناریو a قرار دارند، با رتبه‌بندی آماره‌های این معادلات و دو مدل KNN و درخت تصمیم M5 مشخص شد که در تمامی این معادلات به غیر از معادله مایر، مدل درختی M5 پیش‌بینی مطلوب‌تری را حاصل می‌نماید و مدل KNN حتی نسبت به معادلات تجربی، نتایج ضعیف‌تری را نشان می‌دهد. حال‌آنکه فرمول مایر اصلاح‌شده برای منطقه بیرجند دارای همبستگی (0/81) بیش‌تری می‌باشد. این بیانگر آن است که فرمول مایر در برآورد تبخیر از سطح آزاد در منطقه بیرجند مناسب‌تر می‌باشد. در مورد معادلات USBR و ایوانف که به ترتیب تحت تاثیر سناریو c و b می‌باشند مدل KNN پیش‌بینی مناسب‌تری را ارایه می‌دهد. با توجه به مطالب فوق می‌توان به این نتیجه رسید که زمانی که سرعت باد برای پیش‌بینی تبخیر در نظر گرفته شود، مدل درختی M5 و در غیر این صورت مدل KNN نسبت به معادلات تجربی نتایج دقیق‌تری را حاصل می‌کنند. همچنین در بین تمامی معادلات و الگوریتم‌ها صرف‌نظر از پارامترهای موثر بر آن، مدل درخت تصمیم M5در سناریو a که تحت تاثیر پارامترهای سرعت باد، دمای هوا،

رطوبت نسبی و تبخیر می‌باشد، برای پیش‌بینی تبخیر پیشنهاد می‌شود. در ادامه خروجی مدل درخت تصمیم M5 تحت سناریو a آورده‌شده است.

```
T <= 18.05 :
|T <= 9.7
|   |T <= 6.05: LM1 (88/18.122%)
|   |T >  6.05
|   |   |RH <= 43.65 :
|   |   |   |U <= 5.305: LM2 (18/17.127%)
|   |   |   |U >  5.305: LM3 (13/23.776%)
|   |   |RH >  43.65: LM4 (79/30.103%)
|T >  9.7: LM5 (285/38.877%)
T >  18.05: LM6 (581/42.314%)
LM num1: E = 0.0068*U + 0.0656*T- 0.0069*RH+ 0.4281
LM num 2: E = -0.0873*U + 0.13*T- 0.0495*RH+ 3.4364
LM num 3: E = -0.1041*U + 0.13*T- 0.0125*RH+ 1.5778
LM num 4: E = 0.0068*U + 0.095*T- 0.0095*RH+ 0.8108
LM num 5: E = 0.0068*U + 0.3591*T- 0.0755*RH+ 2.9726
LM num 6: E = 0.3879*U + 0.0937*T- 0.0937*RH- 0.1763
```

جدول4- نتایج معادلات تجربی اصلاح‌شده منطقه بیرجند در پیش‌بینی تبخیر

MAE	RMSE	R^2	معادله
1/63	2/06	0/81	Mayer
1/94	2/53	0/74	Hefner
1/72	2/21	0/8	Shahtin
1/94	2/54	0/74	Marciano
2/15	2/68	0/71	USBR
1/91	2/36	0/72	Ivanof
1/66	2/1	0/81	Tichomirof

جدول5- نتایج مدل KNN در پیش‌بینی تبخیر

MAE	RMSE	R^2	سناریو
2/07	2/7	0/67	A
1/93	2/35	0/76	B
2/18	2/63	0/74	C

در شکل 2 مقادیر برآورد شده در مقابل مقادیر مشاهده شده تبخیر از سطح آب تحت 3 سناریو مختلف نشان داده شد. همان‌طور که مشاهده می‌شود در مدل درخت تصمیم M5 سناریو a دارای بیش‌ترین مقدار ضریب همبستگی، بیش‌ترین شیب‌خط و نزدیک‌ترین حالت به خط 1:1می‌باشد. این نتایج حاکی از مطلوب بودن این مدل و سناریو در مقایسه با سایرین می‌باشد.

جدول6- نتایج مدل درختی M5 در پیش‌بینی تبخیر

MAE	RMSE	R^2	سناریو
1/59	1/97	0/8	A
1/94	2/37	0/75	B
2/19	2/65	0/74	C

جدول7- نتایج آزمون گاما

رتبه	سناریو	پارامتر			
		گاما	گرادیان	انحراف معیار	نسبتv
1	19	0/0269	0/0345	0/0023	0/1074
2	1	0/0283	0/0253	0/0013	0/1132
3	18	0/0311	0/0355	0/0014	0/1244
4	4	0/0316	0/063	0/0012	0/1267
5	2	0/032	0/0209	0/002	0/128
6	3	0/0329	0/0196	0/008	0/1318
7	7	0/0384	0/175	0/0013	0/1537
8	6	0/0389	0/0955	0/002	0/1558
9	5	0/0406	0/0312	0/0013	0/1625
10	9	0/0445	0/8099	0/0022	0/1781
11	8	0/046	0/1495	0/002	0/1843
12	10	0/049	0/2506	0/0012	0/1962
13	17	0/0578	0/0407	0/0019	0/2314
14	16	0/0626	0/0407	0/0024	0/2508
15	14	0/0967	0/0932	0/004	0/3871
16	15	0/1001	0/0381	0/0047	0/4006
17	13	0/1309	0/3259	0/0042	0/5237
18	12	0/1641	0/2855	0/0162	0/6569
19	11	0/2133	0/4303	0/0125	0/8533

نتایج آزمون گاما:

در این تحقیق با در نظر گرفتن یازده پارامتر ورودی موثر بر تبخیر، تعداد 19 ترکیب مختلف ایجاد شد که با استفاده از آزمون گاما و بررسی آماره‌های آن، ترکیبی که دارای مقدار آماره گاما، گرادیان، انحراف معیار و نسبت V کم‌تر بود برای برآورد تبخیر از سطح آزاد آب انتخاب شد. نتایج آزمون گاما و آماره‌های آن در جدول 7 ارایه شده است. در این جدول ترکیب‌هایی ورودی بر اساس کم‌ترین میزان آماره گاما رتبه‌بندی شده‌اند. نتایج نشان می‌دهد ترکیبی که دارای همه پارامترهای ورودی به‌جز دمای متوسط ماهانه باشد، آماره‌های

پایین‌تری دارد و بهترین برآورد از سطح آزاد را حاصل می‌کند و بدترین پیش‌بینی زمانی رخ می‌دهد که فقط پارامترهای جهت بیشینه باد و تبخیر مدنظر قرار گیرد.

الگوریتم KNN و درخت تصمیم M5 در حالت بهترین ترکیب پارامترهای ورودی انتخاب شده از آزمون گاما اجرا گردید و نتایج و آماره‌های این دو مدل در جدول 8 مورد تحلیل قرار گرفت. همان‌طور که مشاهده می‌شود در حالت بهترین ترکیب ورودی مدل درخت تصمیم M5 به‌دلیل داشتن همبستگی زیاد و خطای کم‌تر نتایج مطلوب‌تری را برآورد می‌کند.

جدول8- الگوریتم KNN و درخت تصمیم M5 در حالت بهترین ترکیب پارامترهای ورودی

مدل	R^2	RMSE	MAE
KNN	0/67	2/78	1/19
درخت تصمیم M5	0/83	1/75	1/39

نتیجه‌گیری

در بررسی‌ها مشخص شد با در نظر گرفتن پارامتر سرعت باد توانایی مدل درختی M5 در پیش‌بینی تبخیر بیش‌تر می‌شود. همچنین در بین تمامی معادلات و الگوریتم‌ها مدل درخت تصمیمM5 در سناریو a برای پیش‌بینی تبخیر پیشنهاد می‌شود. نتایج آزمون گاما نشان می‌دهد ترکیبی که دارای همه پارامترهای ورودی به‌جز دمای متوسط ماهانه باشد، بهترین برآورد و ترکیبی که فقط پارامترهای جهت بیشینه باد و تبخیر را مدنظر قرار دهد بدترین پیش‌بینی را ارایه می‌دهد.

منابع

بسالت‌پور،ع.ا، حاج عباسی،م،ع.، ایوبی،ش.ا. 1392. استفاده از آزمون گاما برای انتخاب ورودی‌های بهینه در مدل‌سازی مقاومت برشی خاک با استفاده از شبکه‌های عصبی مصنوعی. مجله پژوهش‌های حفاظت آب و خاک جلد بیستم. 1 : 97-114.

رحمتی،ع.، منتظری،م.، گندمکار،ا.، لشنی‌زند،م. 1394. پیش‌بینی تبخیر با استفاده از شبکه عصبی مصنوعی و سیگنال‌ها اقلیمی در حوضه دز. فصل‌نامه تحقیقات جغرافیایی. سال 30. 2. 117 :262- 274.

سامتی،م.، قهرمان،ن.، قربانی،خ. 1390. کاربرد مدل داده‌کاوی M5 در پیش‌بینی تبخیر - تعرق پتانسیل (مطالعه موردی :ایستگاه شیراز). اولین کنفرانس ملی هواشناسی و مدیریت آب کشاورزی،کرج، ایران.

ستاری،م،ت.، رضازاده جودی،ع.، نهرین،ف. 1393. پیش‌بینی مقادیر بارش ماهانه با استفاده از شبکه‌های عصبی مصنوعی و مدل درختی M5 (مطالعه موردی: ایستگاه اهر). پژوهش‌های جغرافیای طبیعی. سال 46. 2 : 247-260.

شریفان،ح.، قربانی،خ. 1393. بهبود برآورد تبخیر - تعرق پتانسیل با استفاده از ضریب اصلاحی به کمک مدل درخت تصمیم M5. نشریه آبیاری و زهکشی ایران. 8. 1 :53-61.

شریف آذری،س.، عراقی‌نژاد،ش. 1392. توسعه مدل ناپارامتری شبیه‌ساز داده‌های ماهانه هیدرولوژیکی. مجله مدیریت آب و آبیاری. 3. 1: 83-95.

شیرزاد،ا.، سلطانی،ف.، زارع ابیانه،ح. 1387. شبیه‌سازی آب‌شستگی در پایاب سازه‌های مستهلک کننده انرژی با استفاده از الگوریتم K - نزدیک‌ترین همسایگی (KNN) و سیستم استنتاج تطبیقی عصبی- فازی (ANFIS). اولین اجلاس بین‌المللی بحران آب. دانشگاه زابل.

طالبی،ع.، اکبری،ز. 1392. بررسی کارایی مدل درختان تصمیم‌گیری در برآورد رسوبات معلق رودخانه‌ای (مطالعه موردی: حوضه سد ایلام). مجله علوم و فنون کشاورزی و منابع طبیعی. علوم آب و خاک. سال 17. 63 : 109-121.

علیزاده،ا. 1383. اصول هیدرولوژی کاربردی. انتشارات دانشگاه امام رضا. ص291 و 237.

پنجمین کنفرانس سراسری آبخیزداری و مدیریت منابع آب و خاک کشور. کرمان.

قمرنیا،ه.، رضوانی،و.، فتحی،پ. 1391. ارزیابی و واسنجی مدل‌های تبخیر و تعرق گیاه مرجع با توجه به اثر دوره محاسباتی برای اقلیم نیمه‌خشک سرد. مجله مدیریت آب و آبیاری. 2. 2: 37-25.

مشکانی،ع.، ناظمی،ع. 1388. مقدمه‌ای بر داده‌کاوی. مشهد. موسسه چاپ و انتشارات دانشگاه فردوسی.

یوسف،م.، طالبی،ع.، پورشرعیاتی،ر. 1393. کاربرد هوش مصنوعی در علوم آب و خاک انتشارات دانشگاه یزد.

Agalbjorn,S., Konar,N and Jones,A.J. 1997. A note on the gamma test. Neural Comput ing & Applications. 5. 3: 131-133.

Ahmadi,A., Han,D., Karamouz,M and Remesan,R. 2009. Input data selection for solar radiation estimation. Hydrological Processes. 23: 2754-2764.

Alberg,D.,Last,M and Kandel,A. 2012. Knowledge discovery in data streams with regression tree methods. WIREs Data Mining Know Discover. 2: 69-78.

artificial neural networks and adaptive neurofuzzy inference system techniques. Advances in Water Resources. 32: 88-97.

Quinlan,J.R. 1992. Learning with continuous classes. In: Adams, Sterling (Eds.). Proceedings of AI'92. World Scientific. 343-348.

Remesan,R. Shamim,M.A. and Han,D. 2008. Model data selection using gamma testfor daily solar radiation estimation. Hydrological Processes. 22:4301-4309.

Rahimi Khoob,A. 2008. Artificial neural network estimation of reference evapo transpiration from pan evaporation in a semi-arid environment. Journal of Irrigation Science. 27.1: 35-39.

Singh,D. Ganju,A and Singh,A. 2005. Weather prediction using nearest-neighbor model. Current science. 88: 8. 25.

Twarakavi,N.K.C., Šimůnek,J and Schaap,M.G. 2009. Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. Soil oil Science Society of America Journal. 73:1443-1452.

Ghabaei sough,M., Mosaedi,A., Hesam,M and Hezarjaribi,A. 2010. Evaluation effect of input parameters preprocessing in artificial neural networks (ANNs) by using stepwise regression and Gamma Test techniques for fast estimation of daily evapotranspiration. Journal of Water and Soil. 24.3: 610-624.

Hand.D., Heikki,M and Padhraic,S. 2001. Principles of Data Mining. A Bradford book. The MIT press. Cambridge, Massachusetts, London, England.

Jagtap,S.S., Lall,U., Jones,J.W., Gijsman,A.J and Ritchie,J.T. 2004. A Dynamic nearest-neighbor method for estimating soil water parameters. Trans ASAE. 47:1437-1444.

Jain,S.K. Das,A and Srivastsva,D.k. 1999. Application of ANN for reservoir inflow prediction and operation. Journal of Water Resources Planning and Management. 125.5: 263-271.

Moghaddamnia,A., Remesan,R., Hassanpour Kashani,M., Mohammadi,M., Han,D and Piri,J. 2009. Comparison of LLR, MLP, Elman, NNARX and ANFIS Models with a case study in solar radiation estimation. Journal of Atmospheric and Solar-Terrestrial Physics. 71: 975-982.

Moghaddamnia,A., GhafariGousheh,M., Piri,J., Amin,S and Han,D. 2008. Evaporation estimation using

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 11, Agu.-Sept. 2017, p. 356-366

نشریه آبیاری و زهکشی ایران
شماره3، جلد 11، مرداد - شهریور 1396، ص. 366-356

Compare KNN and M5 Decision Tree Models in Anticipation of Evaporation and Comparison With Empirical Equations (Case Study of Birjand)

A.Khalili Naft Chali[1*], A. KHashei Siuki[2], A. Shahidi[3]
Recived: Jul.05, 2016 Accepted: Feb.12, 2017

Abstract

Evaporation from the soil surface is an important component of the water balance in the arid and semi-arid areas .In this study, to estimate the evaporation from open water surface, the empirical equations of Hefner, Shahtin, Marciano, Tichomirof, Ivanof, America Land Development Authority and Meyer have been used and modified in the study area. KNN model and M5 decision tree model were built ones data mining method to estimate the evaporation from open water surface. In so doing, data from meteorological stations in Birjand, Iran were used for finding the best effective input data on evaporation. Meanwhile, Gamma test was conducted to find the best combination of input parameters for estimating evaporation. To compare the results, the statistics of root mean square error (RMSE), correlation coefficient (R2) and mean absolute error (MAE) were analyzed. The results showed that M5 decision tree model reached the most optimum results than KNN model in most cases but the modified Meyer's equation for Birjand area represented better estimation of open water surface evaporation due to 0.81 correlation coefficient and 2.06 errors.

Keywords: Empirical equations, Evaporation, KNN model, M5 decision tree model.

1 - PhD student of water Science Engineering, University of Birjand
2 - Associate professor, Department of water Science Engineering, University of Birjand
3- Associate professor, Department of water Science Engineering, University of Birjand
(* - Corresponding Author Email: Atenakhalili_2014@yahoo.com)