نشریه پژوهش‌های حفاظت آب و خاک
جلد بیست و دوم، شماره چهارم، ۱۳۹٤
http://jwsc.gau.ac.ir

مقایسه عملکرد مدل‌های درختی و شبکه عصبی به‌منظور یافتن داده‌های گمشده تبخیر از تشت در استان خوزستان

مرجان وهابی‌مشهور[۱] و *علی رحیمی‌خوب[۲]

[۱]دانش‌آموخته کارشناسی‌ارشد گروه مهندسی آبیاری و زهکشی، پردیس ابوریحان، دانشگاه تهران،
[۲]استاد گروه مهندسی آبیاری و زهکشی، پردیس ابوریحان، دانشگاه تهران

تاریخ دریافت: ۹۲/۸/۸ ؛ تاریخ پذیرش: ۹۳/٦/۲۲

چکیده

سابقه و هدف: داده‌های تبخیر از تشت برای برآورد نیاز آبی گیاهان استفاده می‌شود ولی این داده‌ها در بعضی موارد به‌علت عدم دقت لازم در اندازه‌گیری‌ها و یا خرابی تجهیزات ناقص شده و به‌عنوان داده‌های گمشده از آن یاد می‌گردد. از آن‌جایی‌که پیوستگی داده‌ها برای برنامه‌ریزی آبیاری مهم است، بنابراین ضرورت دارد این نواقص آماری برطرف گردد. روش‌های زیادی برای یافتن داده‌های گمشده مورد استفاده قرار گرفتند در این میان مدل‌های درختی و شبکه‌های عصبی از دقت خوبی برخوردار بوده‌اند ولی تاکنون این دو روش مورد مقایسه و ارزیابی قرار نگرفتند. هدف از انجام این پژوهش مقایسه دو مدل درختی و شبکه‌های عصبی برای بازسازی داده‌های گمشده تبخیر روزانه ٤ ایستگاه هواشناسی استان خوزستان است.

مواد و روش‌ها: در این پژوهش داده‌های مورد نیاز از ٤ ایستگاه هواشناسی شامل آغاجاری، بندر ماهشهر، ایذه و بستان واقع دراستان خوزستان جمع‌آوری شد. اقلیم این ایستگاه‌ها براساس طبقه‌بندی دومارتن خشک می‌باشد. داده‌ها مربوط به سال‌های ۱۹۹۷ تا ۲۰۰۸ میلادی و شامل مقادیر روزانه تبخیر از تشت، سرعت باد، حداکثر و حداقل دمای هوا، رطوبت نسبی هوا، ساعات آفتابی و تابش برون‌زمینی بودند. این داده‌ها به دو دوره چهار ساله (۲۰۰۵ تا ۲۰۰۸) و ۱۲ ساله (۱۹۹۷ تا ۲۰۰۸) تقسیم شدند و در هر دوره پس از حذف عمدی ۵٪، ۱۰٪ و ۲۰٪ داده‌های اندازه‌گیری شده، مقادیر آن‌ها با استفاده از مدل‌های درختی و شبکه عصبی برآورد شدند و نتایج این دو مدل با استفاده از شاخص‌های آماری مورد مقایسه قرار گرفتند.

یافته‌ها: در مدل درختی مقادیر ضریب تبیین برای دوره ٤ ساله با حذف ۵٪، ۱۰٪ و ۲۰٪ داده‌ها به‌ترتیب برابر ۰/۸۵، ۰/۷۵ و ۰/۸۵ و برای دوره ۱۲ ساله برابر ۰/۹۰، ۰/۸۳ و ۰/۸٤ به‌دست آمد. در مورد مدل شبکه عصبی مقادیر ضریب تبیین برای دوره ٤ ساله با حذف ۵٪، ۱۰٪ و ۲۰٪ داده‌ها به‌ترتیب برابر ۰/۸۵، ۰/۷۵ و ۰/۸۵ و برای دوره ۱۲ ساله برابر ۰/۹۰، ۰/۸۲ و ۰/۸۵ به‌دست آمد. بیش‌تر بودن مقادیر ضریب تبیین برای دوره آماری ۱۲ ساله نشان داد، مدل‌ها هنگام تخمین داده‌های گمشده برای دوره‌های آماری طولانی‌تر دارای عملکرد بهتری هستند. با افزایش داده‌های گمشده از

* مسئول مکاتبه: akhob@ut.ac.ir

۵٪ به ۲۰٪ نیز از دقت این مدل‌ها کاسته شد. همچنین مقایسه این دو مدل نشان داد که هردو دارای دقت مشابهی در برآورد داده‌های گمشده می‌باشند.

نتیجه‌گیری: با توجه به نتایج به‌دست آمده از این پژوهش، هر دو مدل شبکه عصبی و درختی در صورت موجود بودن بیش از ۱۰ سال داده‌های آماری، نتایج نسبتاً مطلوبی خواهند داشت. همچنین هنگامی که تعداد داده‌های گمشده کم‌تر باشند و یا در بازه‌های کوتاه‌تری گم شده باشند، مقادیر تخمین زده شده به مقادیر واقعی نزدیک‌تر خواهد بود. به‌منظور بهبود و تکمیل نتایج این پژوهش پیشنهاد می‌شود تخمین داده‌های گمشده برای دوره‌های آماری مختلف نظیر ۸ و یا ۱۵ سال نیز تکرار و بهترین دوره آماری برای بهترین عملکرد مدل‌ها مشخص گردد.

واژه‌های کلیدی: داده‌های گمشده هواشناسی، بازسازی داده‌ها، مدل درختی، مدل شبکه عصبی مصنوعی، خوزستان

مقدمه

روش تشت تبخیر به‌عنوان یک روش ساده و مناسب برای برآورد تبخیر و تعرق مرجع استفاده می‌شود. اما در بعضی موارد، داده‌های تبخیر به‌دست آمده از تشت تبخیر به‌علت دقت پایین در برداشت داده‌ها و یا نقص درتجهیزات هواشناسی، ناقص می‌باشند و چون برای انجام برنامه‌ریزی‌های صحیح در منابع آب، پیوستگی داده‌ها دارای اهمیت می‌باشد، بنابراین لازم است این نواقص آماری به طریقی برطرف گردد. تخمین داده‌های گمشده که با عنوان‌های پر کردن داده‌های گمشده (2)، بازسازی داده‌های گمشده (1) و تکمیل سری داده‌های هواشناسی (4) نیز شناخته می‌شود، موضوع مهم و شایان توجهی در مطالعات هواشناسی و هیدرولوژیکی می‌باشد (2, 1, 4).

در میان روش‌های متنوع یافتن داده‌های گمشده تاکنون مدل درختی چندین بار برای برآورد داده‌های گمشده هواشناسی مورد استفاده قرار گرفته است. مدل درختی را برای تخمین داده‌های گمشده سرعت باد با استفاده از داده‌های ایستگاه‌های مجاور به‌کار بردند و نتایج مطلوبی به‌دست آوردند (3). وو کیم و پاچپسکی (2010) دو مدل شبکه عصبی مصنوعی و درختی را به‌صورت جداگانه و به‌صورت ترکیبی برای یافتن داده‌های گمشده بارش روزانه در حوضه چساپیک با استفاده از داده‌های بارش روزانه ۷ سال آماری ۳۹ ایستگاه هواشناسی واقع در حوضه به‌کار بردند (6). نتایج این مطالعات نشان داد که مدل شبکه عصبی- درختی دارای دقت بیش‌تری در بازسازی داده‌های گمشده بارش نسبت به مدل درختی و مدل شبکه عصبی به‌صورت جداگانه می‌باشد. سنگون و همکاران (2010) مدل شبکه عصبی را برای برآورد داده‌های گمشده دمای متوسط ماهانه خاک با استفاده از داده‌های موجود دمای ماهانه خاک در ایستگاه‌های مجاور به‌کار بردند (5). آنها بدون استفاده از هیچ داده یا متغیر دیگری و تنها با داشتن متوسط دمای ماهانه در هشت ایستگاه همسایه، داده‌های دمای خاک را با استفاده از این مدل تخمین زدند. ضریب تبیین بین داده‌های اندازه‌گیری شده دمای خاک و داده‌های تخمین زده شده با استفاده از مدل شبکه عصبی ۰/۹۹ به‌دست آمد. نتایج این پژوهش‌ها نشان داد که می‌توان از مدل شبکه عصبی برای تخمین داده‌های گمشده دمای ماهانه خاک با دقت بسیار بالا استفاده کرد.

هدف از انجام این پژوهش، مقایسه دو مدل درختی و شبکه عصبی برای یافتن داده‌های گمشده تبخیر از تشت در استان خوزستان است. به این منظور داده‌های گمشده با استفاده از داده‌های هواشناسی

روزانه سرعت باد، حداکثر و حداقل دمای هوا، رطوبت نسبی هوا و تابش برون زمینی به‌عنوان داده‌های ورودی مدل‌ها، تخمین زده شدند.

مواد و روش‌ها

منابع داده‌ها: در این پژوهش داده‌های مورد نیاز از ٤ ایستگاه هواشناسی سینوپتیک شهرستان‌های آغاجاری، بندر ماهشهر، ایذه و بستان واقع دراستان خوزستان جمع‌آوری شد. اقلیم این ایستگاه‌ها براساس طبقه‌بندی دومارتن خشک می‌باشد. داده‌ها مربوط به سال‌های ۱۹۹۷ تا ۲۰۰۸ میلادی و شامل مقادیر روزانه سرعت باد، حداکثر و حداقل دمای هوا، رطوبت نسبی هوا، ساعات آفتابی و تابش برون‌زمینی می‌باشد. مشخصات ایستگاه‌های فوق در جدول ۱ ملاحظه می‌شود. پراکنش ایستگاه‌های انتخابی نیز در شکل ۱ نمایش داده شده است.

جدول ۱- مشخصات ایستگاه‌های هواشناسی منطقه مورد مطالعه.

Table 1. Profile of weather stations used in the study.

ایستگاه (Station)	طول جغرافیایی (E) Lon. (E)	عرض جغرافیایی (N) Lat. (N)	ارتفاع (Elavation)	دوره آماری (period)
آغاجاری (Aghajari)	49°، 39′	30°، 45′	27	2008-1997
بندرماهشهر (Bandar Mahshahr)	49°،9′	30°، 33′	6.2	2008-1997
ایذه (Izeh)	49°، 51′	31°، 51′	767	2008-1997
بستان (Bostan)	48°، 00′	31°، 42′	7.8	2008-1997

شکل ۱- موقعیت ایستگاه‌های هواشناسی در استان خوزستان.

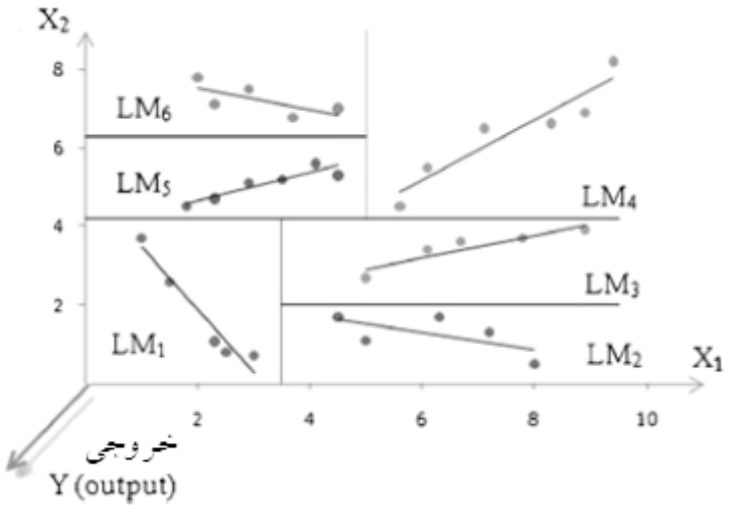
Figure 1. Location of weather stations used in the study area.

روش‌های برآورد داده‌های گمشده تبخیر از تشت

مدل درختی M5: اولین بار مدل درختی M5 توسط کوئینلن براساس روش طبقه‌بندی درختی برای ایجاد رابطه بین متغیرهای مستقل و وابسته ارائه شد. این مدل بر خلاف مدل درخت تصمیم که برای داده‌های کیفی استفاده می‌شود، برای هر دو نوع داده‌های کمی و کیفی قابل استفاده است. مدل M5 مشابه توابع خطی جدا شده است که ترکیبی از مدل‌های رگرسیون خطی و رگرسیون درختی است که کاربردهای زیادی در علوم مختلف دارد. مدل رگرسیون برای کل فضای داده‌ها یک معادله رگرسیون ارائه می‌دهد، ولی در مدل رگرسیون درختی، محدوده داده‌ها به زیرناحیه‌هایی که برگ نامیده می‌شوند تقسیم شده و به هر یک از این برگ‌ها یک برچسب عددی نسبت داده می‌شود. جایگزینی معادله رگرسیون خطی به‌جای برچسب در گره‌ها، شیوه‌ای است که در مدل M5 اجرا می‌شود که می‌تواند متغیرهای عددی پیوسته را برآورد کند. ساختار درخت تصمیم‌گیری شبیه یک درخت بوده که از ریشه، شاخه‌ها، گره‌ها و برگ‌ها تشکیل یافته است. درخت تصمیم از بالا به پایین ترسیم می‌شود. ریشه به‌عنوان اولین گره در بالا قرار گرفته و زنجیره‌ای از شاخه‌ها و گره‌ها به برگ‌ها ختم می‌شود. هر گره مربوط به یک متغیر پیش‌بینی‌کننده بوده و به‌وسیله شاخه‌ها عمل انشعاب در گره انجام می‌گیرد. شاخه‌ها شامل بازه‌ای عددی است که از گره والد منشعب شده و به یک گره فرزند می‌رسد. در مدل M5 از هر گره والد دو شاخه منشعب می‌شود. ساخت مدل درخت تصمیم‌گیری در دو مرحله انجام می‌شود. در مرحله اول درخت تصمیم‌گیری با انشعاب‌سازی داده‌ها تشکیل می‌شود. معیار انشعاب در مدل M5 بیشینه‌سازی کاهش انحراف معیار داده‌ها در گره فرزند است. زمانی که امکان کاهش انحراف معیار داده‌های گره فرزند میسر نبود، گره والد آن منشعب نشده و به گره پایانی و یا برگ رسیده است. کاهش انحراف معیار از رابطه زیر برآورد می‌شود:

$$SDR = sd(T) - \sum \frac{|Ti|}{|T|} sd(Ti) \qquad (۱)$$

که در آن، SDR: کاهش انحراف معیار در گره فرزند، T: مجموعه داده‌های ورودی به گره والد، T_i: زیرمجموعه از داده‌های ورودی به گره والد و sd: انحراف معیار است. به‌علت فرآیند انشعاب، انحراف معیار داده‌ها در گره فرزند کم‌تر از گره واحد بوده و بنابراین از همگنی بیش‌تری برخوردارند. در مدل M5 بعد از آزمودن تمام انشعابات ممکن از یک گره، انشعابی که بیش‌ترین کاهش انحراف معیار را تولید کند، انتخاب می‌شود. این گونه انشعاب‌سازی در اغلب موارد منجر به ایجاد درخت بزرگی شده و امکان دارد برازش بیش از حد روی داده‌های آموزشی رخ دهد. برازش بیش از حد باعث کاهش یافتن کلیت بخشی مدل شده به‌طوری‌که مدل فقط برای داده‌هایی که برای ساخت آن استفاده شدند اعتبار دارد و برای داده‌های جدید از دقت لازم برخوردار نیست. بنابراین مرحله دوم طراحی مدل درختی، شامل کوچک کردن درخت بیش از حد بزرگ شده از طریق هرس‌سازی شاخه‌ها و جایگزین شدن با توابع رگرسیون خطی است. شکل ۲ روند تقسیم فضای داده‌های ورودی X_1 و X_2 (متغیرهای مستقل) را به‌وسیله مدل درختی نشان می‌دهد. در مدل‌سازی رگرسیون مربوط به هر زیرحوزه، خطی و به‌صورت $y=a_0+a_1x_1+a_2x_2$ می‌باشد.



شکل ۲- شمای تقسیم فضای ورودی در مدل درختی.

Figure 2. Diagram of splitting the input space in model tree.

در این بررسی از نرم‌افزار Weka که در دانشگاه Waikato کشور نیوزیلند توسعه‌یافته، استفاده شده است. ساخت مدل M5 با استفاده از داده‌های تبخیر از تشت موجود سال‌های ۱۹۹۷ تا ۲۰۰۸ انجام شد و داده‌های گمشده تبخیر از تشت این سال‌ها برای تست و ارزیابی استفاده شدند.

مدل شبکه عصبی: در این پژوهش از شبکه‌های چندلایه پیشرونده با الگوریتم آموزشی پس‌انتشار خطا استفاده شد، که جزء روش‌های آموزش با ناظر است. ساختار این شبکه شامل یک لایه ورودی، یک لایه میانی و یک لایه خروجی است. در هر لایه یک یا چند عنصر پردازشگر (نرون) وجود دارد که با تمامی نرون‌های لایه بعدی با اتصالات وزن‌دار بهم مربوط می‌شوند. بردار داده‌های ورودی مدل به نرون‌های لایه اول نگاشت می‌شوند و در این لایه هیچ‌گونه پردازشی انجام نمی‌گیرد و نرون‌های لایه خروجی به بردار خروجی مدل نگاشت می‌گردند. تعداد نرون‌های لایه‌های ورودی و خروجی بستگی به تعداد متغیرهای ورودی و خروجی مدل دارد ولی انتخاب تعداد نرون‌های لایه میانی به‌صورت سعی و خطا تعیین می‌شود. در این بررسی، یک ساختار از پارامترهای هواشناسی مؤثر بر تبخیر از تشت شامل داده‌های روزانه کمینه و بیشینه دمای هوا، سرعت باد، تابش فرازمینی و رطوبت نسبی هوا، به‌عنوان ورودی مدل شبکه عصبی مورد بررسی قرار گرفت و داده‌های گمشده تبخیر از تشت روزانه، خروجی این مدل را تشکیل داد.

در شبکه‌های عصبی، نرون‌های هر لایه به همه نرون‌های لایه قبل از طریق یک اتصال جهت‌دار مرتبط می‌شوند. به هر یک از این اتصالات وزنی داده می‌شود که مقدار آن تعیین‌کننده تأثیر هر نرون بر روی نرون لایه خروجی است. مجموع وزنی مقادیر ورودی به هر نرون محاسبه می‌شود و در یک تابع ریاضی قرار می‌گیرد و خروجی نرون از طرق این تابع محاسبه می‌شود. این تابع ریاضی را اصطلاحاً تابع محرک، تابع آستانه و یا تابع انتقال گویند. در این پژوهش، تابع سیگموئید (S)، مورد بررسی قرار گرفت.

وزن‌های ارتباط‌دهنده نرون‌های شبکه، با آموزش تعیین می‌شوند و در شبکه‌های چندلایه از الگوریتم

آموزش پس‌انتشار خطا استفاده می‌شود. در این الگوریتم ابتدا مقادیر تصادفی برای وزن‌ها انتخاب می‌شود و خروجی شبکه به‌دست می‌آید. خطای بین خروجی شبکه با مقدار مطلوب آن به‌سمت عقب انتشار می‌یابد و بر این اساس، وزن‌ها تعدیل می‌شوند. این فرایند تکرار می‌شود تا خروجی شبکه به یک مقدار قابل‌قبولی برسد، این فرایند را آموزش شبکه عصبی گویند. در صورت تکرار زیاد فرایند آموزش، اوزان شبکه به‌صورتی تعدیل می‌شوند که فقط برای داده‌هایی که برای آموزش استفاده شدند، عملکرد خوبی دارند. ولی برای داده‌هایی که در آموزش شبکه از آن‌ها استفاده نشده، عملکرد ضعیفی دارند. این اتفاق را آموزش بیش از حد می‌گویند. برای جلوگیری از آموزش بیش از حد و تصمیم برای توقف مرحله آموزش از یک سری داده به‌عنوان داده‌های ارزیابی استفاده می‌شود. پس از هر بار تکرار فرایند یادگیری، شبکه با اوزان جدید برای داده‌های ارزیابی اجرا می‌شود. به‌طور معمول در مراحل اولیه آموزش، خطای برآورد خروجی مدل برای داده‌های ارزیابی کاهش می‌یابد. ولی زمانی که آموزش بیش از حد داده‌ها اتفاق می‌افتد، این خطا افزایش می‌یابد. با شروع افزایش این خطا، آموزش داده‌ها متوقف می‌شود و بنابراین وزن‌های شبکه در شرایط حداقل خطا برای داده‌های ارزیابی تعیین می‌شوند. به‌عبارتی آموزش شبکه با استفاده از داده‌های آموزش و ارزیابی صورت می‌گیرد. بعد از آموزش، شبکه با داده‌هایی که در آموزش و ارزیابی از آن‌ها استفاده نشده، آزمایش شده و عملکرد آن با استفاده از شاخص‌های آماری بررسی می‌گردد.

الگوریتم پس‌انتشار خطا دارای توابعی مختلف بوده که تفاوت آن‌ها در نحوه تنظیم وزن‌های ارتباط‌دهنده لایه‌های شبکه عصبی است. تابع لونبرگ مارکوارت (LM) از رایج‌ترین توابع می‌باشند که در این پژوهش برای آموزش شبکه عصبی مورد استفاده قرار گرفت.

مشابه مدل درختی، داده‌های تبخیر از تشت سال‌های ۱۹۹۷ تا ۲۰۰۸ برای آموزش و ارزیابی مدل‌های شبکه عصبی و داده‌های گمشده تبخیر از تشت این سال‌ها به آزمون مدل اختصاص یافت. از میان داده‌های آموزش و ارزیابی به‌طور تصادفی، ۷۰ درصد آن به آموزش و ۳۰ درصد بقیه به ارزیابی شبکه اختصاص داده شدند.

آماده‌سازی داده‌ها: برای انجام این پژوهش داده‌های تمامی ٤ ایستگاه به دو دوره ٤ و ۱۲ ساله تقسیم شدند. داده‌های چهار ساله شامل داده‌های هواشناسی و تبخیر از تشت روزانه سال‌های ۲۰۰۵ تا ۲۰۰۸ و داده‌های دوازده ساله شامل داده‌های هواشناسی و تبخیر از تشت روزانه سال‌های ۱۹۹۷ تا ۲۰۰۸ بودند. سپس در هر دوره، داده‌های تبخیر از تشت مطابق جدول ۲ به‌طور تصادفی حذف شدند. حذف داده‌ها در این پژوهش به‌صورت تصادفی با استفاده از RANDBETWEEN در نرم‌افزار Excel انجام شد. به این صورت که برای حذف ۵٪ داده‌ها در گروه ٤ ساله، پنج بازه نیم ماهه ماهه حذف شدند. انتخاب این پنج بازه هر بار با استفاده از RANDBETWEEN مشخص شد که بازه حذف شده باید در کدام سال (۲۰۰۵ تا ۲۰۰۸) باشد. سپس مجدداً تعیین شد که در سال انتخاب شده برای حذف، حذف نهایی داده‌ها از چه ماهی از آن سال باید صورت گیرد.

جدول ۲- چگونگی حذف داده‌ها.

Table 2. How to remove data.

بازه زمانی حذف داده‌ها (ماه) (The period of data removal)	0.5	1	1.5
درصد داده‌های حذف شده (The percentage of deleted data)	5	10	20

در این پژوهش داده‌های حذف شده به این صورت نشانه‌گذاری شد: 0.5 - ٪5 - 4 year، به این معناست که دوره آماری ٤ ساله است و ۵ درصد کل داده‌ها با بازه زمانی ۰/۵ ماهه حذف شدند و یا 1.5 -٪20 - 12 year بیان‌کننده این است که دوره آماری داده‌ها ۱۲ ساله است و ۲۰ درصد کل داده‌ها با بازه زمانی ۱/۵ ماهه حذف شدند. این روند برای تمامی دوره‌های ٤ و ۱۲ سال و در تمامی ایستگاه‌ها تکرار شد. این داده‌های حذف شده، داده‌های گمشده نامید شدند. داده‌های حذف شده را داده‌های گمشده نامیدیم. داده‌های گمشده، خروجی مورد انتظار مدل‌ها بودند و از آن‌ها با عنوان داده‌های آزمون استفاده شدند. داده‌های تبخیر از تشت موجود در هر ایستگاه بعد از حذف داده‌های گمشده با عنوان داده‌های آموزشی برای ورودی مدل‌ها به‌کار رفتند. دلیل حذف درصدهای مختلف از داده‌ها (۵ تا ۲۰ درصد) و در بازه‌های ۰/۵ تا ۱/۵ ماهه و نیز طبقه‌بندی داده‌ها به دو دوره ٤ و ۱۲ ساله، بررسی دقت مدل‌ها با تغییر تعداد داده‌های آموزشی و آزمون بود.

ارزیابی روش برآورد داده‌های گمشده تبخیر از تشت: صحت مقادیر برآوردی با محاسبه ضریب تبیین بین داده‌ها[1]، میانگین انحراف خطا[2] و ریشه میانگین مربعات خطا[3] براساس روابط زیر ارزیابی شد. ارزیابی عملکرد مدل‌ها براساس کم‌ترین مقدار میانگین انحراف خطا و ریشه میانگین مربعات خطا و نیز بیش‌ترین مقدار ضریب تبیین انجام شد.

$$R^2 = \frac{\sum_{i=1}^{n}\left(s_i - \overline{c_i}\right)^2}{\sum_{i=1}^{n}\left(c_i - \overline{c_i}\right)^2} \qquad (۲)$$

$$MBE = \frac{\sum_{i=1}^{n}\left(s_i - c_i\right)}{n} \qquad (۳)$$

$$RMSE = \left(\frac{\sum_{i=1}^{n}\left(s_i - c_i\right)^2}{n}\right)^{0.5} \qquad (٤)$$

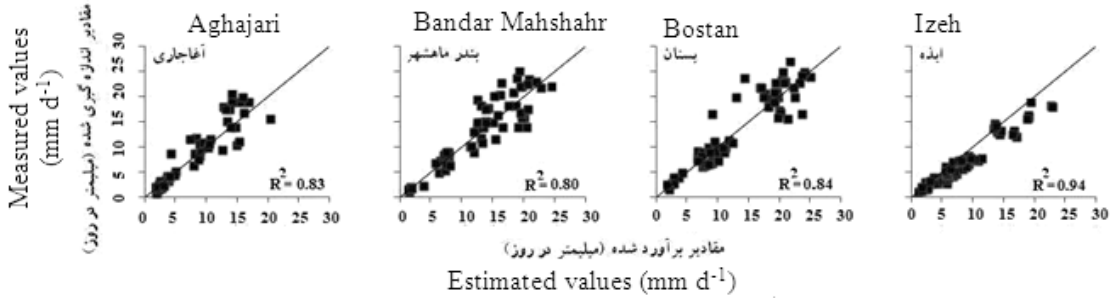
که در آن‌ها، $\overline{c_i}$: میانگین مقدار مشاهداتی متغیر، c_i: مقدار مشاهداتی (واقعی) متغیر، s_i: مقدار محاسبه شده متغیر توسط مدل و n: تعداد داده‌های مشاهداتی می‌باشند.

نتایج و بحث

مدل درختی: به‌منظور تولید درخت بهینه از پارامترهای روزانه دمای کمینه و بیشینه هوا، رطوبت نسبی هوا، سرعت باد و تابش فرازمینی استفاده شد. پراکنش نتایج مدل درختی برای تمامی ایستگاه‌ها در شکل‌های ۳ تا ۸ نمایش داده شده است. با توجه به این اشکال با افزایش تعداد داده‌های گمشده از ۵٪ به ۲۰٪ از دقت مدل کاسته می‌شود. زیرا از تعداد داده‌های آموزشی ورودی به مدل کاسته می‌شود در حالی‌که تعداد داده‌های آزمون افزایش می‌یابند. با

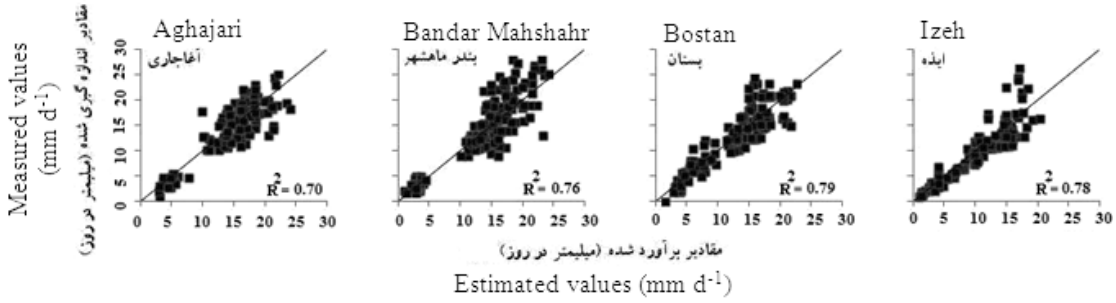
1- R^2
2- MBE
3- RMSE

افزایش دوره آماری از ٤ سال به ۱۲ سال نیز به دلایل ذکر شده، دقت مدل افزایش یافته و نتایج برآورد شده به مقادیر واقعی نزدیک‌تر می‌باشند. همان‌طورکه در این اشکال ملاحظه می‌شود در اکثر گروه‌ها، مقادیر ضریب تبیین در ایستگاه ایذه دارای بیش‌ترین مقدار نسبت به سایر ایستگاه‌ها می‌باشند که نشان‌دهنده دقت بالاتر داده‌های برآورد شده در این ایستگاه می‌باشد. این مقادیر برای ایستگاه ایذه در شش گروه مورد بررسی برابر با ۰/۹٤، ۰/۷۸، ۰/۸۹، ۰/۹۲، ۰/۸۱ و ۰/۸۸ می‌باشند. نکته دیگری که در این اشکال قابل ملاحظه است، برآورد کم‌تر مدل درختی نسبت به مقدار واقعی در اکثر گروه‌ها می‌باشد.



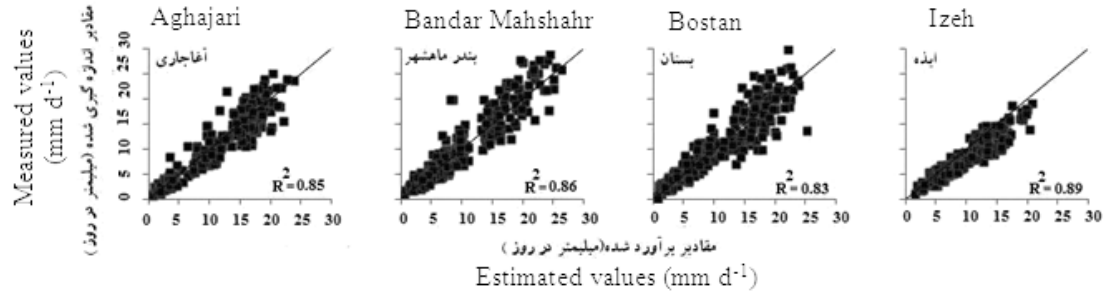
شکل ۳- پراکنش مقادیر برآورد شده از مدل درختی و اندازه‌گیری شده برای گروه 4year-%5-0.5.

Figure 3. Dispersion of estimated data by tree model and measured data for 4year-%5-0.5 group.



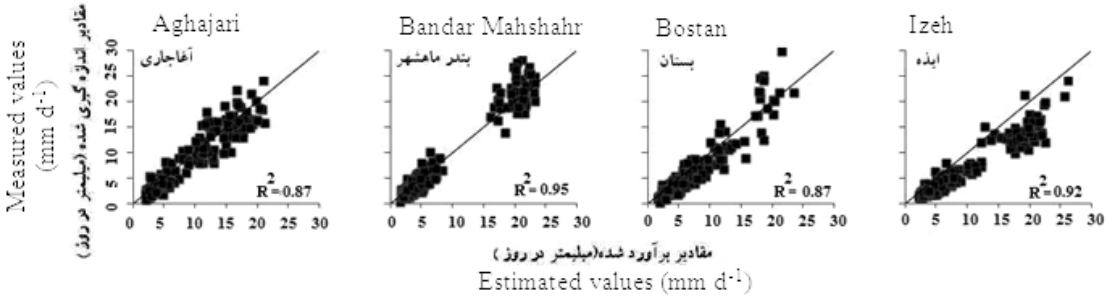
شکل ٤- پراکنش مقادیر برآورد شده از مدل درختی و اندازه‌گیری شده برای گروه 4year-%10-1.

Figure 4. Dispersion of estimated data by tree model and measured data for 4year-%10-1 group.



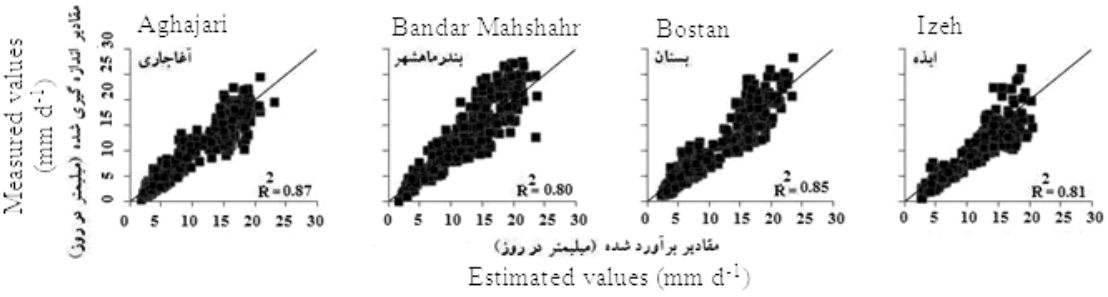
شکل ٥- پراکنش مقادیر برآورد شده از مدل درختی و اندازه گیری شده برای گروه 4year-%20-1.5.

Figure 5. Dispersion of estimated data by tree model and measured data for 4year-%20-1.5 group.



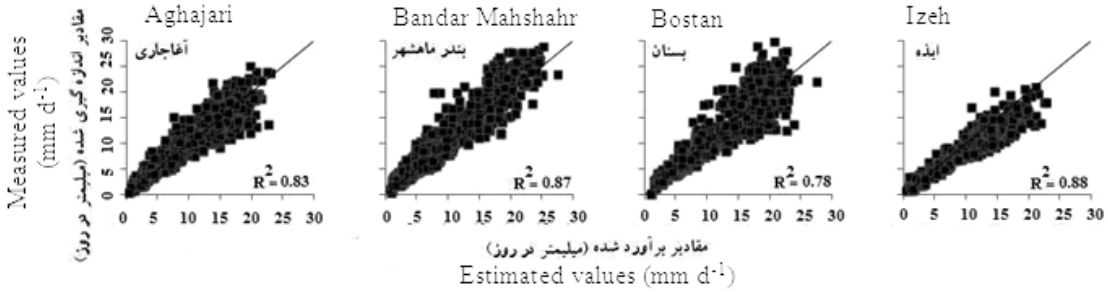
شکل ٦- پراکنش مقادیر برآورد شده از مدل درختی و اندازه‌گیری شده برای گروه 12year-%5-0.5.

Figure 6. Dispersion of estimated data by tree model and measured data for 12year-%5-0.5 group.



شکل ۷- پراکنش مقادیر برآورد شده از مدل درختی و اندازه‌گیری شده برای گروه 12year-%10-1.

Figure 7. Dispersion of estimated data by tree model and measured data for 12year-%10-1 group.



شکل ۸- پراکنش مقادیر برآورد شده از مدل درختی و اندازه‌گیری شده برای گروه 12year-%20-1.5.

Figure 8. Dispersion of estimated data by tree model and measured data for 12year-%20-1.5 group.

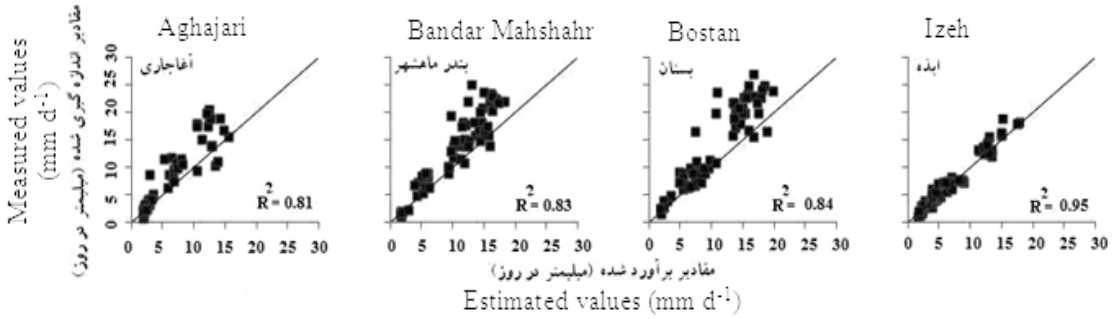
در جدول ۳ مقادیر میانگین انحراف خطا (MBE) و ریشه میانگین مربعات خطا (RMSE) برای تمامی ایستگاه‌ها ودر همه گروه‌ها نشان داده شده است. همان‌طورکه ملاحظه می‌شود، برای دوره آماری ٤ ساله، کم‌ترین و بیش‌ترین خطا MBE به‌ترتیب برابر با ۰/۸- تا ۱/۵۳ میلی‌متر در روز برآورد شده که عدد منفی نشان‌دهنده کم برآوردی مدل می‌باشد. در همین دوره کم‌ترین و بیش‌ترین خطا RMSE به‌ترتیب برابر با ۰/۸- تا ۱/۵۳ میلی‌متر در روز برآورد شده ۲/۰۲ و ۳/۳۵ میلی‌متر در روز برآورد شده است. مقادیر بالای خطای فوق نشان‌دهنده این است که برای دوره آماری ٤ ساله، نمی‌توان با استفاده از مدل درختی داده‌های گمشده را با دقت مطلوب تعیین کرد.

جدول ۳- نتایج مدل درختی برای تمامی گروه‌ها.

Table 3. The results of tree model for all groups.

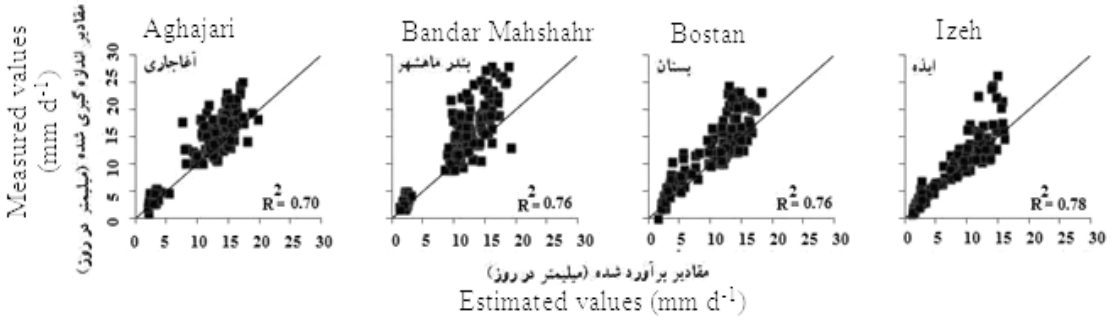
نام ایستگاه Name of Station	نام گروه Group Name	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)	نام ایستگاه Name of Station	نام گروه Group Name	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)
آغاجاری Aghajari	4-%5-0.5	-0.8	2.58	آغاجاری Aghajari	12-%5-0.5	0.13	2.13
	4-%10-1	-0.23	2.72		12-%10-1	0.007	2.16
	4-20%-1.5	-0.009	2.3		12-%20-1.5	-0.36	2.32
بندر ماهشهر Bandar Mahshahr	4-%5-0.5	-0.52	2.86	بندر ماهشهر Bandar Mahshahr	12-%5-0.5	-0.25	1.87
	4-%10-1	-0.65	3.35		12-%10-1	-0.89	3.01
	4-20%-1.5	-0.71	2.59		12-%20-1.5	-1.24	2.81
بستان Bostan	4-%5-0.5	-0.62	2.91	بستان Bostan	12-%5-0.5	0.47	2.08
	4-%10-1	-0.08	2.62		12-%10-1	-0.12	2.42
	4-20%-1.5	-0.31	2.74		12-%20-1.5	-0.42	2.97
ایذه Izeh	4-%5-0.5	1.42	2.02	ایذه Izeh	12-%5-0.5	2.94	3.75
	4-%10-1	0.46	2.49		12-%10-1	0.99	2.28
	4-20%-1.5	1.53	2.14		12-%20-1.5	0.97	1.98

مدل شبکه عصبی: در مدل شبکه عصبی از میان داده‌ها، ۷۰ درصد آن به آموزش و ۳۰ درصد بقیه به ارزیابی شبکه اختصاص داده شدند. برآورد تبخیر از تشت با استفاده از الگوریتم آموزشی لونبرگ مارکوارت و تابع انتقال سیگموئیداکسون انجام شد واز پارامترهای روزانه دمای کمینه و بیشینه هوا، رطوبت نسبی هوا، سرعت باد و تابش فرازمینی به‌عنوان ورودی مدل استفاده شد. پراکنش نتایج مدل شبکه عصبی برای تمامی ایستگاه‌ها در اشکال ۹ تا ١٤ نمایش داده شده است. با توجه به این اشکال با افزایش تعداد داده‌های گمشده از ۵٪ به ۲۰٪ از دقت مدل کاسته می‌شود. زیرا از تعداد داده‌های آموزشی ورودی به مدل کاسته می‌شود در حالی‌که تعداد داده‌های آزمون افزایش می‌یابند. با افزایش دوره آماری از ٤ سال به ۱۲ سال نیز به دلایل ذکر شده، دقت مدل افزایش یافته و نتایج برآورد شده به مقادیر واقعی نزدیک‌تر می‌باشند. همان‌طورکه در این اشکال ملاحظه می‌شود در اکثر گروه‌ها، مقادیر ضریب تبیین در ایستگاه ایذه دارای بیش‌ترین مقدار نسبت به سایر ایستگاه‌ها می‌باشند که نشان‌دهنده دقت بالاتر داده‌های برآورد شده در این ایستگاه می‌باشد. این مقادیر برای ایستگاه ایذه در شش گروه مورد بررسی برابر با ۰/۹۵، ۰/۷۸، ۰/۸۹، ۰/۹۳، ۰/۸۱ و ۰/۸۹ می‌باشند. نکته دیگری که در این اشکال قابل ملاحظه است، برآورد کم‌تر مدل شبکه عصبی مصنوعی نسبت به مقدار واقعی در اکثر گروه‌ها می‌باشد.



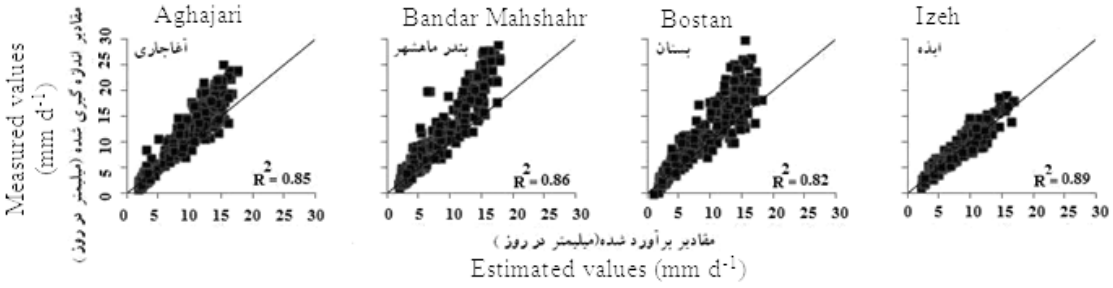
شکل ۹- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 4year-%5-0.5.

Figure 9. Dispersion of estimated data by neural network model and measured data for 4year-%5-0.5 group.



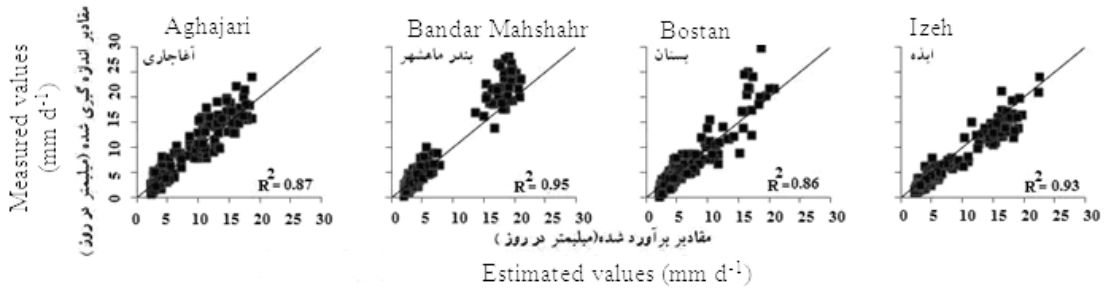
شکل ۱۰- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 4year-%10-1.

Figure 10. Dispersion of estimated data by neural network model and measured data for 4year-%10-1 group.



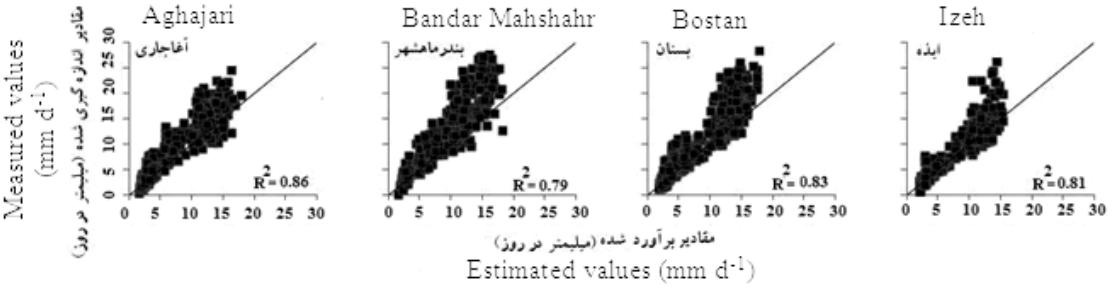
شکل ۱۱- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 4year-%20-1.5.

Figure 11. Dispersion of estimated data by neural network model and measured data for 4year-%20-1.5 group.



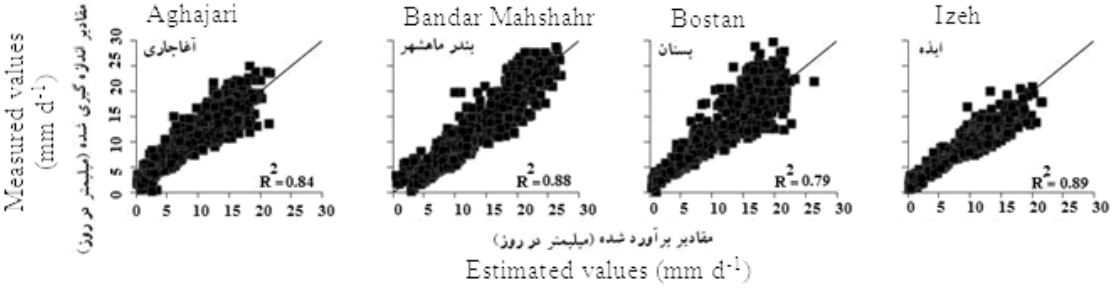
شکل ۱۲- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 12year-%5-0.5.

Figure 12. Dispersion of estimated data by neural network model and measured data for 12year-%5-0.5 group.



شکل ۱۳- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 12year-%10-1.

Figure 13. Dispersion of estimated data by neural network model and measured data for 12year-%10-1.



شکل ١٤- پراکنش مقادیر برآورد شده از مدل شبکه عصبی و اندازه‌گیری شده برای گروه 12year-%20-1.5.

Figure 14. Dispersion of estimated data by neural network model and measured data for 12year-%20-1.5.

در جدول ٤ مقادیر میانگین انحراف خطا (MBE) و ریشه میانگین مربعات خطا (RMSE) برای تمامی ایستگاه‌ها ودر همه گروه‌ها نشان داده شده است. با توجه به این جدول مقادیر RMSE برای ایستگاه ایذه، کم‌تر از سایر ایستگاه‌ها می‌باشد که این نشان‌دهنده انحراف کم‌تر مقادیر برآورد شده از مقادیر واقعی در این ایستگاه می‌باشد. این مقادیر در شش گروه مورد بررسی برای ایستگاه ایذه برابر با ۱/۲۵، ۲/۹، ۱/٤۷، ۱/۷۷، ۲/۵۷ و ۱/۲۰ میلی‌متر در روز می‌باشند. همچنین منفی بودن مقدار MBE در اکثر گروه‌ها نشان‌دهنده برآورد کم‌تر مدل شبکه عصبی نسبت به مقادیر واقعی تبخیر از تشت می‌باشد.

مقایسه مدل‌ها: شاخص‌های آماری ناشی از مقایسه نتایج مدل‌های شبکه عصبی و رگرسیون درختی برای دوره ٤ ساله و ۱۲ ساله آماری در جدول‌های ۵ و ٦ نمایش داده شده است. همان‌طورکه مشاهده می‌شود مدل‌های رگرسیون درختی و شبکه عصبی برآوردی مناسب و نزدیک به هم را از میزان تبخیر از تشت نشان می‌دهند با این تفاوت که مدل درختی نشان‌دهنده خطای محاسباتی کم‌تری می‌باشد.

نتیجه‌گیری

در این پژوهش برای اولین بار در ایران از مدل درختی M5 و مدل شبکه عصبی برای یافتن داده‌های گمشده تبخیر در استان خوزستان استفاده شد. با توجه به نتایج به‌دست آمده، این مدل‌ها در صورت موجود بودن بیش از ۱۰ سال داده‌های آماری تبخیر از تشت، نتایج نسبتاً مطلوبی به‌دست می‌دهند. همچنین هنگامی که تعداد داده‌های گمشده کم‌تر باشند و یا در بازه‌های کوتاه‌تری گم شده باشند، مقادیر تخمین زده شده به مقادیر واقعی نزدیک‌تر می‌باشند. همان‌گونه این پژوهش نشان داد، مدل‌ها برای دوره ۱۲ ساله آماری نسبت به دوره ٤ ساله و برای گروه ۰/۵-۵٪ نسبت به گروه ۱/۵-۲۰٪ مقادیر داده‌های گمشده را با دقت بالاتری برآورد کردند. به‌منظور بهبود و تکمیل نتایج این پژوهش پیشنهاد می‌شود تخمین داده‌های گمشده برای دوره‌های آماری مختلف نظیر ۸ و یا ۱۵ سال نیز تکرار شود و بهترین دوره آماری برای بهترین عملکرد مدل‌ها مشخص شود. در این پژوهش از پارامترهای هواشناسی سرعت باد، دمای کمینه و بیشینه هوا، رطوبت نسبی هوا، تابش برون زمینی و ساعات آفتابی به‌عنوان ورودی‌های مدل‌های درختی و شبکه عصبی استفاده شد. توصیه می‌شود، دقت مدل‌ها با تغییر این پارامترها و انتخاب ترکیب‌های مختلف از آن‌ها به‌عنوان ورودی، سنجیده شود. همچنین حساسیت مدل نسبت به ورودی‌های مختلف مورد تحلیل قرار گیرد.

جدول ٤- نتایج مدل شبکه عصبی برای تمامی گروه‌ها.

Table 4. The results of neural network model for all groups.

نام ایستگاه Name of Station	نام گروه Group Name	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)	نام ایستگاه Name of Station	نام گروه Group Name	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)
آغاجاری Aghajari	4-%5-0.5	-2.46	3.7	آغاجاری Aghajari	12-%5-0.5	-0.89	2.3
	4-%10-1	-2.84	3.93		12-%10-1	-2.19	3.26
	4-20%-1.5	-2.33	3.47		12-%20-1.5	-1.41	1.47
بندر ماهشهر Bandar Mahshahr	4-%5-0.5	-3.31	4.39	بندر ماهشهر Bandar Mahshahr	12-%5-0.5	-1.17	2.56
	4-%10-1	-3.38	4.9		12-%10-1	-3.49	4.78
	4-20%-1.5	-3.06	4.39		12-%20-1.5	1.95	2.04
بستان Bostan	4-%5-0.5	-3.75	5	بستان Bostan	12-%5-0.5	-0.28	2.21
	4-%10-1	-2.37	3.69		12-%10-1	-2.67	3.89
	4-20%-1.5	-2.6	4		12-%20-1.5	-1.09	1.09
ایذه Izeh	4-%5-0.5	-0.68	1.25	ایذه Izeh	12-%5-0.5	0.97	1.77
	4-%10-1	-1.54	2.9		12-%10-1	-1.52	2.57
	4-20%-1.5	-0.62	1.47		12-%20-1.5	-1.19	1.2

جدول ۵- مقایسه نتایج مدل‌ها برای دوره آماری ٤ ساله.

Table 5. The comparison between models for 4 years statistical period.

ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ضریب تبیین Coefficient of Determination	گروه Group Name	نام مدل Name of Model
2.59	-0.13	0.85	%5-0.5	Tree Model
2.79	-0.12	0.75	%10-1	
2.44	0.12	0.85	20%-1.5	
3.58	-2.55	0.85	%5-0.5	Neural Network
3.85	-2.53	0.75	%10-1	
3.33	-2.15	0.85	20%-1.5	

جدول ٦- مقایسه نتایج مدل‌ها برای دوره آماری ۱۲ ساله.

Table 6. The comparison between models for 12 years statistical period.

ریشه میانگین مربعات خطا (میلی‌متر در روز) RMSE(mm/d)	میانگین انحراف خطا (میلی‌متر در روز) MBE(mm/d)	ضریب تبیین Coefficient of Determination	گروه Group Name	نام مدل Name of Model
2.45	0.82	0.90	%5-0.5	Model Tree
2.46	-0.003	0.83	%10-1	
2.52	-0.26	0.84	20%-1.5	
2.21	-0.34	0.90	%5-0.5	Neural Network
3.62	-2.46	0.82	%10-1	
1.45	-1.41	0.85	20%-1.5	

منابع

1.Abebe, A.J., Solomatine, D.P., and Venneker, R.G.W. 2000. Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrol. Sci. J. 45: 3. 425-436.

2.Coulibaly, P., and Evora, N.D. 2007. Comparison of neural network methods for infilling missing daily weather records. J. Hydrol. 341: 1-2. 27-41.

3.Faucher, M., Burrows, W.R., and Pandolfo, L. 1999. Empirical-statistical reconstruction of surface marine winds along the western coast of Canada. Climate Research. 11: 3. 173-190.

4.Ramos-Calzado, P., Gomez-Camacho, J., Perez-Bernal, F., and Pita-Lopez, M.F. 2008. A novel approach to precipitation series completion in climatological data sets: application to Andalusia. Inter. J. Climatol. 28: 11. 1525-1534.

5.Sangun, L., Sahin, B., and Biligili, M. 2012. Estimating soil temperature using neighboring station data via artificial neural network models. J. Environ. Monitor. Assess. 185: 347-358.

6.Woo Kim, J., and Pachepsky, Y.A. 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J. Hydrol. 394: 305-314.

Gorgan University of Agricultural Sciences and Natural Resources

J. of Water and Soil Conservation, Vol. 22(4), 2015
http://jwsc.gau.ac.ir

Comparison between neural network and M5 model tree for reconstructing missing evaporation data of Khuzestan

M. Vahabi Mashhor[1] and *A. Rahimi Khoob[2]

[1]M.Sc. Graduate, Dept. of Irrigation and Drainage Engineering, Aburaihan College, University of Tehran,
[2]Professor, Dept. of Irrigation and Drainage Engineering, Aburaihan College, University of Tehran
Received: 10/30/2013; Accepted: 09/13/2014

Abstract

Background and Objectives: Pan Evaporation data used to estimate crop water requirements, but in some cases due to lack of accurate data on measurements or defective equipments failure and lost data as referred in missing data. Since data integration is important for irrigation planning, it is necessary to correct the statistical errors. Many methods were used for finding missing data, In the meantime, neural network and tree models have high degree of accuracy, however these models were not compared and evaluated. The aim of this study is to compare the model tree and neural network for reconstructing missing daily evaporation data for four stations in Khuzestan province.

Materials and Methods: In this study, the required data from four stations include Aghajari, Bandar Mahshahr, Izeh and Bostan located in Khuzestan province were collected. On the basis of the Koppen climate classification, the climate of these stations is arid. The data related to the years 1997 to 2008 that include daily values of pan evaporation, wind speed, maximum and minimum air temperature, relative humidity, sunshine and the extraterrestrial radiation. The data were divided to two four-year period (2005 to 2008) and 12 years (1997 to 2008) and at any period after intentional removal of 5%, 10% and 20% of the measured data, their values were estimated with the use of tree and neural network models. The results of the model were compared using Statistical indices.

Results: In tree model coefficient of determination for four years period were: 85%, 75% and 85% and for 12 years period were: 90%, 83% and 84% respectively. In neural network model coefficient of determination for 4years period were: 85%, 75% and 85% and for 12years period were: 90%, 82% and 85% respectively. A higher coefficient value for 12 years period showed that models are more accurate to estimate missing data for longer term statistical data. By increasing missing data from 5% to 20%, accuracy of models was diminished. This research also indicated that both models have similar accuracy in the estimation of missing data.

Conclusion: According to the results of this study, when more than 10 years of data are available, both neural network and tree models will have relatively good results. Also, when the number of missing data are less or missed in shorter periods, estimated values will be closer to the actual values. In order to improve and complete results, it is suggested that statistical estimates of missing data for different periods, such as 8 or 15 years are repeated and the best period to determine for best practice models.

Keywords: Missing data, Reconstructing data, Tree model, Neural network model, Khuzestan

* Corresponding Author; Email: akhob@ut.ac.ir