



UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO

**DIPARTIMENTO PER L'INNOVAZIONE NEI SISTEMI BIOLOGICI,
AGROALIMENTARI E FORESTALI (DIBAF)**

Corso di Dottorato di Ricerca in

ECOLOGIA FORESTALE – Ciclo XXVIII.

**Analisi del ruolo delle variabili climatiche nei flussi di CO₂ tra ecosistemi e
atmosfera tramite modelli empirici e serie storiche di misure eddy covariance**

s.s.d. AGR/05

Tesi di dottorato di:

Dott. DANIELA CAVALLI

Coordinatore del corso

Prof. Paolo De Angelis

Firma

Tutore

Prof. Dario Papale

Firma

16/06/2016

Sommario

1.	Introduzione	3
1.1.	Il ciclo del carbonio nel contesto dei cambiamenti climatici	3
1.2.	I controlli climatici della GPP	6
1.3.	Analisi delle relazioni tra flussi e variabili climatiche	7
1.4.	Reti Neurali Artificiali.....	8
1.5.	Metodologie per individuare l'importanza dei driver all'interno delle ANN	11
1.6.	Random Forests	14
1.7.	SSA (Singular System Analysis).....	16
1.8.	Scopo della ricerca	19
2.	Materiali e Metodi	21
2.1.	Drivers e Siti eddy-covariance selezionati	21
2.2.	Selezione della tipologia e parametrizzazione dei modelli	25
2.3.	Test e selezione dei metodi per la stima dell'importanza dei driver	27
2.4.	SSA – modalità di applicazione ed uso degli output.....	30
3.	Risultati.....	36
3.1.	Performances di ANN e RF addestrate con variabili non decomposte	36
3.2.	Selezione dei metodi per la stima dell'importanza dei drivers	38
3.3.	SSA – Segnali ottenuti dalla scomposizione	42
3.4.	Performance dei modelli ANN e RF addestrati usando variabili decomposte.....	48
3.5.	Stima importanza delle variabili driver e sintesi dei risultati.....	50
4.	Discussioni.....	64
4.1.	Performance di ANN e RF	64
4.2.	Selezione dei metodi più robusti per la selezione delle variabili guida	65
4.3.	SSA – segnali ottenuti dalla scomposizione delle variabili	66
4.3.	Performance ANN e RF addestrate con le componenti estratte dalle variabili.....	67
4.4.	Importanza dei driver	68
5.	Conclusioni	71
	Bibliografia.....	73

1. Introduzione

1.1. Il ciclo del carbonio nel contesto dei cambiamenti climatici

Nelle ultime decadi si è assistito ad un crescente interesse per il bilancio del carbonio negli ecosistemi terrestri (Janssens & Pilegard, 2003; Smith, 2004). Tale interesse è scaturito principalmente dal rapido riscaldamento globale e da altri problemi legati ai cambiamenti climatici, considerati diretta conseguenza dell'aumento della concentrazione dei gas serra atmosferici, specialmente del diossido di carbonio (IPCC, 2007).

L'attività umana ha giocato un ruolo fondamentale nell'aumento dell'anidride carbonica (CO₂) nell'atmosfera soprattutto per il forte utilizzo dei combustibili fossili. L'uomo ha infatti man mano accresciuto l'uso di combustibili fossili (petrolio, gas naturale, carbone) a scopo combustivo, rilasciando il carbonio in atmosfera in forma ossidata (la CO₂, appunto). La CO₂ è quindi il risultato della ossidazione dei combustibili fossili, i quali effettivamente costituiscono la principale fonte umana di emissione atmosferica di CO₂ dell'era moderna la quale costituisce oggi il principale gas responsabile dell'aumento dell'effetto serra.

Oggi tali cambiamenti impongono alla società lo sviluppo di nuove tecnologie in grado di ridurre le emissioni di gas serra e portano alla massimizzazione dell'efficienza con cui la biosfera è in grado di sequestrare il carbonio.

Queste problematiche sono assai sentite a livello mondiale ed europeo ed i modelli previsionali indicano che, sebbene il potenziale di sottrazione di carbonio da parte del suolo e della vegetazione non sia da solo in grado di compensare gli aumenti delle emissioni di origine antropica, la capacità di accumulo di carbonio nella biosfera nei prossimi 20-30 anni è una misura essenziale per poter mitigare gli aumenti di temperatura e poter stabilizzare le emissioni (Smith, 2004).

Lo studio del ciclo del carbonio e dei fattori che lo influenzano risulta quindi di fondamentale importanza. La GPP (Gross Primary Production) è la quantità di carbonio fissata dalla vegetazione attraverso la fotosintesi e rappresenta il flusso di carbonio dall'atmosfera alla superficie terrestre. Gli scambi netti di carbonio tra ecosistema e atmosfera derivano in pratica dal bilancio tra la GPP e la respirazione ad opera delle piante (autotrofa), dei microrganismi (soprattutto del suolo) e degli altri organismi eterotrofi (eterotrofa). Il bilancio tra entrate e perdite di carbonio, costituisce la produttività netta dell'ecosistema (NEP).

La tecnica Eddy Covariance

A livello metodologico, la produttività netta di un ecosistema può essere misurata attraverso misure micrometeorologiche in particolare utilizzando la tecnica dell'eddy covariance.

L'eddy covariance permette la misura continuativa dei flussi netti di CO₂, acqua, energia ed altri gas serra.

Utilizzando strumenti installati alla sommità di una torre che sovrasta di qualche metro la canopy. I vantaggi principali di questa tecnica di misura sono i seguenti:

- non è una misurazione invasiva, infatti non altera le condizioni ambientali;
- è una misurazione media dei flussi riferita ad una determinata area di base di riferimento (footprint) che si estende per alcune centinaia di metri attorno al punto di monitoraggio;
- è una misurazione che avviene in continuo e in molti casi su base multiannuale (*long-term measurement*).

La misura viene eseguita in un solo punto sopra la canopy, dove vengono misurate, ad alta frequenza (10-20 Hz) le componenti del vettore vento (u , v , w), le fluttuazioni di temperatura e la concentrazione di CO₂ e H₂O, rispettivamente attraverso un anemometro sonico (vento e T) ed un analizzatore di gas (di solito ad infrarossi, l'IRGA – Infra Red Gas Analyzer). I flussi vengono calcolati sulla base della covarianza tra la velocità verticale del vento e lo scalare (T , CO₂ etc.) (Aubinet et al., 2012). Per spiegare concettualmente il meccanismo consideriamo un ipotetico caso con turbolenza sviluppata rispetto a due misure nel tempo: la prima misura della componente verticale del vento ha direzione verso l'alto e concentrazione maggiore di CO₂; la seconda misura registra direzione verso il basso, a pari velocità, ma concentrazione inferiore di CO₂. Il risultato è uno scambio medio di CO₂, con perdita verso l'alto. Queste variazioni, integrate lungo il tempo di mediazione, danno il risultato di scambio netto medio tra ecosistema sottostante ed atmosfera.

La misura eseguita viene riportata in valori di scambio netto del parametro in esame (CO₂, calore latente, calore sensibile, ecc.) rispetto alla superficie e nel tempo (metro quadro al secondo).

I flussi di CO₂ misurati con questa tecnica sono il risultato di due processi ecofisiologici fondamentali:

1. di giorno, lo scambio netto di ecosistema (Net Ecosystem Exchange, NEE, equivalente alla NEP), è la differenza tra la respirazione dell'ecosistema (R) e la

fotosintesi (GPP) (negli ecosistemi decidui, in autunno ed inverno, anche di giorno può essere presente solamente il processo respirativo):

$$\text{NEE} = \text{R} - \text{GPP}$$

2. di notte, la NEE rappresenta la sola respirazione dell'ecosistema (R):

$$\text{NEE} = \text{R}$$

Per convenzione micrometeorologica il flusso netto NEE è negativo se in entrata nell'ecosistema (CO₂ sottratta all'atmosfera) mentre sia GPP che R vengono di solito riportati con valori positivi.

La misura diretta della NEP è di sicuro un vantaggio della tecnica ma per una comprensione completa dei meccanismi che ne regolano la dinamica è importante poter stimare anche le due componenti GPP e R.

A questo scopo sono stati sviluppati diversi approcci per la stima di GPP e R a partire dai dati di NEE misurata con l'eddy covariance (metodi di partitioning, vedi Desai et al., 2008). I due più utilizzati sono basati entrambi sulla parametrizzazione di modelli:

- Estrapolazione respirazione notturna: si basa sulla parametrizzazione di un modello di respirazione (dove la respirazione è funzione della temperatura ed in alcuni casi di contenuto idrico del suolo e tempo) utilizzando i dati di NEE misurati di notte. Il modello viene poi applicato anche di giorno per stimare la R e di conseguenza la GPP come differenza tra NEE e R. Uno dei modelli più applicati di questo genere è quello descritto in Reichstein et al., 2005.

- Modelli di NEE: questi modelli descrivono la NEE esplicitamente come differenza tra la GPP e R, dove la GPP è funzione della radiazione incidente (ed in alcuni casi del VPD) mentre la R è funzione della temperatura. Questi modelli vengono parametrizzati con i dati di giorno e poi applicati anche ai dati notturni e danno come risultati direttamente le due componenti GPP e R. Il più usato di questi modelli è quello descritto da Lasslop et al., 2010.

Con la tecnica dell'eddy covariance si ricava quindi una misura dello scambio netto dell'ecosistema ed è possibile stimare anche le due componenti di fotosintesi e respirazione. Questi dati possono essere utilizzati non solo per il calcolo della capacità di sequestro di carbonio ma anche per lo studio della sensibilità dell'ecosistema alle variazioni del clima, ovvero si calcola la risposta dell'ecosistema sia in termini di breve periodo (su base semioraria) che di lungo periodo (su base stagionale o annua) alle variazioni ambientali nei processi di controllo dell'assorbimento di anidride carbonica e sullo scambio di vapore acqueo.

Oggi questa tecnica di misura si è molto diffusa, grazie soprattutto alla commercializzazione di strumenti scientifici con alte performance di analisi e acquisizione a costi relativamente accessibili.

Esistono circa 700 siti eddy covariance in tutto il mondo in cui vengono monitorati gli scambi di gas serra e energia a scala ecosistemica. Tali siti coprono differenti land-use e condizioni climatiche e sono coordinati a livello mondiale da reti regionali e dall'iniziativa FLUXNET.

Il database FLUXNET contiene dati eddy-covariance registrati ad una frequenza di 30 minuti e dati aggregati giornalieri, settimanali e mensili. Tali dati sono accompagnati da un indice di qualità che indica la qualità del dato e quindi la sua affidabilità.

La vegetazione esaminata include foreste temperate di conifere e latifoglie (decidue e sempreverdi), foreste tropicali e boreali, crops, grasslands, chaparral, wetlands e tundra. Le torri eddy covariance operano su cinque continenti e il loro range di distribuzione latitudinale varia da 70 gradi nord a 30 gradi sud.

Studi basati sull'utilizzo di stazioni di misura locali hanno permesso di caratterizzare con precisione le dinamiche giornaliere, intra ed inter-annuali su un sempre più vasto spettro di ecosistemi e condizioni ambientali e lo studio delle relazioni tra flussi e clima, tipo di vegetazione, disturbi e gestione degli ecosistemi (e.g. Falge et al., 2002; Reichstein et al., 2007; M. Aubinet et al., 2001; Clark et al., 2004; Dolman et al., 2002).

1.2.I controlli climatici della GPP

Diversi sono i fattori che possono incidere sul bilancio del carbonio a livello di ecosistema: la tipologia di vegetazione, la sua densità, le specie presenti, le tecniche di gestione, ma anche elementi del clima quali temperatura, illuminazione, piovosità e la loro variabilità.

Molti sono stati gli studi che hanno approfondito il controllo dei driver ambientali. Reichstein et al. (2007) hanno individuato la disponibilità di acqua come il controllo principale dell'assorbimento di carbonio. A livello *above-ground la net primary productivity* è fortemente influenzata per diversi biomi dall'evapotraspirazione (Rosenzweig, 1968; Lieth & Whittaker, 1975) e anche dalle precipitazioni (Huxman et al., 2004; Lieth & Whittaker, 1975).

I controlli sulla variabilità spaziale della GPP sono stati approfonditi da diversi autori: Valentini et al. (2000) ha dimostrato che la GPP non dipende dal cambiamento di latitudine per una vasta gamma di foreste europee. Lavori successivi (Law et al., 2002) mostrano che la temperatura media annuale e il bilancio idrico spiegano la maggior parte della variazione lungo differenti biomi.

Globalmente the GPP delle foreste aumenta con l'aumento delle temperature e delle precipitazioni (Luyssaert et al., 2007). Altri autori hanno individuato nella disponibilità d'acqua il principale controllo della GPP per le foreste del sud Europa, mentre la temperatura è considerata il più importante controllo per le foreste più a nord dell'Europa (Reichstein et al., 2007).

Lungo una grande diversità di ecosistemi asiatici, sia le precipitazioni che le temperature sono considerate i principali drivers della GPP (Kato & Tang, 2008).

La variabilità stagionale dello stoccaggio di carbonio dipende dai limiti fisiologici e strutturali dell'ecosistema. Il forte legame positive tra area fogliare e fotosintesi si verifica in tipi vegetazionali come la tundra (Boelman et al., 2005), foreste temperate decidue (Waring et al., 1995) e cropland annuali, dove area fogliare, PAR e limitazioni climatiche covariano per tutta la stagione di crescita.

La mancanza di stagionalità nell'FPAR è associata ad una bassa correlazione con l'assorbimento di carbonio nella maggior parte delle tipologie sempreverdi di vegetazione, come le foreste pluviali o le foreste mediterranee (Sims et al., 2006; Garbulsky et al., 2008). La crescita e l'adattamento della vegetazione all'ambiente circostante sono più influenzati dalla disponibilità d'acqua piuttosto che dalla temperatura (Garbulsky et al., 2010).

La variabilità annuale dell'efficienza d'uso della radiazione è largamente spiegata delle precipitazioni piuttosto che dal tipo di vegetazione (Garbulsky et al., 2010). Inoltre nelle variazioni intra-annuali della fotosintesi, VPD e temperatura risultano i driver più importanti (Garbulsky et al., 2010).

1.3. Analisi delle relazioni tra flussi e variabili climatiche

Per approfondire le dinamiche ecofisiologiche all'interno di diversi ecosistemi è importante stimare le relazioni tra flussi eddy covariance e variabili climatiche.

Tali relazioni possono essere analizzate tramite due tipologie di modelli, una di tipo ipotetico-deduttivo e un'altra di tipo induttivo. Il primo approccio si basa su un'ipotesi iniziale su come le variabili climatiche lavorano all'interno dell'ecosistema. I processi dei controlli sono poi implementati in un modello ecosistemico come equazioni parametrizzate (deduzione). I dataset dei flussi di carbonio sono usati per settare i parametri e testare la validità del modello (Moffat, Beckstein, Churkina, Mund, & Heimann, 2010).

In questo studio è stato utilizzato un approccio di tipo induttivo dove sono evitate assunzioni a priori e si basa su modelli empirici quali le ANN (*Artificial Neural Network*) e le RF (*Random Forests*). Le relazioni funzionali tra i controlli climatici (drivers) e i flussi di carbonio sono dedotte solamente e direttamente dalle osservazioni.

Queste relazioni puramente empiriche sono poi utilizzate per caratterizzare la risposta dell'ecosistema ai driver climatici, ad esempio la gerarchia dei controlli, le dipendenze multivariate e la sensibilità delle risposte (Moffat et al., 2010).

Ad oggi esistono serie di dati eddy covariance che ricoprono, in alcuni siti, più di 15 anni e possono essere usati per la parametrizzazione dei modelli empirici e trattati con metodologie statistiche per l'individuazione di pattern, soglie, trend legati in particolare alla variabilità climatica e della gestione degli ecosistemi.

1.4.Reti Neurali Artificiali

Le reti neurali artificiali sono dei sistemi di elaborazione dell'informazione che cercano di simulare all'interno di un sistema informatico il funzionamento dei sistemi nervosi biologici che sono costituiti da un gran numero di cellule nervose o neuroni collegati tra di loro in una complessa rete (Gallo, 2007). Nel cervello umano ogni neurone è collegato mediamente con una decina di migliaia di altri neuroni. Si hanno quindi centinaia di miliardi di connessioni. Il comportamento intelligente emerge dalle numerose interazioni tra le unità interconnesse. Alcune di queste unità ricevono informazioni dall'ambiente, altre emettono risposte nell'ambiente e altre ancora - se ve ne sono - comunicano solamente con le unità all'interno della rete.

Nelle reti neurali artificiali i neuroni sono distribuiti su diversi strati e in base al loro posizionamento all'interno degli strati sono definiti rispettivamente unità di ingresso (input), unità di uscita (output) e unità nascoste (hidden) (Figura 1).

Ogni unità è collegata alle unità dello strato successivo tramite delle connessioni. Ciascuna unità svolge un'operazione molto semplice che consiste nel diventare attiva se la quantità totale di segnale che riceve supera una certa soglia di attivazione. Se un'unità diventa attiva, essa emette un segnale che viene trasmesso lungo i canali di comunicazione sino alle altre unità cui essa è connessa.

All'interno di ogni neurone vi è una funzione di trasferimento (memoria locale) che consiste in una relazione tra il livello di attivazione interno del neurone (chiamata funzione di attivazione) e l'output. Una tipica funzione è quella sigmoide che varia tra 0 e 1 per un range di input. Una funzione $f(x)$ sarà una funzione sigmoide se essa è delimitata e il suo valore crescerà con l'incremento di x (Melesse & Hanley, 2005).

La funzione di trasferimento, trasforma il messaggio ricevuto in un segnale inibitorio o eccitatorio aumentandone o diminuendone nel contempo l'intensità a seconda delle proprie caratteristiche individuali.

Tale segnale è rappresentato da un peso (W) che viene associato alla connessione tra i vari neuroni della rete.

Lo strato intermedio somma gli input pesati e usa la funzione di trasferimento per creare un valore di output.

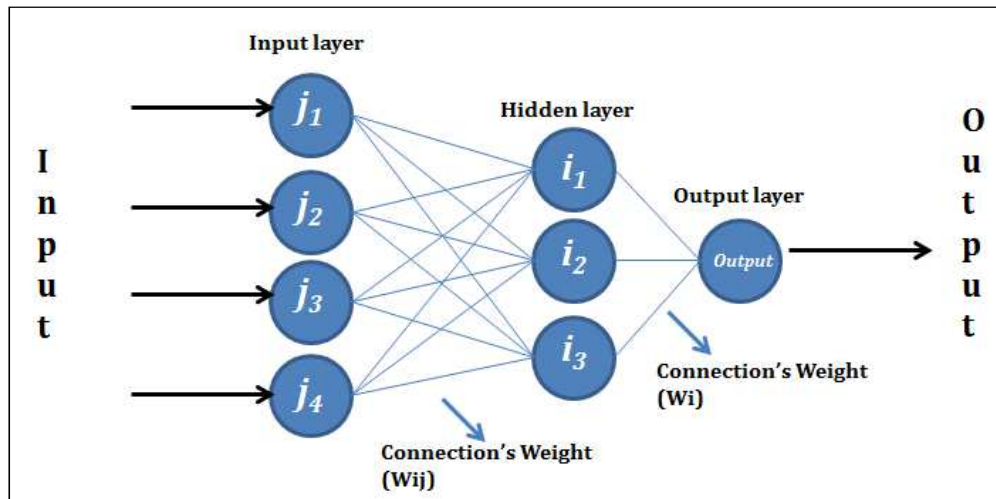


Figura 1: Esempio di rete neurale artificiale.

Il legame input-output, non viene programmato ma è semplicemente ottenuto da un processo di apprendimento basato su dati empirici che può essere:

- a) Supervisionato (*supervised learning*), qualora si disponga di un insieme di dati per l'addestramento (o training set), comprendente esempi tipici di ingressi con le relative uscite loro corrispondenti: in tal modo la rete può imparare ad inferire la relazione che li lega. Successivamente, la rete è addestrata mediante un opportuno algoritmo (tipicamente il *backpropagation* che è appunto un algoritmo d'apprendimento supervisionato), il quale usa tali dati allo scopo di modificare i pesi ed altri parametri della rete stessa in modo tale da minimizzare l'errore di previsione relativo all'insieme di addestramento. Se l'addestramento ha successo, la rete impara a riconoscere la relazione incognita che lega le variabili d'ingresso a quelle d'uscita, ed è quindi in grado di fare previsioni anche laddove l'uscita non è nota a priori; in altri termini, l'obiettivo finale dell'apprendimento supervisionato è la previsione del valore dell'uscita per ogni valore valido dell'ingresso, basandosi soltanto su un numero limitato di esempi di corrispondenza (vale a dire, coppie di valori input-output). Per fare ciò, la rete deve essere infine dotata di un'adeguata capacità di generalizzazione, con riferimento a casi ad essa ignoti. Ciò consente di risolvere problemi di regressione o classificazione.

- b) Non supervisionato (*unsupervised learning*), basato su algoritmi d'addestramento che modificano i pesi della rete facendo esclusivamente riferimento ad un insieme di dati che include le sole variabili d'ingresso. Tali algoritmi tentano di raggruppare i dati d'ingresso e di individuare pertanto degli opportuni cluster rappresentativi dei dati stessi, facendo uso tipicamente di metodi topologici o probabilistici. L'apprendimento non supervisionato è anche impiegato per sviluppare tecniche di compressione dei dati.

Le Reti Neurali sono in sostanza un sofisticato sistema di tipo statistico, altamente non lineare, dotato di una buona immunità al rumore; se alcune unità del sistema dovessero funzionare male, la rete nel suo complesso avrebbe delle riduzioni di prestazioni ma difficilmente andrebbe incontro ad un blocco del sistema. Lek & Guuégan (1999) hanno riportato le diverse applicazioni delle ANN e ne hanno sottolineato l'importanza. Infatti le ANN rappresentano uno strumento potente per modellizzare relazioni sconosciute all'interno di dataset complessi.

L'applicazione principale delle Reti Neurali Artificiali (ANNs) è rappresentata dallo sviluppo di modelli predittivi. Tali modelli vengono creati per prevedere valori futuri di una particolare variabile di risposta da un dato insieme di variabili indipendenti.

La valutazione dei modelli può essere effettuata tramite l'analisi della *Nash–Sutcliffe model efficiency - meff* (Nash & Sutcliffe, 1970). La *meff* rappresenta infatti una misura della capacità predittiva di un modello rispetto ad un riferimento. Varia tra 0 e 1. Tanto più tende ad 1, tanto maggiore è la capacità predittiva del modello rispetto ai valori di riferimento.

Diversi studi hanno utilizzato le ANN per spiegare le relazioni non lineari tra le variabili in ecologia, essendo uno dei metodi più idonei per illustrare queste interazioni. Ad esempio Lek et al., (1996) le hanno applicate nel loro studio proprio per individuare diverse relazioni ecologiche; Recknagel, 1997 le ha utilizzate per predire l'abbondanza e la successione di specie di alghe in un lago del Giappone. Mastorillo et al. (1998); Brosse et al. (1999) e Laë et al. (1999) le hanno testate per individuare le relazioni tra la ricchezza di specie di pesci all'interno di diversi ecosistemi acquatici e le caratteristiche chimiche e topologiche di tali ecosistemi. Maier & Dandy (2000) e Liong et al. (2000) hanno applicato le ANN come strumento di predizione delle risorse idriche.

Partendo da una o più variabili in input, le ANNs hanno la capacità di predire la variabile di output ma i meccanismi che si verificano all'interno della rete sono sconosciuti. Anche se il valore predittivo delle ANNs è di grande aiuto a molti ecologi, i ricercatori hanno spesso criticato il valore esplicativo delle reti neurali, definendolo un approccio "black box" per modellare fenomeni ecologici (Paruelo & Tomasel, 1997; Lek & Guégan, 1999).

Diversi autori hanno esplorato questo problema e proposto algoritmi per illustrare il ruolo delle variabili all'interno dei modelli delle reti. Ci sono differenti metodi per individuare i contributi delle variabili indipendenti all'interno delle ANN (Gevrey et al., 2003; Olden & Jackson, 2002).

1.5. Metodologie per individuare l'importanza dei driver all'interno delle ANN

Esistono diversi metodi che hanno lo scopo di individuare una gerarchia tra i driver in input per vedere quali guidano maggiormente la variabile in output.

1.5.1. Garson's algorithm

Questo metodo usa i valori assoluti dei pesi delle connessioni per calcolare i contributi delle variabili. Il metodo essenzialmente comporta la partizione dei pesi delle connessioni *hidden-output* in componenti associate con ogni neurone in input (Garson, 1991; Gevrey et al., 2003).

I contributi di ogni neurone di input j (Q_j) sono calcolati come il prodotto tra il peso della connessione *input-hidden* (W_{ij}) e il peso associato alla connessione *hidden-output* (W_i). Per ogni input, si ottengono tanti valori quanti sono i nodi del *hidden layer*.

Il contributo relativo di ogni variabile in input viene calcolato prima per ogni neurone del *hidden layer* in relazione all'output (P_{ij}):

$$1. \quad P_{ij} = |W_i| \times |W_{ij}|$$

dove:

W_i = peso della connessione tra il neurone dell'*hidden layer* i e l'output

W_{ij} = peso della connessione tra il neurone dell'*hidden layer* i e il neurone dell'input j

Poi viene calcolato il peso relativo di ogni neurone dell'*hidden layer* in relazione all'input (Q_{ij}):

$$2. \quad Q_{ij} = \frac{P_{ij}}{\sum_{i=1}^n P_{ij}}$$

Dalla somma dei contributi relativi ottenuti per ogni neurone di input (Q_{ij}), si ottiene l'importanza di ogni variabile in input (S_j):

$$3. \quad S_j = \sum_{i=1}^n Q_{ij}$$

Infine viene calcolata l'importanza relativa di ogni variabile in input (RI_j):

$$4. \quad RI_j = \frac{S_j}{\sum_{j=1}^n S_j}$$

L'algoritmo di Garson può essere utilizzato solo in presenza di Reti Neurali con un solo *hidden layer*. Gevrey et al. (2003) ha suggerito di semplificare questo algoritmo in modo da renderlo indipendente dal numero di strati intermedi, ma considerando solo le connessioni tra lo strato di input e il primo *hidden layer*, ottenendo lo stesso identico risultato:

$$1. \quad Q_{ij} = \frac{W_{ij}}{\sum_{i=1}^n W_{ij}}$$

Viene quindi eliminato il primo passaggio dal metodo di Garson (1991), poi il procedimento per calcolare le RI di ogni variabile in input rimane lo stesso.

1.5.1.1. Connection Weight:

In questo approccio proposto successivamente da Olden et al. (2004) viene rivisto il metodo proposto da Gevrey et al. (2003). Vengono riprese in considerazione le connessioni *hidden-output*, che contengono un'informazione importante che non può essere trascurata. Il procedimento per arrivare all'importanza relativa di ogni driver (RI) è il seguente:

$$1. \quad P_{ij} = W_i \times W_{ij}$$

$$2. \quad Q_j = \sum_{i=1}^n P_{ij}$$

$$3. \quad RI_j = \frac{Q_j}{\sum_{j=1}^n Q_j}$$

In questo modo, il metodo dei pesi delle connessioni appare più robusto e tiene in considerazione anche lo strato di output, permettendo un'analisi più accurata dei pesi che la rete attribuisce alle variabili in input (Olden et al., 2004).

1.5.2. Stepwise methods

Questo metodo consiste nell'aggiungere o eliminare una variabile in input volta per volta ed osservare gli effetti sul risultato in output (Gevrey et al., 2003).

1.5.2.1. Forward stepwise

La prima rete viene addestrata con tutte le variabili a disposizione nel dataset in input. Assumendo che rappresenti la ANN migliore che si potesse creare, la sua performance (rappresentata dal coefficiente R^2) viene presa come misura di riferimento rappresentando il modello più affidabile in quanto costituito da tutte le informazioni a disposizione (Moffat et al., 2010).

Dopo aver determinato il modello di riferimento, N modelli vengono generati, ognuno usando solamente una delle variabili input. Successivamente $N-1$ modelli vengono creati, combinando la variabile il cui modello risultava avere il più piccolo errore con ognuna delle rimanenti variabili. Questa procedura viene ripetuta utilizzando modelli con tre variabili in input, poi quattro e così via fino ad aggiungere tutte le variabili a disposizione in input.

L'ordine di integrazione delle variabili nella rete rappresenta l'ordine di importanza dei loro contributi nella creazione dell'output.

1.5.2.2. Backward method

N modelli vengono creati, ognuno utilizzando $N-1$ variabili a disposizione in input. La rete con la perdita di performance maggiore rispetto al modello di riferimento (addestrato con tutte le variabili in input a disposizione) sarà quello in cui mancherà la variabile in input più importante. Successivamente $N-1$ reti vengono create, combinando $N-2$ variabili (eliminando per ogni modello, la variabile risultata più importante dai modelli precedenti e una delle altre variabili rimaste). Questo procedimento viene ripetuto creando ANNs con $N-3$, $N-4$, ..., fino a $N-(N-1)$ variabili (Gevrey et al., 2003).

L'ordine in cui vengono eliminate le variabili in input nella rete corrisponderà all'ordine di importanza dei loro contributi.

1.5.3. Perturbation

Questo metodo si basa sull'effetto, sull'output della rete neurale, di piccoli cambiamenti in ogni variabile di input. L'algoritmo modifica il valore di una variabile in input per volta, lasciando i

valori delle altre invariati. Vengono verificate le risposte della variabile di output per ogni cambiamento sugli input. L'errore medio quadratico (MSE) dell'output della rete neurale aumenterà in modo proporzionale all'aggiunta o alla rimozione di rumore alla variabile di input.

Questi cambiamenti possono assumere la forma di $x_i = x_i \pm \delta$ dove x_i è la variabile di input selezionata e δ è il cambiamento. δ può essere aumentato o diminuito in steps del 5% del valore della variabile da modificare fino al 50%. In base all'MSE può essere ottenuta una classificazione di importanza delle singole variabili di input (Gevrey et al., 2003; Olden et al., 2004).

1.5.4. Destroy order

Questa metodologia distrugge l'ordine dei valori di una variabile di input per volta, lasciando le altre invariate. Vengono quindi analizzate le risposte della variabile di output ai diversi livelli di distruzione delle variabili in input.

L'MSE della rete neurale sarà tanto più grande, quanto maggiore sarà il livello di distruzione della variabile input selezionata.

La quantità di dati della variabile di input che viene distrutta (riordinata in modo random) varia in questo modo: 25%, 50%, 75% and 100%.

Il concetto è lo stesso del precedente metodo. Anche in questo caso, viene ottenuta, sulla base dell'MSE, una classificazione di importanza delle variabili di input.

1.6. Random Forests

Le *Random Forests* (RF) sono dei popolari ed efficienti algoritmi, introdotti da Breiman (2001) e basati su un modello di aggregazione di due tipologie: modelli di regressione e modelli di classificazione. Appartengono alla famiglia dei metodi di *ensemble*, entrando a far parte delle macchine di apprendimento alla fine degli anni novanta (Dietterich, 2000).

Considerando un set di apprendimento $L = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ costituito da n osservazioni di un vettore random (X, Y) . Il vettore $X = (X^1, \dots, X^n)$ contiene i predittori o le variabili esplicative (Genuer, Poggi, & Tuleau-Malot, 2010).

Il principio delle RF è combinare molti alberi di decisione binaria (*split decision*) costruiti usando differenti campioni estratti dal campione di apprendimento L (*bagging*) e scegliendo in modo random ad ogni nodo un subset di variabili esplicative X .

La risposta di ogni albero dipende dal subset L scelto indipendentemente (con replicazione) e con la stessa distribuzione per tutti gli alberi nella foresta. La lunghezza ottimale di L dipende dal numero di input del dataset originale.

Per i problemi di classificazione, preso un set di alberi semplici e un set di predittori random, la RF definisce una funzione margine che misura il limite per cui il valore medio associato dall'albero alla classe corretta supera il valore medio per ciascuna altra classe presente nelle variabili dipendenti.

Per i problemi di regressione, le Random Forest sono formate dall'accrescimento di ogni singolo albero, ciascuno in grado di produrre un valore numerico di risposta. Tale valore numerico di risposta è volto a minimizzare l'errore dell'albero. Anche in questo caso il set di predittori è selezionato casualmente e con la stessa distribuzione per tutti gli alberi.

Il *mean-square error (mse)* per una Random Forest è così ottenuto:

$$mse = (OOB - tree\ response)^2$$

Dove il campione **OOB** è l'insieme di osservazioni che non vengono utilizzate per costruire l'albero corrente (*Out of Bag*). Viene utilizzato per stimare l'errore di predizione e quindi valutare l'importanza delle variabili all'interno del modello.

I due parametri principali da settare per costruire una RF sono l'*mtry*, il numero di variabili di input scelte in modo casuale ad ogni divisione (ramificazione) e l'*ntree*, il numero di alberi nella foresta.

L'algoritmo RF è molto potente e viene utilizzato in molte applicazioni differenti (Díaz-Uriarte & Alvarez de Andrés (2006) lo hanno utilizzato per l'analisi di dati di espressione genica).

Genuer et al. (2010) e Grömping (2009) hanno utilizzato le Random Forest per stimare il rank di importanza di alcune variabili all'interno di differenti processi e confrontare la loro efficacia rispetto alle regressioni lineari.

Anche per le Random Forest, un utile parametro per valutarne le performance è la *meff*.

1.6.1 Stima dell'importanza dei driver con le RF

Nelle foreste casuali, il parametro con cui viene valutata l'importanza di una data variabile consiste quindi nell'incremento medio dell'errore di un albero (*mse*) per la regressione e il tasso di errata classificazione per la classificazione (*misclassification rate*) (Genuer et al., 2010).

Per ottenere un rank delle variabili in input si considera l'importanza della variabile X_j che è definita come segue:

- per ogni albero t della foresta, viene preso in considerazione il campione OOB_t associato (set di dati non inclusi nel campione estratto utilizzato per la costruzione di t).
- Viene poi calcolato l'errore err_{OOB_t} (*mse* per la regressione o *misclassification rate* per la classificazione) di un singolo albero t sul campione OOB_t .

- In modo random vengono poi permutati i valori della variabile X_j nell'OOBt per calcolare l'errore OOBt j , cioè l'errore del predittore t sul campione perturbato.
- L'importanza della variabile sarà data dall'incremento dell'errore del modello (differenza tra mse dopo la perturbazione della variabile X_j e mse dell'albero prima della perturbazione):

$$\text{Imp}(X_i) = \text{ERR}(\text{perm}(X_i)) - \text{ERR}(X)$$

dove:

$X_i = \text{driver}$

$\text{Imp}(X_i) = \text{importanza stimata del driver } X_i \text{ sull'insieme dei dati OOB}$

$\text{ERR}(\text{perm}(X_i)) = \text{errore (MSE) stimato sull'insieme OOB permutando } X_i$

$\text{ERR}(X) = \text{errore sull'insieme OOB mantenendo l'ordine originale dei dati}$

Tanto più risulterà grande la differenza tra i due errori, più importante sarà il driver perturbato (Genuer et al., 2010).

1.7.SSA (Singular System Analysis)

Le variabili climatiche contengono al loro interno dei *patterns* di breve, medio e lungo periodo. Tali andamenti possono spiegare le interazioni tra ecosistema e atmosfera su diverse scale temporali.

Disponendo di serie climatiche lunghe, è possibile estrarre i sottosegnali delle variabili ecologiche per indagare quali drivers incidono maggiormente sui vari *patterns* della GPP.

Le dipendenze a scala temporale tra le variabili sono state indagate per il breve periodo (da andamenti orari fino a quelli stagionali) per diversi aspetti ecofisiologici, statistici e micro-metereologici (Baldocchi, 2003). Inoltre, è noto che alcune variabili driver (come temperature e precipitazioni) possono mostrare oscillazioni di bassa frequenza e trend (Mahecha et al., 2007).

Grazie alla tecnica dell'SSA si possono individuare, all'interno di dati di flussi di medio e lungo periodo, se presenti, tali pattern ad alta e bassa frequenza.

Lo scopo dell'SSA (*Singular System Analysis*) è infatti quello di identificare i sottosegnali di una determinata serie storica $X(t)$, dove t va da 1 a N , e proiettarli alle corrispondenti scale temporali. La serie temporale (centrata a media zero) viene sottoposta alla SSA, che può essere descritta come una procedura costituita da due fasi principali: una decomposizione del segnale e una ricostruzione del segnale (Golyandina & Korobeynikov, 2001).

La decomposizione mira a trovare rilevanti funzioni ortogonali, che consentono la parziale o, se necessaria, la completa ricostruzione delle serie storiche. Le analisi hanno bisogno di definire a priori un *embedding dimension*, che consiste in una finestra di lunghezza P . Lo Scorrimento della finestra lungo la serie temporale porta alla costruzione di una *trajectory matrix* formata dalla sequenza di $K = N - P + 1$ vettori *time-lagged* della serie originale.

I vettori della *trajectory matrix* $\mathbf{Z}$ sono create come segue (Golyandina & Korobeynikov, 2001):

$$\mathbf{Z}_i = (X_i; \dots; X_{i+P-1})^T$$

dove:

$$1 \leq i \leq K$$

Viene costruita una matrice di covarianza $\mathbf{C} = \{c_{i,j}\}$ di dimensioni $P \times P$ sulla base della *trajectory matrix* $\mathbf{Z}$, che in accordo con Vautard, Yiou, & Ghil (1992) può essere stimata direttamente dai dati nella forma di una matrice di Toeplitz:

$$c_{i,j} = \frac{1}{(N - |i - j|) \times \sum_{t=1}^{N-|i-j|} (X_t \times X_{t+|i-j|})}$$

Le voci della risultante matrice $P \times P$ rappresentano la covarianza catturata e dipendono solamente dal ritardo $|i-j|$, dove $i, j = 1, \dots, P$. Broomhead & King (1986), in accordo con Ghil et al. (2002) ha proposto un approccio leggermente diverso per la costruzione della matrice di covarianza.

$$\mathbf{C} = \frac{1}{K} \times \mathbf{Z}^T \times \mathbf{Z}$$

Tale approccio è più appropriato per segnali non stazionari. Basandosi sulle due equazioni precedenti si può determinare la base ortonormale dalla seguente equazione:

$$\mathbf{E}^T \times \mathbf{C} \times \mathbf{E} = \mathbf{\Lambda}$$

In questa equazione, $\mathbf{E}$ è una matrice $P \times P$ contenente gli *eigenvectors* E_i , anche chiamati funzioni ortogonali empiriche (EOFs) di $\mathbf{C}$.

La matrice $\mathbf{\Lambda}$ nella diagonale contiene i rispettivi *eigenvalues*, ordinati per convenzione in ordine decrescente:

$$diag(\lambda) = (\lambda_1, \dots, \lambda_P)$$

dove:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_P$$

Può essere dimostrato che per le proprietà della matrice di covarianza $\mathbf{C}$ – preservando la simmetria ed essendo semidefinita da valori reali e positivi – tutti gli *eigenvectors* e *eigenvalues* sono valori reali, dove questi ultimi sono scalari non negativi. Gli *eigenvalues* sono proporzionali alla frazione di varianza spiegata corrispondente ad ogni EOF. In analogia alla meglio conosciuta *Principal Component Analysis*, la decomposizione consente la costruzione delle componenti principali (PCs) come serie storiche generate rappresentanti le mode ortogonali estratte. Questo è il motivo per cui l'SSA è spesso chiamata una “*PCA in the time domain*”.

Le componenti principali $\mathbf{A}^k$ sono così calcolate:

$$A_t^K = \sum_{j=1}^P X_{t+j-1} \times E_j^K$$

dove:

$$1 \leq K \leq P$$

L'ultimo passaggio della SSA è la ricostruzione delle serie storiche attraverso le componenti principali. Il segnale originale può essere ricostruito interamente o parzialmente. Questo è uno *step* selettivo e l'analista può decidere quale $\mathbf{A}^k$ combinare per ottenere una combinazione interpretabile delle componenti principali.

Questo consente la separazione del segnale dal rumore e la ricostruzione di componenti con frequenze specifiche selezionate, come illustrato dalla seguente equazione:

$$R_t^K = \frac{1}{M_t} \sum_{k \in K} k \times \sum_{j=L_t}^{U_t} A_{t-j+1}^K \times E_j^K$$

Nella tabella (Tabella 1) seguente sono riportati i valori del fattore di normalizzazione $\mathbf{M}_t$ e dei limiti, più basso $\mathbf{L}_t$ e più alto $\mathbf{U}_t$, della somma:

Temporal locations	M_t	L_t	U_t
for $1 \leq t \leq P-1$	t^{-1}	1	t
for $P \leq t \leq K$	P^{-1}	1	P
for $K+1 \leq t \leq N$	$(N-t+1)^{-1}$	$t-N+P$	P

Tabella 1: Parametri per la ricostruzione delle serie storiche.

In questa procedura di ricostruzione, è un insieme di indici che determina la selezione delle modalità utilizzate per la ricostruzione, M_t è un fattore di normalizzazione, mentre l'estensione corrispondente per i limiti delle serie è ottenuta da L_t e U_t .

La ricostruzione selettiva delle serie temporali permette di rappresentare il comportamento delle serie a diverse scale temporali.

La scala temporale della variazione corrispondente a un EOF o PC può essere trovata dall'analisi del rispettivo spettro di potenza. Le singole mode solitamente hanno uno spettro molto semplice, essendo dominate solamente da una singola frequenza dominante.

Vautard et al. (1992) ha sottolineato che la somma degli spettri di potenza delle PCs conserva l'andamento principale dello spettro di potenza della serie originale.

Tuttavia, l'*embedding dimension* P pone alcuni limiti: la più bassa frequenza ricoperta dalle mode individuali ha un periodo P (Ghil et al., 2002). Le periodicità di lunghezza P o corrispondono ad oscillazioni con periodo P o sono indotte da trend (possibilmente non lineari) (Yiou, Sornette, & Ghil, 2000).

Dato che uno degli obiettivi di questo lavoro è quello di esplorare anche le strutture a lungo raggio nei dati, si è cercato il massimo valore attendibile della dimensione di *embedding*. La scelta di P è infatti un *trade-off*. Da un lato, la massimizzazione del contenuto informativo dell'analisi richiede una grande finestra P ; d'altra parte è cruciale per ottimizzare la confidenza statistica della decomposizione usare un alto numero di canali K . Questo bilanciamento è espresso dal rapporto N/P . Il più basso rapporto riportato in letteratura è $N/P = 2.5$ (Golyandina & Korobeynikov, 2001).

1.8.Scopo della ricerca

La ricerca è volta a studiare i fattori ambientali che influenzano la GPP di ecosistemi terrestri tramite metodi empirici e tecniche di data-mining, anche scomponendo i segnali in differenti frequenze tramite l'SSA.

I dati di partenza provengono da diversi siti eddy-covariance, che ricoprono vari ecosistemi e latitudini e sono caratterizzati da differenti tipologie di vegetazione e regimi climatici.

Lo scopo finale è quello di valutare queste metodologie per individuare una gerarchia delle variabili climatiche nel regolare la GPP e classificare conseguentemente i siti in funzione della loro dipendenza da specifici fattori in modo da poterne analizzare la sensibilità e comprendere i meccanismi ambientali che influenzano i flussi di anidride carbonica a lungo termine in differenti siti a livello globale.

Per raggiungere l'obiettivo, il lavoro è stato svolto in diverse fasi:

1. Selezione delle variabili driver e dei siti eddy-covariance.
2. Scelta delle strutture e procedura di parametrizzazione dei modelli (ANN e RF).
3. Validazione e selezione dei metodi per la stima dell'importanza dei drivers tramite dataset artificiale.
4. Applicazione dei metodi selezionati alle serie giornaliere di GPP e drivers nei siti eddy-covariance selezionati.
5. Scomposizione, tramite SSA, delle serie di GPP e drivers nelle principali componenti (stagionalità, trend e residui).
6. Applicazione dei modelli e dei metodi per la stima dell'importanza dei drivers alle componenti trend e residui per individuare i driver che guidano maggiormente la GPP nel breve e nel lungo periodo al netto della stagionalità.
7. Sviluppo ed analisi di un indice sintetico in grado di poter confrontare i risultati ottenuti impiegando le diverse metodologie e tra i diversi siti.

2. Materiali e Metodi

2.1. Drivers e Siti eddy-covariance selezionati

In questo studio sono stati utilizzati dati eddy-covariance misurati in siti principalmente in Europa e Stati Uniti, e disponibili nel database FLUXNET 2015 (<http://fluxnet.fluxdata.org>). Il database contiene dati già filtrati e processati ed anche le componenti GPP e R ottenute tramite diversi metodi di partitioning. Nello specifico è stata utilizzata la GPP ottenuta con il partitioning basato sul metodo di Lasslop et al. (2010) (GPP_DT_VUT_REF) mentre i dati meteo impiegati sono gapfilled con il metodo MDS (Marginal Distribution Sampling – Reichstein et al., 2005).

I dati utilizzati sono a risoluzione giornaliera e sono forniti assieme ad un indicatore di qualità che varia tra 0 e 1 ed è relativo alla percentuale di dati di buona qualità utilizzati per il calcolo del valore giornaliero. In questo studio sono stati utilizzati solo i dati con qualità superiore a 0.6 (almeno il 60% di misure di buona qualità).

Per predire la GPP a livello di sito, sono stati utilizzati i seguenti drivers: temperatura dell'aria (TA), precipitazioni (P), radiazione solare incidente (SW_IN), deficit di pressione di vapore (VPD), temperatura del suolo (TS), contenuto idrico del suolo (SWC). Queste variabili coprono le principali caratteristiche climatiche legate alla produttività della vegetazione (temperatura, radiazione, acqua) e sono generalmente disponibili nella maggior parte dei siti.

Le precipitazioni sono caratterizzate da un valori discontinui e non tengono conto dell'effetto distribuito nel tempo. Per questo motivo le precipitazioni sono state trasformate in una variabile continua, distribuendo nei 10 giorni successivi all'evento di pioggia, la quantità d'acqua precipitata, secondo questa funzione:

$$P_{Tj} = P_{Tj-1} \times (10) + \frac{P_j}{10}$$

dove:

P_{Tj} : è il valore j relativo alle precipitazioni dopo la trasformazione

P_{Tj-1} : è il valore trasformato corrispondente al dato $j-1$

P_j : è il valore j relativo alle precipitazioni prima della trasformazione

Per la selezione dei siti eddy-covariance da utilizzare nello studio sono state prese in considerazione due condizioni:

1. I siti dovevano avere una lunghezza della serie storica di dati pari o superiore a sei anni, in modo da poter valutare le dinamiche interannuali.
2. Tali siti dovevano contenere tutte e sei le variabili climatiche selezionate.

In questo modo sono stati individuati 38 siti (Figura 2) con i seguenti i codici identificativi (SITE_ID) riportati in Tabella 2. Le caratteristiche di ogni sito (climatiche, pedologiche, vegetazionali e gestionali) sono descritte in appendice A.

SITE_ID	SITE_NAME	STATE
AT-Neu	Neustift/Stubai Valley	Austria
AU-DaP	Daly River Savanna	Australia
AU-DaS	Daly River Cleared	Australia
AU-Tum	Tumbarumba	Australia
BE-Lon	Lonze	Belgio
BE-Vie	Vielsalm	Belgio
CA-Qfo	Eastern Boreal	Canada
CA-SF3	Saskatchewan - Western Boreal	Canada
CH-Oe1	Oensingen1 grass	Svizzera
DE-Hai	Hainich	Germania
DE-Kli	Klingenberg – cropland	Germania
DE-Obe	Oberbarenburg	Germania
DE-Tha	Anchor Station Tharandt - old spruce	Germania
DK-Sor	Soroe- LilleBogeskov	Danimarca
DK-ZaH	Zackenbergh Heath	Groenlandia
FI-Hyy	Hyytiala	Finlandia
FR-Gri	Grignon	Francia
GF-Guy	Guyana	Guyana Francese
IT-Lav	Lavarone	Italia
IT-Noe	Sardinia/Arca di Noè	Italia
IT-Ren	Renon/Ritten (Bolzano)	Italia
IT-Ro1	Roccarespampani	Italia
IT-SRo	San Rossore	Italia
IT-Tor	Torgnon	Italia
NL-Hor	Horstermeer	Olanda
NL-Loo	Loobos	Olanda
RU-Fyo	Fyodorovskoye wet spruce stand	Russia
US-ARM	ARM Southern Great Plains site- Lamont	Oklahoma - USA
US-Blo	Blodgett Forest	California -USA
US-IB2	Fermi National Accelerator Laboratory- Batavia	Illinois - USA
US-MMS	Morgan Monroe State Forest	Indiana - USA
US-SRM	Santa Rita Mesquite	Arizona - USA
US-Ton	Tonzi Ranch	California - USA
US-UMd	UMBS Disturbance	Michigan - USA
US-Var	Vaira Ranch- Ione	California - USA
US-Whs	Walnut Gulch Lucky Hills Shrub	Arizona - USA
US-Wkg	Walnut Gulch Kendall Grasslands	Arizona - USA
ZA-Kru	Skukuza- Kruger National Park	Sudafrica

Tabella 2: Codici indentificativi dei siti utilizzati nello studi.

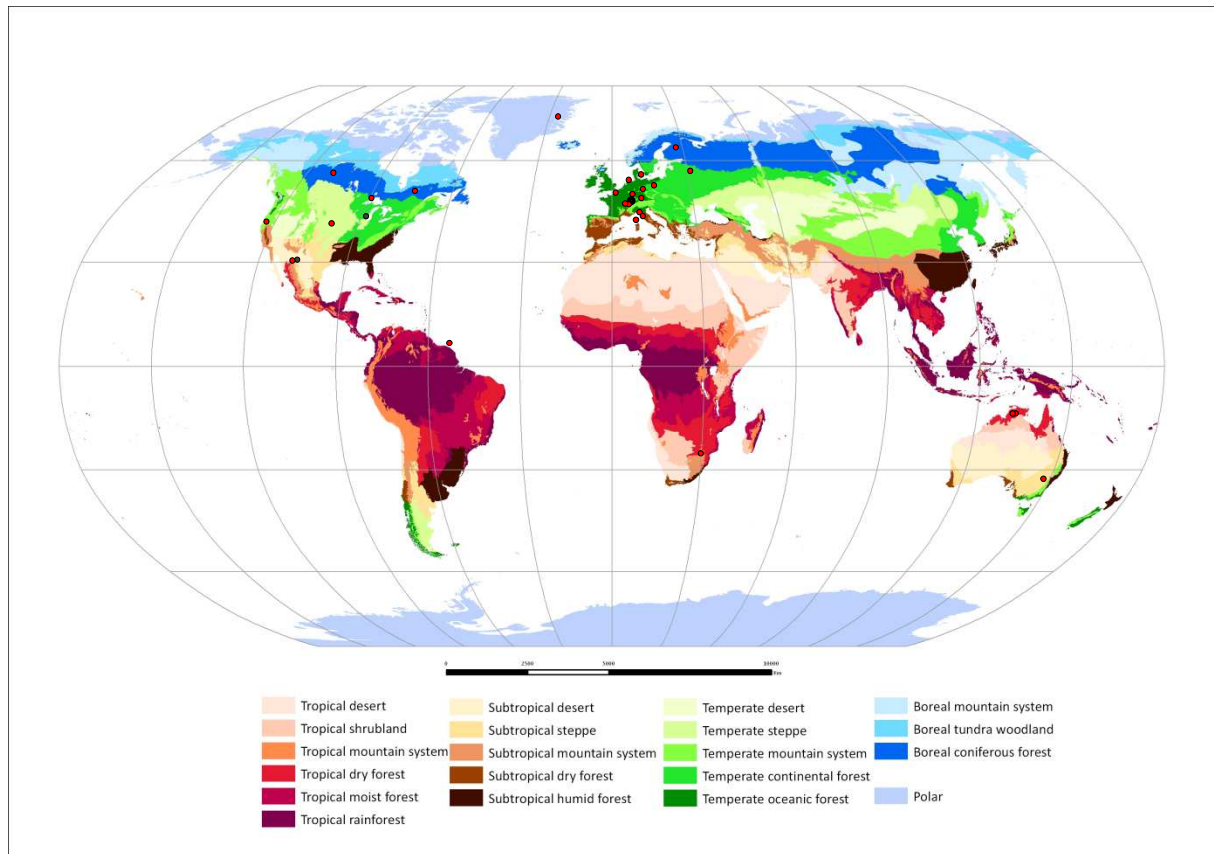


Figura 2: Localizzazione dei siti eddy covariance utilizzati nello studio.

2.1.1 Caratteristiche dei Siti

Per rendere possibile un confronto tra i risultati ottenuti con le diverse tipologie di siti, sono state prese in considerazione le seguenti caratteristiche: latitudine (Lat), altitudine (Alt), temperatura media annua (MAT), precipitazioni medie annue (MAP), leaf area index (LAI) e la tipologia vegetazionale (IGBP) (Figura 3).

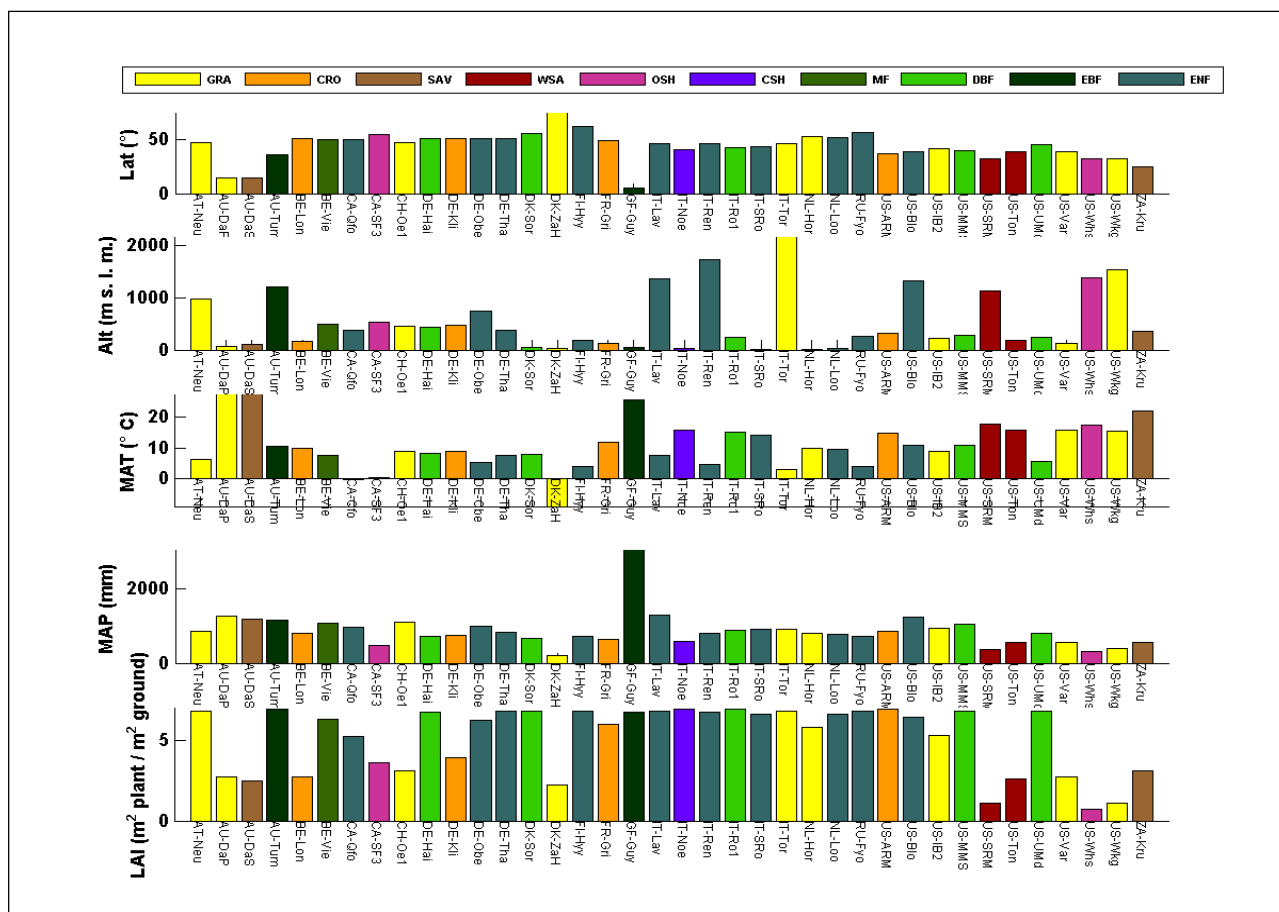


Figura 3: Caratteristiche dei siti distinti per IGBP: Latitudine (Lat), Altitudine (Alt), Temperatura media annua (MAT), Precipitazioni medie annue (MAP), Leaf Area Index (LAI).

Per ottenere un'ulteriore informazione climatica si è calcolato l'indice di aridità (Ia) per ogni sito, utilizzando la formula di De Martonne (1923) (Tabella 3):

$$Ia = \frac{P}{T + 10}$$

dove:

P : precipitazione media annua (mm)

T : temperatura media annua (°C)

CLIMA	Ia
Periumido	➤ 60
Umido	60-30
Subumido	30-20
Semiarido (di tipo mediterraneo)	20-15
Arido (steppe)	15-5
Arido estremo (deserto)	0-5

Tabella 3: Indice di Aridità di De Martonne (1923).

2.2. Selezione della tipologia e parametrizzazione dei modelli

2.2.1. ANN

Le ANN possono avere diverse architetture. Possono essere composte da diversi strati interni (*hidden layers*) e tali strati possono contenere un differente numero di nodi (neuroni). Un maggior numero di nodi comporta un aumento dei gradi di libertà del modello e una maggiore difficoltà nella parametrizzazione dello stesso. Per individuare quale architettura fosse più idonea al caso di studio, sono state testate diverse strutture delle reti.

In questo modo è stato possibile verificare se vi fosse una architettura standard da impiegare per le metodologie. Tale verifica è stata effettuata sulla base dell'errore totale dei modelli nella stima della GPP a partire dai driver selezionati.

La selezione dei parametri delle reti neurali su cui applicare i metodi per la selezione dei driver è stata effettuata tramite due fasi:

- **FASE 1 (analisi dell'architettura)**

Al fine di valutare la possibilità di selezionare una architettura comune delle ANN per tutti i siti sono stati fatti degli addestramenti utilizzando la tecnica del *n-folds*: il dataset corrispondente ad un anno di dati per volta è stato messo da parte per la validazione finale della rete e tutti gli altri usati per la parametrizzazione. In questo modo sono stati effettuati tanti addestramenti quanti gli anni a disposizione.

L'addestramento delle reti è stato effettuato dividendo il rimanente set di dati con un rapporto pari a 60:20:20 (rispettivamente *training set*, *test set* e *validation set*, dove il primo è usato per settare i parametri della rete, il secondo per minimizzare l'*overfitting* ed il terzo

per verificare la capacità di generalizzare). Questa suddivisione è stata ripetuta per dieci volte (10 estrazioni) in maniera random per ogni ciclo.

Sono state scelte sei architetture differenti (6-5-1, 6-7-1, 6-9-1, 6-11-1, 6-5-3-1, 6-7-5-1, dove il primo numero sono le variabili in input, l'ultimo numero la variabile target e i rimanenti numeri corrispondono ai neuroni nei diversi strati intermedi).

Per ogni architettura sono state effettuate 5 inizializzazioni in cui i pesi iniziali sono stati settati in modo random (in modo da ridurre le possibilità di terminare l'addestramento in minimi locali). Ogni rete è stata infine validata con l'anno del dataset messo da parte.

Alla fine di questa prima fase, per ogni ciclo è stata selezionata l'architettura con la migliore performance sulla base dell'errore sulla predizione (*rmse- root mean square error*).

L'obiettivo di questa fase era di individuare una architettura base da poter usare come standard nelle successive fasi. In effetti i risultati ottenuti non hanno permesso di individuare una singola architettura. Per questo motivo si è deciso di applicare le successive fasi a tutte e sei le architetture identificate e mediane i risultati.

- **FASE 2 (addestramento dei modelli e stima degli errori)**

Sull'intero set di dati disponibili e con lo stesso rapporto di suddivisione per i 3 *subset* utilizzato nella fase 1, sono state effettuate 20 estrazioni *random* per la creazione de *training set*, del *test set* e del *validation test*.

L'addestramento delle ANN è stato effettuato usando l'algoritmo *backpropagation* e per minimizzare il rischio di minimi locali sono state effettuate 5 inizializzazioni per ogni addestramento.

Il calcolo dell'errore del modello è stato effettuato utilizzando il *validation set*.

Per ogni sito, alla fine della fase 2 sono state create e addestrate 120 reti (20 estrazioni * 6 architetture) che sono state utilizzate per il calcolo dell'errore del modello (media delle 120 reti) e per la stima dell'importanza dei differenti drivers.

2.2.2. Random Forest

Per quanto riguarda la parametrizzazione delle Random Forest, i parametri necessari sono diversi da quelli richiesti dalle reti neurali. È stato scelto il metodo della regressione. Sono stati costruiti 100 alberi di decisione binaria per la costruzione della foresta. Il numero minimo di osservazioni per foglia è stato definito pari a 1.

È stato attivato il parametro *OOBvarimp* che memorizza le stime dell'importanza delle variabili calcolate sull'*out-of-bag* in previsione dell'applicazione del metodo per l'importanza dei drivers.

Il numero di variabili da selezionare random per ogni decisione binaria è lasciato di default, ovvero, per la regressione, un terzo del numero delle variabili a disposizione.

Sulla foresta viene poi applicata la funzione *oobPredict* con cui vengono calcolate le risposte previste dal modello utilizzando il *bagging* per l'addestramento, con le osservazioni OOB. L'output consiste in una previsione per ogni addestramento.

2.3. Test e selezione dei metodi per la stima dell'importanza dei driver

Per validare i diversi metodi di investigazione dell'importanza dei driver e verificarne l'applicabilità per il successivo utilizzo con i dati eddy-covariance, sono stati utilizzati dataset artificiali di diversa lunghezza in cui il ruolo delle variabili è definito a priori secondo questa funzione (Friedman's function):

$$y = 10 \times \sin(\pi \times X_1 \times X_2) + 20 \times (X_3 - 0.5)^2 + 10 \times X_4 + 5 \times X_5 + \text{rumore}$$

I dataset artificiali sono originati a partire da 10 variabili, delle quali 5 con un ordine di importanza noto (X_4 è la variabile più importante, seguita da X_1 e X_2 che hanno stessa importanza, e da X_3 , X_5 , anch'esse con uguale importanza), e le altre 5 senza nessun ruolo nella definizione dell'output (rumore).

I metodi sono stati applicati a dataset artificiali di lunghezza diversa (100, 1000, 5000 e 10000 dati) per simulare diverse possibili scale di aggregazione temporale di dati eddy covariance e verificare l'applicabilità dei metodi anche in caso di un numero di dati ridotto.

Inoltre è stato creato un dataset da 1000 valori con una variabile ripetuta per osservare come le metodologie rispondessero nel caso di variabili altamente correlate, come accade spesso nel caso delle variabili meteo che seguono un andamento stagionale. La variabile duplicata è la X_4 , che è quella a cui viene attribuita la maggiore importanza all'interno del dataset artificiale.

Per simulare il rumore tipico dei dati eddy covariance e testare la robustezza dei metodi anche in presenza di dati rumorosi, è stato aggiunto del rumore sia alla variabile in output che alle variabili in input nel dataset da 1000 valori.

L'aggiunta del rumore alla variabile output è avvenuta con la seguente formula:

$$y_p = y + \mathbf{1} \times p$$

dove:

y: output non perturbato

p: è un vettore di valori random, con media 0 e deviazione standard 1

Per perturbare gli input è stato aggiunto ad ogni variabile da perturbare la stessa variabile moltiplicata per un fattore di perturbazione (α), che è rappresentato da un valore che varia da -0.5 a 0.5. Questo fattore aggiunge o sottrae alla variabile da perturbare una percentuale random del suo valore che oscilla tra 0 ed il 50%. La formula applicata è la seguente:

$$X_p = X + \alpha \times X$$

dove:

X_p : *variabile in input perturbata*

X : *variabile in input non perturbata*

α : *fattore di perturbazione che varia tra -0.5 e 0.5*

2.3.1 Metodi testati nello studio:

Con l'ausilio del dataset artificiale è stato possibile testare i diversi metodi per individuare una classificazione dei driver.

Le parametrizzazioni delle ANN e delle RF hanno permesso di individuare i modelli base su cui applicare le metodologie. L'errore, calcolato sull'*OOB* per le Random Forest e su il *validation set* per le reti, è stato il valore di riferimento per definire gli addestramenti corretti di ogni modello e per evitare di andare incontro ad overfitting.

Ogni metodo ha bisogno di un modello addestrato con variabili in input e la variabile target su cui le funzioni dei modelli calcolano l'output. Le performance relative all'applicazione dei metodi terranno conto degli R^2 di fitting per verificare la capacità dei modelli di fittare i dati.

In questo lavoro sono stati testati 6 metodi per ricavare l'importanza dei Driver, 5 di essi si basano sulle Reti Neurali:

1. Connection Weight (WM – *Weight Method*)
2. Stepwise - Forward Method (FM – *Forward Stepwise Method*)

3. Stepwise - Backward Method (BM – *Backward Stepwise Method*)

4. Perturbation (PM – *Perturbation Method*)

5. Destroy order (DM – *Destruction Order Method*)

L'ultimo metodo si basa sulle Random Forest (RM – *Random Forest Method*).

Tutti i metodi sono stati applicati in ambiente MATLAB.

I parametri di *performance* di ogni modello (rete o albero) necessari per l'applicazione delle diverse metodologie, sono: R^2 (coefficiente di determinazione), **rmse** (errore quadratico medio) e **meff** (*model efficiency*) calcolati come segue:

$$R^2 = \left\{ \frac{\frac{\sum_{j=1}^n [(y_j - \mu_y) \times (x_j - \mu_x)]}{n_y - 1}}{(\sigma_y \times \sigma_x)} \right\}^2$$

$$rmse = \sqrt{\left\{ \mu [(x_j - y_j)^2] \right\}}$$

$$meff = 1 - \frac{\left\{ \mu [(x_j - y_j)^2] \right\}}{\mu (y_j - \mu_y^2)}$$

dove:

j = dato j_{esimo} del dataset

y = dataset reale

x = dataset predetto

σ = deviazione standard

μ = media

L'uso del dataset artificiale ha permesso di selezionare 3 metodi che si sono rivelati i più robusti (FM, DM e RM) e che sono poi stati applicati alle reti e alle RF addestrate sui siti eddy covariance. Maggiori dettagli riguardo questa fase sono forniti nei risultati.

2.4.SSA – modalità di applicazione ed uso degli output

L' applicazione della Singular Spectrum Analysis (SSA) è stata effettuata in ambiente R tramite la funzione *filterTseriesSSA* (Buttlar, 2015). Questa funzione scompone (o filtra) una serie temporale in un set di ortogonali (ad esempio additive) componenti con varianza su scale temporali distinte e differenti. Tale funzione si basa sul pacchetto *Rssa* di Korobeynikov, Shlemov, Usevich, Golyandina, & Korobeynikov (2015).

La funzione richiede l'inserimento di differenti parametri:

borders.wl, M, harmonics, repeat.extr

dove:

series: vettore numerico (serie in input senza gaps)

borders.wl: lista di vettori numerici (limiti delle differenti bande di periodicità da estrarre). Nel caso dell'estrazione di una singola banda, questo vettore sarà composto da due valori (il limite più basso e quello più alto, mentre nel caso dell'estrazione di più componenti (come in questo studio) il limite più alto di ogni banda è automaticamente il limite inferiore della banda successiva, quindi nel caso dell'estrazione di due sottosegnali, il vettore di input sarà composto da tre valori.

M: lunghezza della finestra o dimensione di embedding (vettore numerico). Per un'ottimale investigazione delle componenti spettrali, la lunghezza della finestra dovrebbe essere un multiplo del periodo di oscillazione che si desidera estrarre.

harmonics: numero di armoniche da includere in ogni componente.

repeat.extr: quanto spesso ripetere l'estrazione. Specialmente per il trend potrebbe essere opportuno ripetere lo step di filtraggio. In molti casi questa procedura aiuta a escludere le componenti ad alta frequenza dal trend.

Lo scopo era quello di estrarre dal segnale della GPP e dei driver la stagionalità, il trend e di conseguenza i residui.

Per individuare quali parametri fossero i più idonei a scomporre i segnali di GPP e driver, sono stati effettuati diversi test su un sito eddy covariance. Il sito selezionato è stato DE-Tha (ENF vicino a Tharant in Germania), caratterizzato da una serie temporale di oltre 16 anni (1997-2014) e già sottoposto alla metodologia SSA da Mahecha (2007).

Per confrontare le componenti risultanti con il lavoro di Mahecha, si è reso necessario inizialmente scomporre la NEE (Net Ecosystem Exchange - variabile utilizzata nello studio del 2007). I risultati

della scomposizione sono stati analizzati visivamente e paragonati alle componenti ottenute da Mahecha per gli anni 1997-2005.

Una volta individuati i parametri più idonei a estrarre le componenti stagionalità e trend, tali parametri sono stati applicati alla GPP e ai driver per il sito DE-Tha. Si sono rese necessarie alcune modifiche nel settaggio dei per ottenere dei segnali visibilmente buoni.

L'algoritmo è stato applicato in modo scalare: per prima cosa viene estratta la stagionalità e sottratta dalla serie originale. Successivamente questo processo viene ripetuto per estrarre il trend che a sua volta viene rimosso dalla serie lasciando infine il segnale residuo. Ciò consente di separare in modo più chiaro il segnale.

Ogni parametro della funzione della SSA necessita di un vettore composto da tanti elementi quanti sono gli steps desiderati (in questo caso due). Per ogni passaggio viene utilizzato il corrispondente elemento del vettore. Per *borders.wl* è necessario inserire i limiti delle bande da estrarre in ogni step (340,380; 356*4, ∞). I due valori relativi alla lunghezza della finestra di scorrimento (M) sono 365*5 e 365*3. Il numero di armoniche da includere in ogni componente (*harmonics*) sono rispettivamente tre per la stagionalità e zero per il trend. Le estrazioni (*repeat.extr*) vengono ripetute due volte per la stagionalità e dieci volte per il trend.

Una volta applicata la funzione al sito DE-Tha, si è proceduto con gli altri 37 siti utilizzando gli stessi parametri e sono state così isolate la componente stagionale, la componente a frequenza più ampia (assimilabile al trend) e i residui.

Per verificare che i residui fossero effettivamente privi di stagionalità è stata effettuata in ambiente R un'analisi dei valori residuali tramite la funzione *acf* che calcola le stime dell'autocovarianza in base alle deviazioni standard della serie di dati.

Una volta ottenute le tre componenti (stagionale, trend e residuo) sia della GPP che dei drivers i modelli (ANN e RF) sono stati addestrati e poi analizzati per la stima dell'importanza dei drivers usando tre diversi schemi e combinazioni:

1. GPP originale stimata a partire dalle tre componenti dei drivers:

$$GPP = Driver_s + Driver_t + Driver_r$$

2. Trend della GPP stimato a partire dai trend dei drivers

$$GPP_t = Driver_t$$

3. Residui della GPP stimati a partire dai residui dei drivers

$$\text{GPP}_r = \text{Driver}_r$$

dove:

_s: stagionalità della serie

_t: andamento di lungo periodo della serie

_r: residui della serie

La stima della stagionalità della GPP non è stata analizzata perché il forte segnale presente veniva facilmente spiegato da tutte le variabili con ciclo simile quali radiazione e temperatura.

2.5. Analisi risultati e sviluppo di un indice sintetico

Ognuno dei tre metodi selezionati rilascia un rank di importanza delle variabili, ma ciascun rank è basato su parametri differenti in base alla metodologia che lo ha generato.

Il metodo FM, rilascia una classificazione dei driver basandosi su R^2 , tanto più alto sarà R^2 del modello, tanto maggiore sarà l'importanza associata dal metodo alla variabile con cui il modello viene creato.

Il metodo DM riporta una gerarchia delle variabili in input basandosi sull'*rmse*. Confrontando l'*rmse* della rete senza modificare le variabili in input con quello ottenuto dalla distruzione dell'ordine di una determinato driver, maggiore sarà l'errore della rete, tanto più importante sarà il driver disturbato.

Il metodo RM si basa sulla differenza di errore, considerando l'importanza delle variabili in base all'aumento dell'errore sul modello dopo la perturbazione dei driver in input. Più alta sarà la differenza tra gli errori prima e dopo la perturbazione, più importante sarà il driver perturbato.

Per superare le limitazioni imposte dai differenti approcci con cui viene stabilita l'importanza dei driver è necessario individuare un indice per confrontare e sintetizzare i risultati riportati da ogni metodo.

Tale indice deve normalizzare i valori riportati dai tre metodi nell'ambito di uno stesso sito rispetto alla performance migliore ottenibile, ovvero quella riferita al modello creato con tutti gli input, senza modifiche nell'ordine dei dati dei driver e senza perturbazione dei valori.

Una volta normalizzati i valori riportati dalle tre metodologie all'interno di uno stesso sito, vanno tenute presenti due considerazioni:

- 1 Il driver selezionato come più importante può avere significato diverso in differenti siti. Può infatti spiegare meglio le dinamiche all'interno di un sito rispetto ad un altro, pur essendo in entrambi selezionato come primo driver.
- 2 Il secondo driver individuato dai metodi potrebbe avere comunque una importanza molto vicina a quella del primo o al contrario molto lontana, non contribuendo realmente alla formazione dell'output.

Tali differenze devono essere quindi rappresentate all'interno dell'indice sintetico.

Infine per comparare i risultati anche tra i diversi siti, serve considerare quanto il modello base, di per sé, era in grado di riprodurre la GPP. La qualità del modello viene rappresentata dalla *meff*, che sarà quindi moltiplicata per l'indice sintetico, abbassando il valore dell'indice quando già il modello di base non funziona.

Il primo passaggio è stato quindi quello di normalizzare i risultati nell'ambito del singolo sito e del singolo metodo attraverso un indice (VRI - Variables Relative Index) che classifica le prime tre variabili selezionate da ogni metodo secondo le seguenti equazioni:

- **Metodo FM:**

L'indice normalizza l' R^2 della variabile selezionata come prima rispetto all' R^2 del modello addestrato con tutte le variabili. Per la seconda variabile selezionata, al fine di tenere in considerazione che la prima variabile è inclusa negli input utilizzati, l'indice VRI è calcolato come prodotto tra VRI della prima variabile ed il rapporto tra gli R^2 . Simile approccio è usato per la terza variabile, secondo le seguenti equazioni:

$$VRI_{j1} = \frac{R_{j1}^2}{R_{tot}^2}$$

$$VRI_{j2} = VRI_{j1} \times \frac{R_{j1+j2}^2}{R_{tot}^2}$$

$$VRI_{j3} = VRI_{j1} \times VRI_{j2} \times \frac{R_{j1+j2+j3}^2}{R_{tot}^2}$$

dove:

R_{j1}^2 : errore della rete addestrata con la prima variabile selezionata dal metodo

R_{j1+j2}^2 : errore della rete addestrata con le prime due variabili più importanti

$R_{j1+j2+j3}^2$: errore della rete addestrata con le prime tre variabili più importanti

R_{tot} : errore delle rete addestrata con tutte le variabili.

- **Metodo DM**

In questo caso l' $rmse$ è relativo alla variabile disturbata. Quindi considerando l' $rmse$ massimo, ovvero l'errore maggiore ottenuto dal cambiamento dell'ordine del driver più importante, l'indice VRI sarà ottenuto dal rapporto tra l'errore del modello con il driver disturbato e l'errore massimo.

$$VRI_{jn} = \frac{rmse_{jn}}{rmse_{max}}$$

dove:

$rmse_{jn}$: errore della rete dopo la distruzione dell'ordine della n variabile selezionata dal metodo (n varia da 1 a 3 e si riferisce alla prima, seconda o terza variabile selezionata dal metodo).

$rmse_{max}$: errore della rete massimo ottenuto dopo la distruzione dell'ordine della variabile più importante.

- **Metodo RM**

Per il metodo RM, la classificazione delle variabili è data dalla differenza tra l'errore del modello prima e dopo la perturbazione del driver. Anche in questo caso viene presa in considerazione la differenza massima ottenuta dalla perturbazione del driver più importante. L'indice VRI, sarà dato dal rapporto tra le due differenze di errore:

$$VRI_{jn} = \frac{\Delta_{jn}}{\Delta_{max}}$$

dove:

Δ_{jn} : importanza stimata del driver j sull'insieme dei dati OOB (n varia da 1 a 3 e si riferisce alla prima, seconda o terza variabile selezionata dal metodo).

Δ_{max} : differenza massima tra l'errore (MSE) stimato sull'insieme OOB permutando i diversi driver j .

Per rendere confrontabili i risultati anche tra i diversi siti è necessario tenere in considerazione le performance dei modelli per considerare anche l'incertezza dei valori riportati dai metodi.

Sono stati quindi mediati i VRI calcolati per i diversi metodi. Ogni VRI medio è stato moltiplicato per la performance media dei modelli ($meff$) creando in questo modo l'indice VAI (Variables Absolute Index):

$$VAI = VRI_{mi} \times meff_{mi}$$

dove:

VRI_{mi} : è il *VRI medio di un dato driver per il sito i*

$meff_{mi}$: è *l'efficienza media dei modelli applicati al sito i (diversa per le varie componenti indagate: trend e residui).*

3. Risultati

3.1. Performances di ANN e RF addestrate con variabili non decomposte

Nella prima fase per la creazione delle ANN (analisi dell'architettura), confrontando le performance, calcolate tramite dataset indipendenti, dei diversi modelli addestrati in base alla riduzione dell'errore sulla predizione (*rmse*- *root mean square error*), non sono state riscontrate differenze significative tra un tipo di architettura e un altro. La frequenza con cui ogni tipologia di rete veniva selezionata all'interno di uno stesso sito e tra i diversi siti, è apparsa omogenea per le sei diverse architetture testate.

Di conseguenza, nella fase 2 (addestramento dei modelli e stima degli errori), sono state utilizzate tutte e sei le architetture.

In questo modo sono state create 120 reti (20 * 6 architetture) per ogni sito.

Le performance medie rappresentate dalla media della *meff* delle 120 reti e dei 100 alberi delle Random Forest sono confrontate nella figura 4 e riportate nella tabella 4. Dalla figura 4, si può osservare come le due tipologie di modello (ANN e RF) riportino valori di performance altamente correlati. Infatti per nessun sito vi sono differenze particolari nelle *meff* dei modelli. La tabella 4 mostra valori di performance in media superiori a 0.6.

I modelli sono quindi in grado di simulare i flussi della GPP, senza andare incontro ad overfitting.

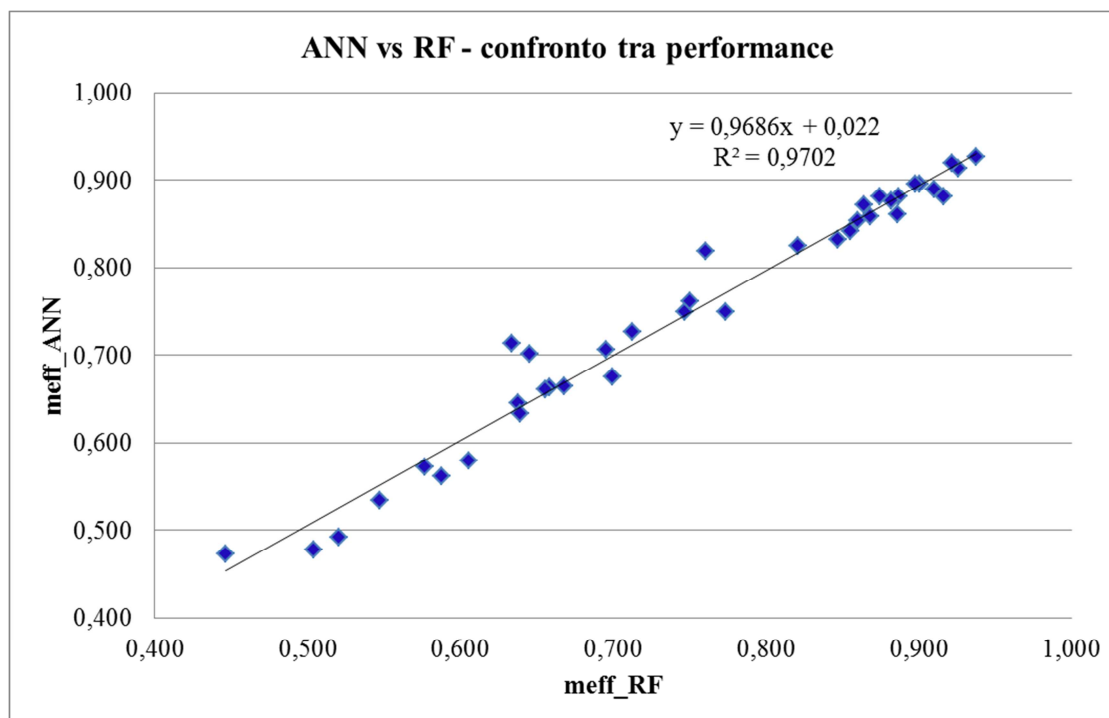


Figura 4: Scatter plot relativo alle *meff* medie dei modelli creati con ANN e RF con driver e GPP non decomposti.

SITE_ID	IGBP	ANN	RF
AT-Neu	GRA	0,747	0,750
AU-DaP	GRA	0,887	0,882
AU-DaS	SAV	0,658	0,664
AU-Tum	EBF	0,700	0,676
BE-Lon	CRO	0,588	0,561
BE-Vie	MF	0,860	0,854
CA-Qfo	ENF	0,926	0,913
CA-SF3	OSH	0,774	0,750
CH-Oe1	GRA	0,668	0,665
DE-Hai	DBF	0,869	0,859
DE-Kli	CRO	0,638	0,646
DE-Obe	ENF	0,882	0,876
DE-Tha	ENF	0,901	0,895
DK-Sor	DBF	0,938	0,927
DK-ZaH	GRA	0,446	0,473
FI-Hyy	ENF	0,922	0,920
FR-Gri	CRO	0,577	0,573
GF-Guy	EBF	0,521	0,492
IT-Lav	ENF	0,865	0,872
IT-Noe	CSH	0,634	0,713
IT-Ren	ENF	0,822	0,825
IT-Ro1	DBF	0,750	0,762
IT-SRo	ENF	0,761	0,819
IT-Tor	GRA	0,910	0,889
NL-Hor	GRA	0,917	0,881
NL-Loo	ENF	0,898	0,895
RU-Fyo	ENF	0,874	0,881
US-ARM	CRO	0,504	0,478
US-Blo	ENF	0,646	0,702
US-IB2	GRA	0,847	0,832
US-MMS	DBF	0,856	0,841
US-SRM	WSA	0,639	0,634
US-Ton	WSA	0,696	0,706
US-UMd	DBF	0,886	0,861
US-Var	GRA	0,713	0,727
US-Whs	OSH	0,548	0,534
US-Wkg	GRA	0,656	0,662
ZA-Kru	SAV	0,605	0,579

Tabella 4: meff medie per ogni sito per i modelli addestrati con driver e GPP non decomposti. In giallo i valori più bassi.

3.2. Selezione dei metodi per la stima dell'importanza dei drivers

L'uso del dataset artificiale con importanza delle variabili nota ha permesso di individuare i metodi più efficienti per la classificazione dei driver.

Nella tabella seguente (Tabella 5) sono riportati i valori di importanza attribuiti dai diversi metodi alle variabili presenti nel modello di Friedman per le diverse lunghezze di dataset (100, 1000, 5000 e 10000 dati).

Nella prima colonna sono riportati i rank corretti delle variabili (dalla più importante con rank pari a 1, alla meno importante con rank pari a 4), nella seconda le variabili corrispondenti ai rank corretti.

In verde sono evidenziate le variabili che contribuiscono alla formazione dell'output nel dataset artificiale (verde scuro la più importante –X4, verde acceso per X1 e X2 che hanno uguale importanza e verde chiaro per X3 e X5, anch'esse con uguale importanza) e in grigio le variabili che non hanno alcuna importanza (rumore).

Rank di importanza	Variabili	Metodi																							
		WM				PM				DM				FM				BM				RM			
		100	1000	5000	10000	100	1000	5000	10000	100	1000	5000	10000	100	1000	5000	10000	100	1000	5000	10000	100	1000	5000	10000
1	X4	X4	X2	X2	X1	X4	X2	X3	X3	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4
2	X1	X1	X1	X1	X2	X2	X1	X1	X1	X1	X1	X2	X1	X2	X1	X2	X2	X1	X1	X2	X1	X2	X1	X2	X1
2	X2	X2	X3	X3	X3	X1	X4	X2	X2	X2	X2	X1	X2	X1	X2	X1	X1	X2	X2	X1	X2	X1	X2	X1	X2
3	X3	X3	X4	X4	X5	X3	X4	X4	X5	X3	X3	X3	X5	X3	X3	X5	X3	X5	X3	X3	X5	X5	X5	X5	X5
3	X5	X5	X5	X5	X5	X3	X5	X5	X5	X3	X5	X5	X5	X6	X5	X5	X5	X3	X5	X5	X3	X3	X3	X3	X3
4	X6	X6	X7	X6	X9	X9	X7	X8	X7	X8	X6	X8	X7	X3	X10	X8	X9	X8	X6	X7	X6	X10	X10	X8	X8
4	X7	X8	X6	X8	X7	X10	X6	X9	X8	X6	X10	X9	X8	X7	X8	X9	X6	X10	X10	X6	X9	X7	X6	X6	X6
4	X8	X10	X10	X10	X10	X7	X10	X7	X9	X9	X7	X7	X10	X10	X6	X10	X7	X6	X9	X9	X10	X9	X7	X9	X7
4	X9	X9	X8	X9	X8	X8	X9	X6	X10	X10	X8	X6	X6	X8	X9	X6	X8	X7	X8	X8	X7	X10	X9	X8	X9
4	X10	X7	X9	X7	X6	X6	X8	X10	X6	X7	X9	X10	X9	X9	X7	X7	X10	X9	X7	X10	X8	X8	X8	X7	X10

Tabella 5: Confronto dei rank corretti con quelli riportati dai metodi per le diverse lunghezze di dataset (100, 1000, 5000 e 10000 dati). In verde sono segnate le variabili che contribuiscono alla formazione dell'output (verde scuro per la più importante, fino a verde più chiaro per X3 e X5), in grigio le altre. WM è il weight method, PM è il perturb method, DM è il destruction order method, FM è il forward stepwise method, BM è il backward stepwisw mwthod e RM è il random forest method.

I risultati mostrano come i metodi WM (weight method) e PM (perturbation method) nella maggior parte dei casi non riescano ad attribuire il rank corretto alle variabili e per questo sono stati esclusi dal resto delle analisi.

I metodi DM (destruction order method), FM (forward stepwise method), BM (backward stepwise method) e RM (random forest method) attribuiscono sempre valori di importanza alle variabili X uguali o prossimi a quelli corretti.

Nella Tabella 6 sono riportati i rank restituiti dai quattro metodi che risultano i più robusti per tutte le lunghezze del dataset (DM, FM, BM, RF). Sono state inserite due colonne che riportano rispettivamente i rank attribuiti dai quattro metodi alle variabili dopo l'aggiunta di:

- rumore sia nell'output che e nell'input al fine di testarne le robustezza al rumore, tipico delle serie eddy covariance (colonna **x**).
- variabile ripetuta (colonna **xx**) che corrisponde alla variabili X4, la più importante.

Rank di importanza	Variabili	Metodi											
		DM						FM					
		100	1000	5000	10000	x	xx	100	1000	5000	10000	x	xx
1	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4	X4
2	X1	X1	X1	X2	X1	X2	X1	X2	X1	X2	X2	X1	X1
2	X2	X2	X2	X1	X2	X1	X2	X1	X2	X1	X1	X2	X2
3	X3	X5	X3	X3	X3	X3	X3	X5	X3	X3	X5	X5	X3
3	X5	X3	X5	X5	X5	X5	X5	X6	X5	X5	X3	X3	X5
		BM						RM					
		100	1000	5000	10000	x	xx	100	1000	5000	10000	x	xx
1	X4	X4	X4	X4	X4	X4	X1	X4	X4	X4	X4	X4	X1
2	X1	X1	X1	X2	X1	X2	X2	X2	X1	X2	X1	X1	X4
2	X2	X2	X2	X1	X2	X1	X3	X1	X2	X1	X2	X2	X2
3	X3	X5	X3	X3	X3	X3	X5	X5	X5	X5	X5	X5	X3
3	X5	X3	X5	X5	X5	X5	X4	X3	X3	X3	X3	X3	X5

Tabella 6: Confronto dei rank corretti con quelli riportati dai metodi più robusti (DM - destruction order method, FM – forward stepwise method, BM – backward stepwise method e RM – random forest method) per l'aggiunta di rumore in input e in output (x) e con l'inserimento della variabile ripetuta X4(xx). In verde sono segnate le variabili che contribuiscono alla formazione dell'output (verde scuro per la più importante, fino a verde più chiaro per X3 e X5), in grigio le altre.

Anche in questo caso, in verde sono evidenziati i rank attribuiti dai metodi in accordo con la giusta classificazione e in giallo i valori che si discostano da quelli attesi.

Tre metodi sono risultati più affidabili anche nei casi dell'aggiunta di rumore e di una variabile ridondante (Random Forest – RM, Destruction of order method – DM e Forward stepwise method – FM). Il metodo BM invece non riesce ad attribuire il giusto rank alle variabili nel caso dell'aggiunta

di una variabile ridondante in quanto al suo interno contiene la variabile gemella e quando viene rimossa la X4, non rileva una perdita di informazione all'interno del modello.

3.2.1 Effetto della lunghezza del dataset sulle performances dei metodi

Al fine di analizzare quanto i tre metodi selezionati per la stima dell'importanza delle variabili fossero influenzati dalla lunghezza del dataset, sono stati effettuati dei test usando lo stesso dataset artificiale ma di lunghezza differente e, nel caso del dataset da 1000 valori, anche con l'aggiunta di rumore e di una variabile ridondante. Per le ANN sono state anche confrontate differenti architetture per vedere se in questo caso ci fossero differenze nelle performances.

Per ogni dataset di lunghezza diversa sono state create 120 reti (ovvero sei architetture diverse, per ogni architettura sono state addestrate 20 reti, come descritto nel paragrafo 2.5) per i metodi DM e FM, e 100 alberi di decisione per il metodo RM. I risultati sulla classificazione dell'importanza delle variabili sono stati confrontati attraverso l'indice di Spearman come illustrato in Figura 4 (Random Forest - RM), Figura 5 (Destruction of Order - DM) e Figura 6 (Forward stepwise method - FM). Il coefficiente di correlazione di Spearman assume i valori tra -1 e $+1$ indicando nel segno e nel valore il tipo e la forza della correlazione. Il segno $+$ indica una correlazione direttamente proporzionale, il segno $-$ indica una correlazione inversamente proporzionale. Valori di rho vicini ad 1 indicano una correlazione perfetta, il valore $\rho = 0$ molto vicino a 0 indica una correlazione nulla.

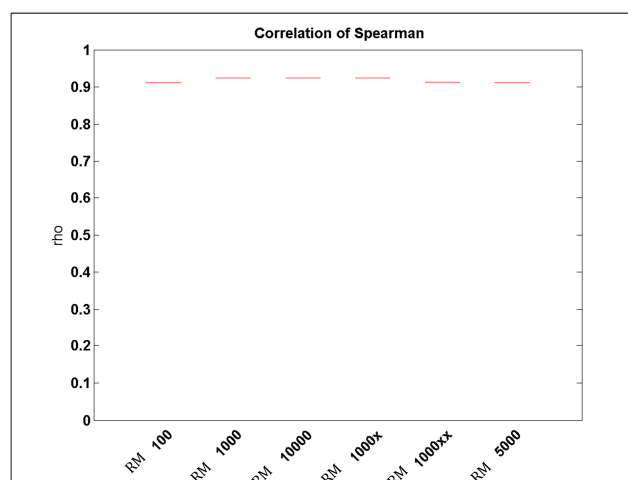


Figura 5: Indice di Spearman relativo alla correlazione tra i risultati ottenuti dall'applicazione del metodo RM per la selezione delle variabili per dataset di lunghezza diversa per il metodo Random Forest (RM). Ogni boxplot, che appare collassato in quanto rho risulta molto alto, contiene i valori riportati da 100 alberi di decisione.

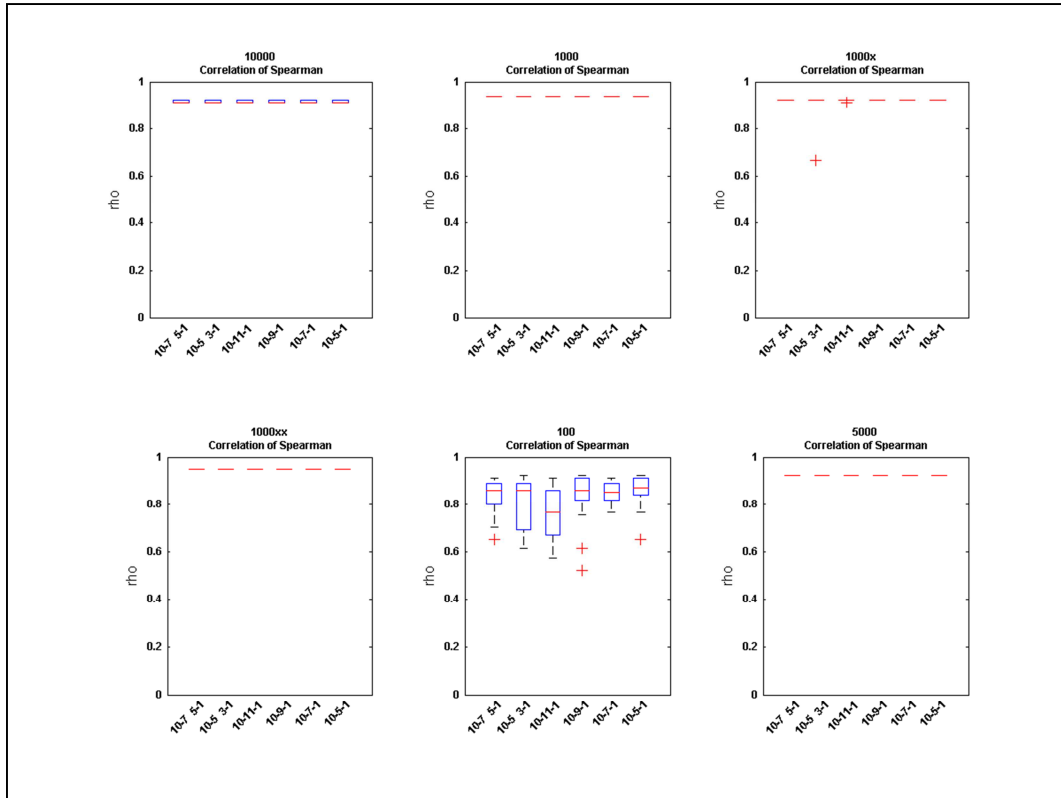


Figura 6: Indice di Spearman relativo alla correlazione nella selezione delle variabili del metodo DM. Ogni boxplot riporta i valori di 20 reti, confrontate in base all'architettura e per dataset di lunghezza diversa.

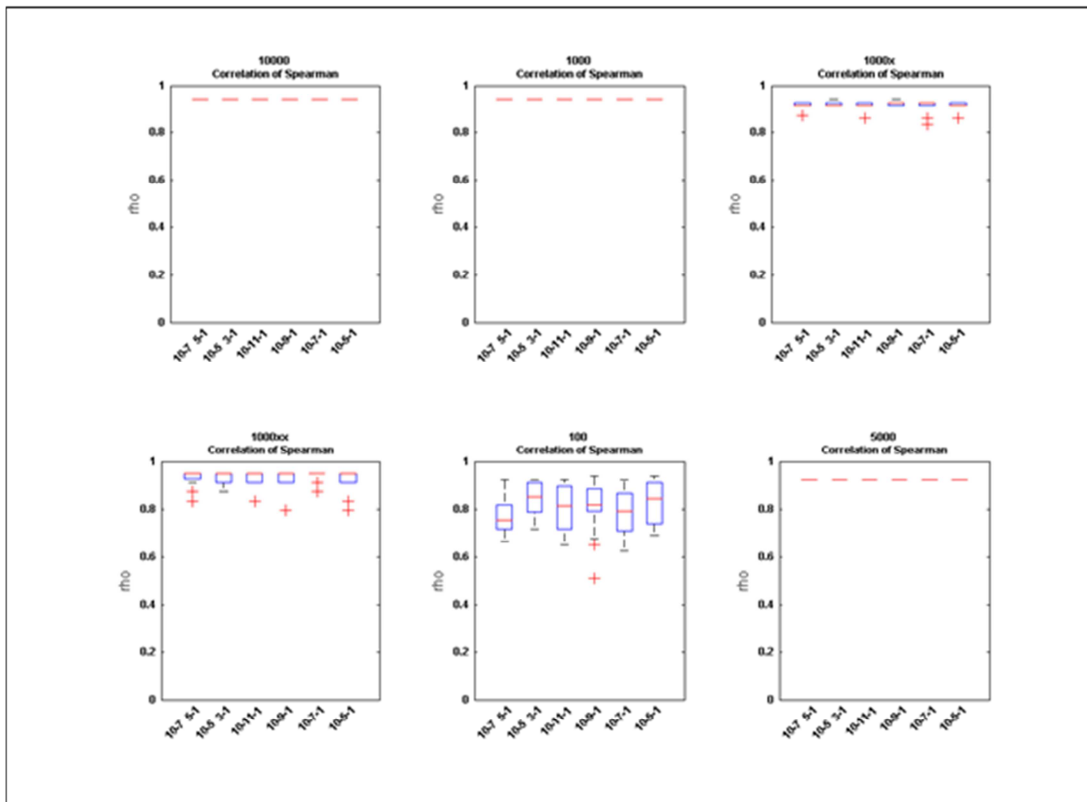


Figura 7: Indice di Spearman relativo alla correlazione nella selezione delle variabili del metodo FM. Ogni boxplot riporta i valori di 20 reti, confrontate in base all'architettura e per dataset di lunghezza diversa.

Il metodo delle Random Forest (Figura X) risulta robusto sia per quanto riguarda le diverse lunghezze di dataset, sia nel caso di aggiunta di rumore al dataset e sia con l'aggiunta di una variabile ripetuta.

I metodi basati sulle ANN (DM – Figura X, FM – Figura X) mostrano un indice di Spearman più basso nel caso del dataset corto (100 valori), mentre rispondono bene in tutte le altre situazioni, con qualche limitazione nel caso della variabile ridondante e l'uso del metodo FM. In ogni caso, l'architettura della rete non influenza il risultato a conferma di quanto già esposto in precedenza.

3.3.SSA – Segnali ottenuti dalla scomposizione

La metodologia SSA (Singular System Analysis) è stata applicata ai 38 siti. Per ogni sito si sono scomposti i segnali delle seguenti variabili: GPP, SW-IN, TA, TS, VPD, SWC e P.

Sono stati quindi ricavati 3 componenti: stagionalità (S), trend (T) e residui (R).

Come primo passaggio è stata estratta la stagionalità. Dal segnale destagionalizzato è stato poi isolato il trend, lasciando come ultima componente quella residuale.

Nelle figure 8 e 9 sono riportati degli esempi di decomposizione relativi al sito DE-Tha. La figura 8 mostra la decomposizione della GPP, mentre la figura 9 la decomposizione del driver TS.

Si può notare come la stagionalità abbia un andamento molto regolare, che si ripete di anno in anno, mentre i residui, che rappresentano la parte non spiegata da stagionalità e trend di lungo termine, sono maggiori durante la stagione vegetativa per quanto riguarda la GPP. Il trend mostra un andamento di lungo periodo che tende a crescere, in TS, per i diversi anni analizzati. Questa componente è sempre rappresentata da una bassa variabilità, quasi lineare.

Nella tabella 7 sono riportati i siti e le sette variabili climatiche (GPP e i sei drivers) ognuna con le sue 3 componenti (S - Stagionalità, T - Trend e R - Residui) derivate dalla scomposizione tramite SSA. Non sempre tutte le componenti sono presenti in tutti i siti. Il trend per molte variabili e per molti siti non viene individuato, così come la stagionalità in alcuni casi. La tabella 7 mostra dove non ci sono segnali (ovvero dove viene individuato solo un valore costante). Le celle nere individuano le variabili per cui non sono presenti o la stagionalità o il trend. I residui sono sempre presenti essendo il risultato della sottrazione di stagionalità e trend dalla serie originale.

Al termine dell'applicazione della metodologia di analisi, sono state individuate un totale di 667 componenti.

La componente di lungo periodo (T) per la GPP non è presente in 7 siti: CH-Oe1, DE-Hai, DE-Obe, FR-Gri, IT-Noe, NL-Loo, US-SRM. Per questi siti non è stato possibile quindi analizzare le risposte dei driver ecologici al trend della GPP.

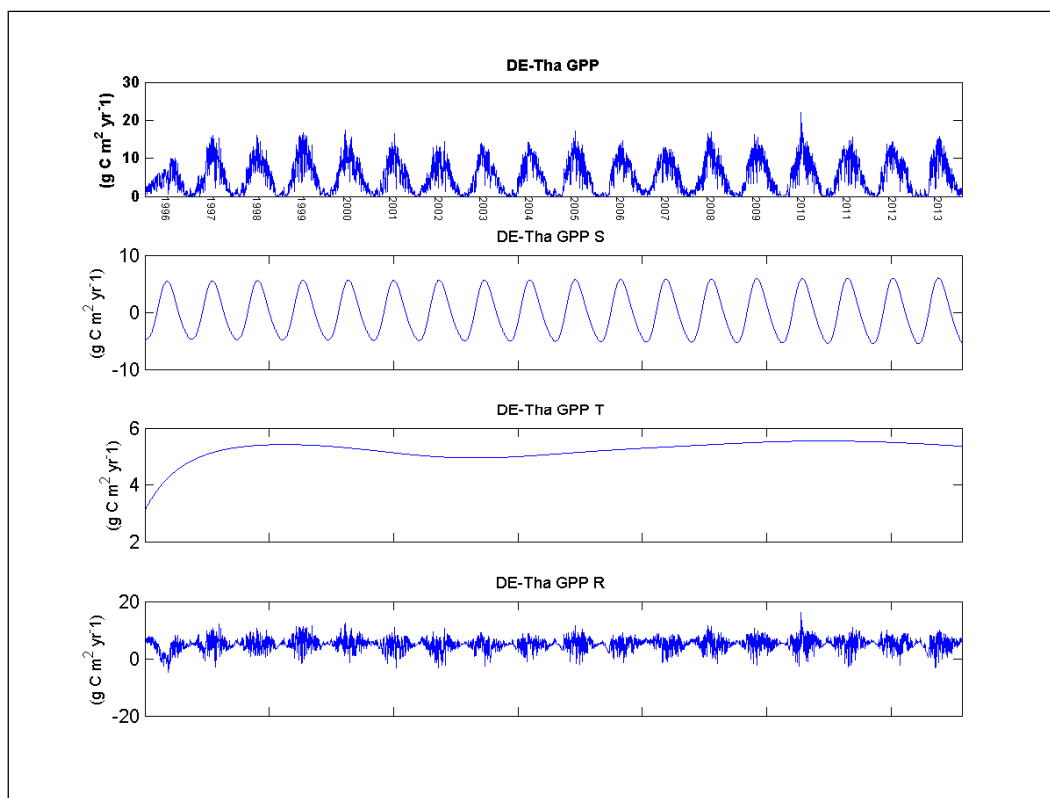


Figura 8: Scomposizione del segnale GPP nelle 3 componenti (S, T e R) per il sito DE-Tha.

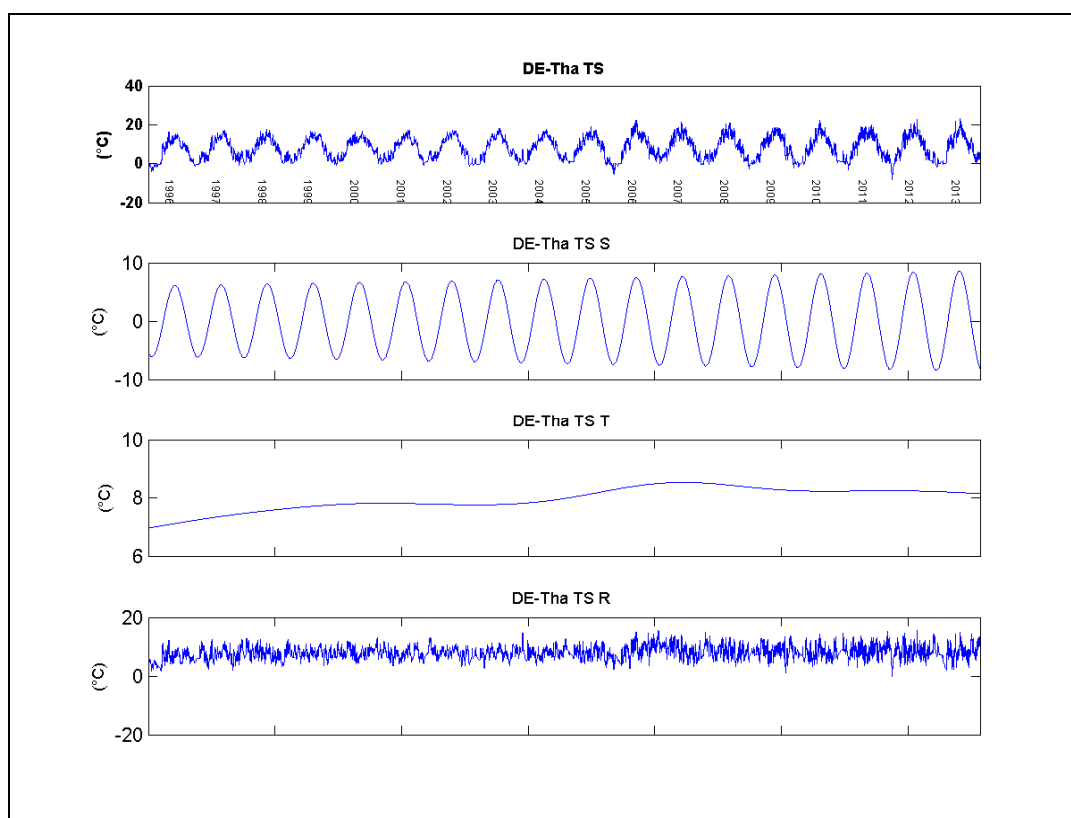


Figura 9: Scomposizione del segnale TS nelle 3 componenti (S, T e R) per il sito DE-Tha.

La variabile che per un maggiore numero di siti non presenta andamenti di lungo periodo è SW-IN. Il trend della radiazione è infatti presente solo in 15 siti su 38 (AU-DaS, AU-Tum, DK-Sor, DK-ZaH, FI-Hyy, GF-Guy, IT-Ren, IT-SRo, NL-Loo, RU-Fyo, US-ARM, US-Blo, US-Ton, US-Var, ZA-Kru).

Il trend di TA è stato isolato in 22 siti su 38. I siti in cui non è stato individuato alcun trend per TA sono: AU-DaP, AU-DaS, CA-Qfo, CH-Oe1, DE-Hai, DE-Obe, DE-Tha, DK-Zah, IT-Lav, IT-SRo, IT-Tor, US-Blo, US-MMS, US-SRM, US-Whs, US-Wkg.

Dodici siti non mostrano andamenti a bassa frequenza per quanto concerne le precipitazioni: AU-DaP, CA-Qfo, CH-Oe1, DE-Tha, DK-ZaH, FR-Gri, IT-Tor, NL-Hor, US-ARM, US-SRM, US-UMd, US-Whs.

Otto siti non presentano i trend per TS e VPD. Per quanto riguarda il driver TS, i siti che non hanno l'andamento di lungo periodo sono: AT-Neu, CA-SF3, CH-Oe1, De-Hai, IT-Tor, US-SRM, US-Whs, US-Wkg. Mentre i siti che non presentano trend per VPD sono: CA-Qfo, CH-Oe1, DE-Tha, IT-Ro1, NL-Hor, RU-Fyo, US-SRM, US-Wkg.

Il SWC è la variabile che per la maggior parte dei siti (34 su 38) mostra la componente di lungo periodo. I 4 siti in cui il SWC non ha trend sono: CA-SF3, FR-Gri, IT-Tor, NL-Hor.

SITE_ID	IGBP	GPP			SW-IN			TA			TS			VPD			SWC			P		
		S	T	R	S	T	R	S	T	R	S	T	R	S	T	R	S	T	R	S	T	R
AT-Neu	GRA	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
AU-DaP	GRA	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1
AU-DaS	SAV	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
AU-Tum	EBF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BE-Lon	CRO	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BE-Vie	MF	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CA-Qfo	ENF	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1
CA-SF3	OSH	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1
CH-Oe1	GRA	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	1
DE-Hai	DBF	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1
DE-Kli	CRO	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DE-Obe	ENF	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1	1
DE-Tha	ENF	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1
DK-Sor	DBF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
DK-ZaH	GRA	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1
FI-Hyy	ENF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
FR-Gri	CRO	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1
GF-Guy	EBF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IT-Lav	ENF	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1
IT-Noe	CSH	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IT-Ren	ENF	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
IT-Ro1	DBF	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
IT-SRo	ENF	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
IT-Tor	GRA	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	1	0	0	1
NL-Hor	GRA	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	1
NL-Loo	ENF	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
RU-Fyo	ENF	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
US-ARM	CRO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1
US-Blo	ENF	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
US-IB2	GRA	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
US-MMS	DBF	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
US-SRM	WSA	1	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	1
US-Ton	WSA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
US-UMd	DBF	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1
US-Var	GRA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
US-Whs	OSH	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1
US-Wkg	GRA	1	1	1	1	0	1	1	0	1	1	0	1	1	0	1	0	1	1	0	1	1
ZA-Kru	SAV	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TOT		38	31	38	37	15	38	38	22	38	38	30	38	38	30	38	26	34	38	28	26	38

Tabella 7: Componenti S (stagionalità), T (trend) e R (residui) ricavate tramite SSA. Le celle grigie rappresentano i segnali mancanti.

Per alcuni siti, SWC e P non mostrano andamenti stagionali. I siti che non hanno stagionalità nel SWC sono 12: DE-Obe, DK-ZaH, FI-Hyy, IT-Lav, IT-Ren, IT-Ro1, IT-Tor, NL-Hor, RU-Fyo, US-ARM, US-IB2 e US-Vkg. Sono invece 10 i siti che non presentano stagionalità nelle precipitazioni: CH-Oe1, DE-Hai, DE-Obe, De-Tha, DK-Sor, FR-Gri, IT-Tor, NL-Loo, US-UMd, US-Wkg.

Per verificare che all'interno dei residui non vi fosse presenza di stagionalità, i dati residuali sono stati sottoposti all'analisi dell'autocovarianza (*acf*). I risultati mostrano, per il sito DE-Tha un'indipendenza dei residui (figura 10).

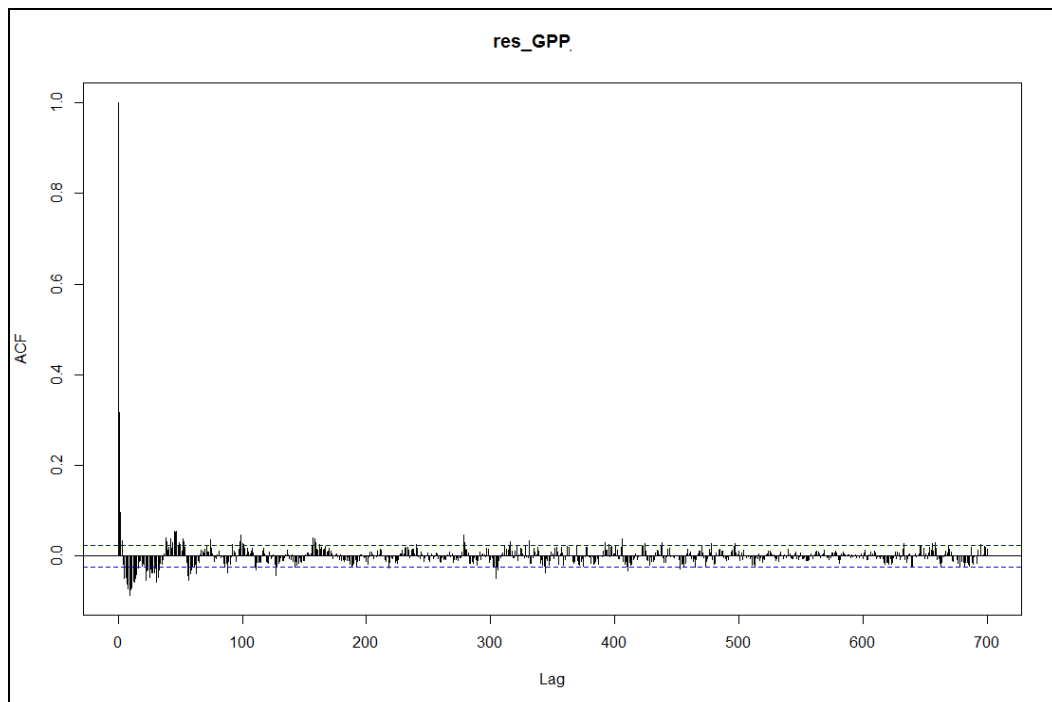


Figura 10: Analisi dei residui tramite la funzione ACF per la variabile GPP del sito DE-Tha. Le due linee tratteggiate indicano i valori limite. I dati che non superano i limiti non mostrano autocorrelazione e quindi presenza di stagionalità.

Per verificare se vi fosse qualche segnale residuale relativo a una determinata variabile che si discostasse dall'andamento medio dei residui di altri siti, sono state analizzate le ampiezze delle varie oscillazioni della serie residuale tramite il confronto tra le deviazioni standard (*Std*). Sono così stati individuati eventuali scostamenti dal valore medio dei residui delle variabili per i diversi siti (Tabella 8 e figura 11). Il risultato mostra come per alcuni siti, il segnale residuale di alcune variabili sia più ampio con una *Std* superiore rispetto agli altri siti (GPP per FR-Gri, , SWC per RU-Fyo, P per GF-Guy e US-Blo), ma comunque non sono presenti valori eccessivamente differenti.

I valori in giallo evidenziano come in alcuni casi la *Std* dei residui sia significativamente inferiore alla media (GPP per CA-SF3, DK-ZaH e US-Whs, TA per GF-Guy e AU-DaS, TS per GF-Guy, VPD per DK-ZaH, SWC per AU-DaS, US-SRM e US-Whs, P per CA-SF3, DK-ZaH, FR-

Gri, US-SRM e US-Whs), ma anche in questo caso, gli scostamenti non risultano mai di grande entità.

Siti	IGBP	GPP	SW-IN	TA	TS	VPD	SWC	P
AU-DaS	SAV	1,40	45,86	1,67	2,25	3,93	0,02	3,04
CA-SF3	OSH	0,60	51,85	5,44	2,07	2,52	3,58	0,67
DK-ZaH	GRA	0,31	39,04	4,97	3,92	0,96	5,03	0,54
FR-Gri	CRO	3,86	53,18	3,33	3,20	2,30	3,26	0,99
GF-Guy	EBF	2,07	48,02	0,71	0,31	1,57	3,01	4,59
RU-Fyo	ENF	1,84	52,44	4,65	1,32	2,47	16,88	1,12
US-Blo	ENF	1,58	54,99	3,85	3,37	3,54	5,15	6,71
US-SRM	WSA	0,75	45,23	3,42	2,74	5,87	1,87	0,92
US-Whs	OSH	0,52	43,49	3,18	2,84	6,06	2,15	1,02
Std media		1,43	48,23	3,47	2,45	3,25	4,55	2,18

Tabella 8: Siti con deviazioni Standard (Std) dei residui che si discostano dalla media. In giallo i valori che sono inferiori alla metà della deviazione standard media, mentre in rosso quelli che superano del doppio il valore medio della Std.

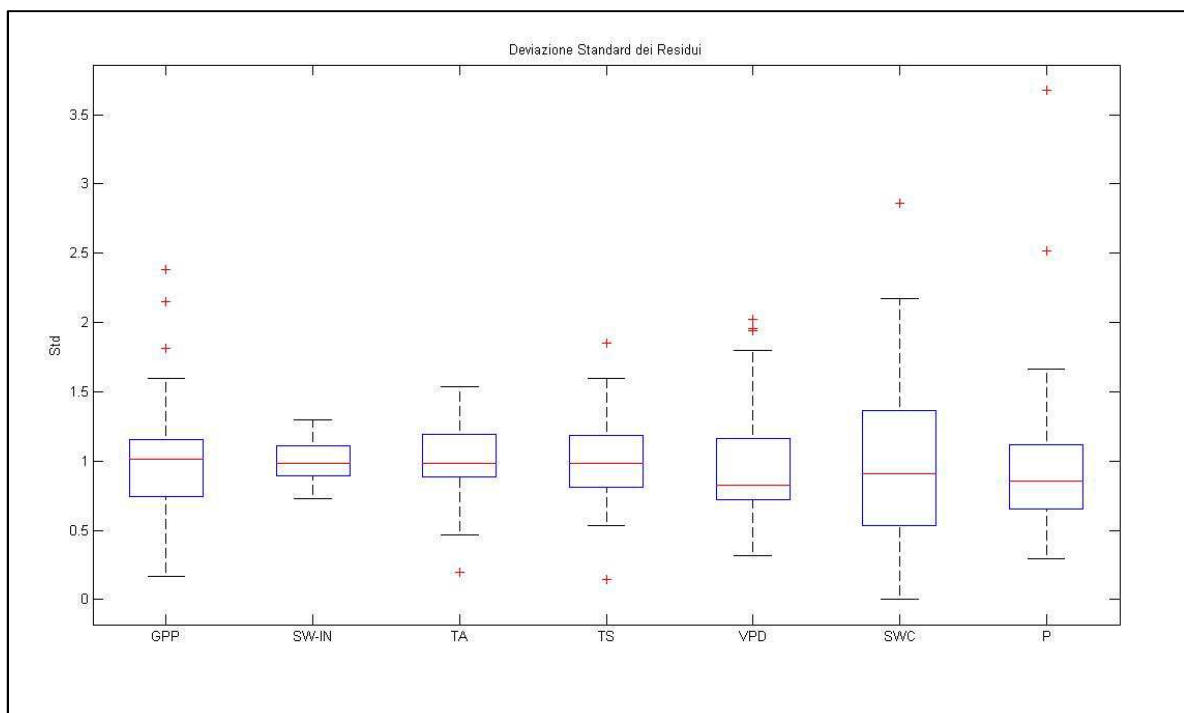


Figura 11: Deviazioni standard (Std) divise per la media della serie residuale su cui sono state calcolate. All'interno di ogni boxplor relativo a una variabile sono presenti le Std di tutti i siti.

3.4. Performance dei modelli ANN e RF addestrati usando variabili decomposte

Le componenti ricavate dall'SSA, sia sui driver che sulla GPP, sono state utilizzate per addestrare diversi modelli.

In Figura 12 sono riportate e rappresentate le performance (R^2) di fitting dei diversi modelli addestrati. I modelli ANN e RF si riferiscono alle reti e agli alberi addestrati senza scomporre le variabili di input e di output; i modelli ANN_all e RF_all sono i modelli addestrati con tutte e tre le componenti dei driver (S + T + R) in input e la GPP non decomposta come target. I modelli ANN_T e RF_T sono i modelli addestrati con la componente trend dei drivers e la componente trend della GPP; mentre i modelli ANN_R e RF_R sono addestrati con le componenti residuali dei drivers e la componente residuale della GPP.

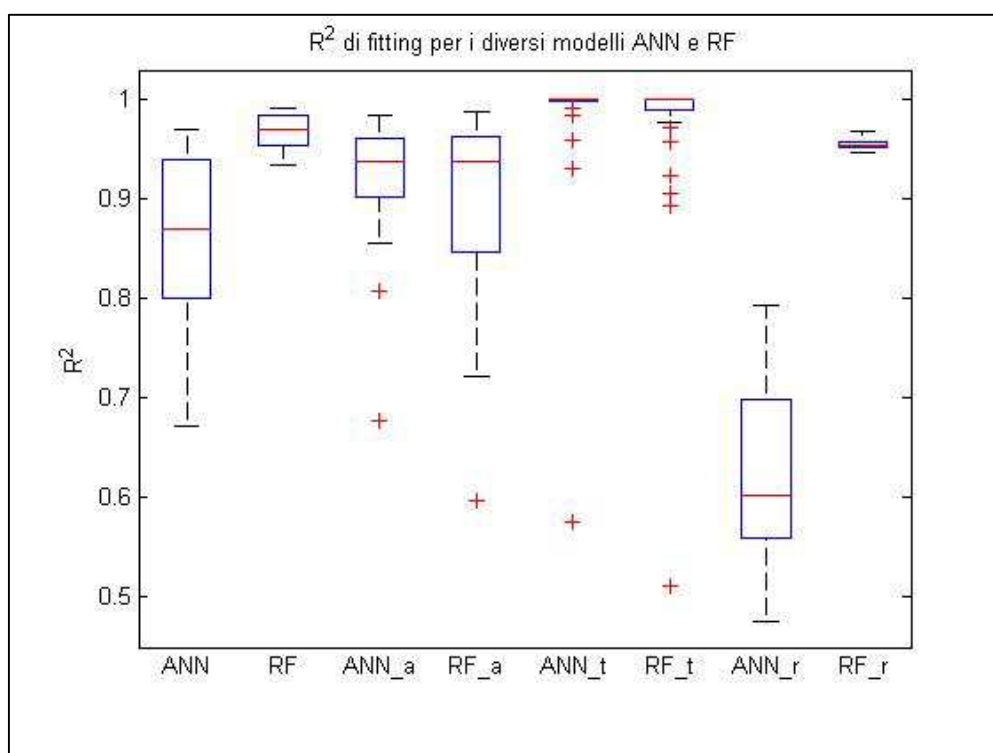


Figura 12: Confronto tra le performance di fitting dei diversi modelli (media delle 120 reti per le ANN e dei 100 alberi per le RF di tutti i siti) dove ANN e RF sono gli R^2 relativi ai modelli addestrati con le variabili originali, ANN_all e RF_all si riferiscono ai modelli creati con tutte le componenti delle variabili (S+T+R) in input e la GPP non decomposta in output. ANN_T e RF_T sono gli R^2 dei modelli addestrati con le componenti di lungo periodo delle variabili in input e il trend della GPP in output e ANN_R e RF_R quelli creati con la componente residuale delle variabili in input e i residui della GPP in output.

SITE_ID	ANN	RF	ANN_a	RF_a	ANN_t	RF_t	ANN_r	RF_r
AT-Neu	0,865	0,964	0,911	0,892	0,998	0,995	0,480	0,946
AU-DaP	0,942	0,985	0,958	0,942	0,930	0,922	0,685	0,961
AU-DaS	0,812	0,961	0,869	0,979	1,000	1,000	0,574	0,957
AU-Tum	0,837	0,953	0,883	0,957	1,000	1,000	0,730	0,957
BE-Lon	0,767	0,954	0,968	0,962	1,000	1,000	0,507	0,960
BE-Vie	0,927	0,974	0,919	0,928	1,000	1,000	0,720	0,953
CA-Qfo	0,962	0,989	0,966	0,845	0,983	0,892	0,637	0,947
CA-SF3	0,880	0,969	0,921	0,807	1,000	0,988	0,578	0,950
CH-Oe1	0,818	0,957	0,807	0,778			0,584	0,952
DE-Hai	0,932	0,983	0,960	0,816			0,667	0,950
DE-Kli	0,799	0,963	0,968	0,934	1,000	1,000	0,596	0,953
DE-Obe	0,939	0,982	0,937	0,942			0,793	0,959
DE-Tha	0,949	0,986	0,943	0,834	1,000	0,972	0,755	0,956
DK-Sor	0,968	0,991	0,983	0,945	1,000	1,000	0,786	0,961
DK-ZaH	0,671	0,946	0,945	0,854	1,000	0,998	0,515	0,968
FI-Hyy	0,960	0,990	0,965	0,914	1,000	1,000	0,699	0,953
FR-Gri	0,760	0,952	0,962	0,845			0,610	0,957
GF-Guy	0,722	0,949	0,677	0,970	1,000	1,000	0,699	0,949
IT-Lav	0,930	0,984	0,914	0,975	1,000	0,999	0,711	0,958
IT-Noe	0,797	0,963	0,860	0,986			0,559	0,954
IT-Ren	0,906	0,978	0,918	0,953	0,990	1,000	0,638	0,952
IT-Ro1	0,866	0,968	0,929	0,956	1,000	0,994	0,585	0,959
IT-SRo	0,872	0,979	0,903	0,975	1,000	1,000	0,647	0,959
IT-Tor	0,954	0,987	0,966	0,897	0,575	0,511	0,711	0,955
NL-Hor	0,959	0,985	0,953	0,597	1,000	0,956	0,643	0,954
NL-Loo	0,948	0,986	0,941	0,956			0,746	0,955
RU-Fyo	0,935	0,985	0,946	0,865	1,000	1,000	0,596	0,951
US-ARM	0,711	0,953	0,902	0,965	1,000	1,000	0,515	0,952
US-Blo	0,804	0,956	0,862	0,964	1,000	1,000	0,608	0,964
US-IB2	0,921	0,980	0,974	0,915	0,990	1,000	0,565	0,948
US-MMS	0,925	0,977	0,944	0,910	1,000	0,997	0,541	0,946
US-SRM	0,800	0,937	0,885	0,753			0,626	0,947
US-Ton	0,834	0,950	0,877	0,985	1,000	1,000	0,547	0,951
US-UMd	0,942	0,980	0,964	0,846	0,959	0,996	0,535	0,945
US-Var	0,844	0,953	0,954	0,965	1,000	1,000	0,560	0,958
US-Whs	0,741	0,933	0,902	0,721	0,991	0,906	0,580	0,954
US-Wkg	0,810	0,962	0,856	0,960	1,000	0,976	0,476	0,954
ZA-Kru	0,778	0,956	0,938	0,955	1,000	1,000	0,534	0,951

Tabella 9: R^2 di fitting dei modelli a confronto. ANN e RF sono i modelli creati con driver e GPP non decomposte, ANN_all e RF_all sono i modelli creati con tutte le componenti dei driver (S+T+R) e con GPP non decomposta, ANN_T e RF_T e ANN_R e RF_R sono i modelli addestrati rispettivamente con i trend e con i residui dei driver e della GPP. In giallo sono evidenziati gli R^2 più bassi. In grigio i siti che non presentano un trend nella GPP. I valori si riferiscono agli R^2 medi di 100 alberi e 120 reti per ogni sito.

Nella Tabella 9 sono riportati i valori relativi ai coefficienti di determinazione (R^2) che rilasciano un'indicazione relativa della capacità di fitting del modello.

I modelli addestrati con i segnali non decomposti, hanno un'ottima capacità di fitting. Le RF risultano più flessibili, mostrando R^2 maggiori rispetto alle ANN.

I modelli addestrati con tutte le componenti delle variabili (ANN_all e RF_all) hanno performance molto alte sia per le ANN che per le RF. Solo in alcuni siti che solo in alcuni siti mostrano R^2 inferiori a 0.8.

Le ANN e le RF create con le componenti relative al trend (T) hanno performance alte riuscendo a fittare in modo ottimale il flusso della GPP sul lungo periodo. (tranne nel sito IT-Tor dove gli R^2 risultano più bassi, riuscendo a fittare correttamente circa il 50% dei dati).

Le performance di fitting dei modelli addestrati solo con i residui (R) risultano migliori per le RF (superiori a 0.9), e più basse nelle ANN, in cui per alcuni siti riescono a fittare circa il 50% dei dati.

I modelli addestrati con le differenti componenti sono poi stati analizzati per l'identificazione delle variabili principali usando i tre metodi come descritto nel paragrafo 2.2 (FM, DM e RM).

3.5. Stima importanza delle variabili driver e sintesi dei risultati

I tre metodi sono stati inizialmente applicati ai modelli addestrati con le variabili originali e calcolati i relativi indici di importanza relativa in modo da rendere i risultati comparabili (VRI). Le figure di questo paragrafo riportano a titolo esemplificativo i risultati relativi al sito DE-Tha.

La figura 13 rappresenta il VRI dei sei driver non decomposti per i tre metodi.

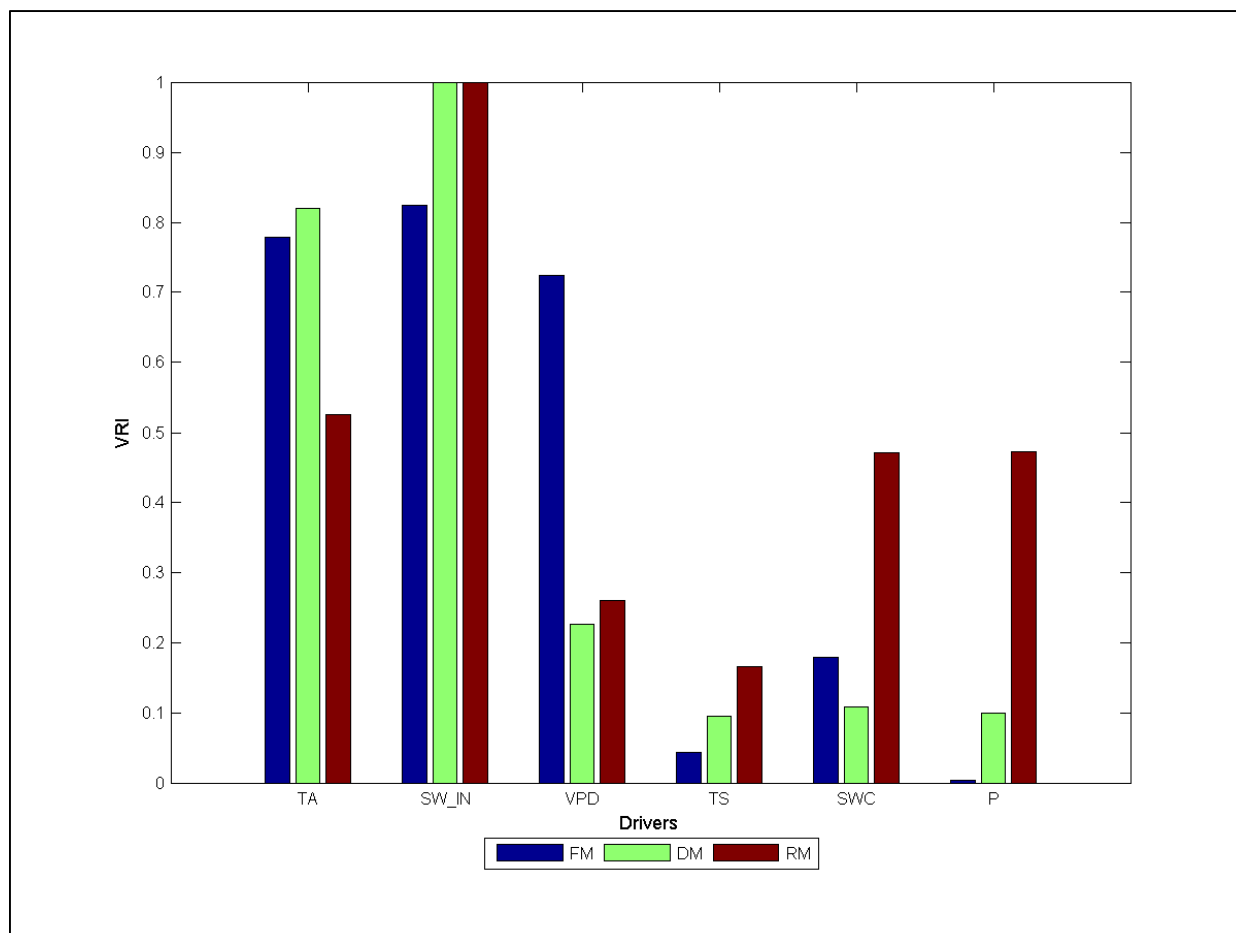


Figura 13: VRI riportato dai 3 metodi (FM – Forward stepwise method; DM- Destruction order method e RM – Random Forest method) per i driver non decomposti, con in output la GPP non decomposta (sito DE-Tha).

La stessa analisi è stata svolta sui modelli addestrati utilizzando tutte e tre le componenti dei driver (Figura 14). La componente stagionale dei driver (S), se presente, è risultata sempre quella che maggiormente guida del segnale GPP non decomposto, con VRI molto più elevati rispetto alle componenti T e R. Questo risultato era atteso in quanto la GPP ha un forte ciclo stagionale che quindi viene facilmente descritto dalle variabili con simile dinamica. Per questo motivo nei passaggi seguenti ci si è concentrati solo sull'analisi delle componenti trend e residui.

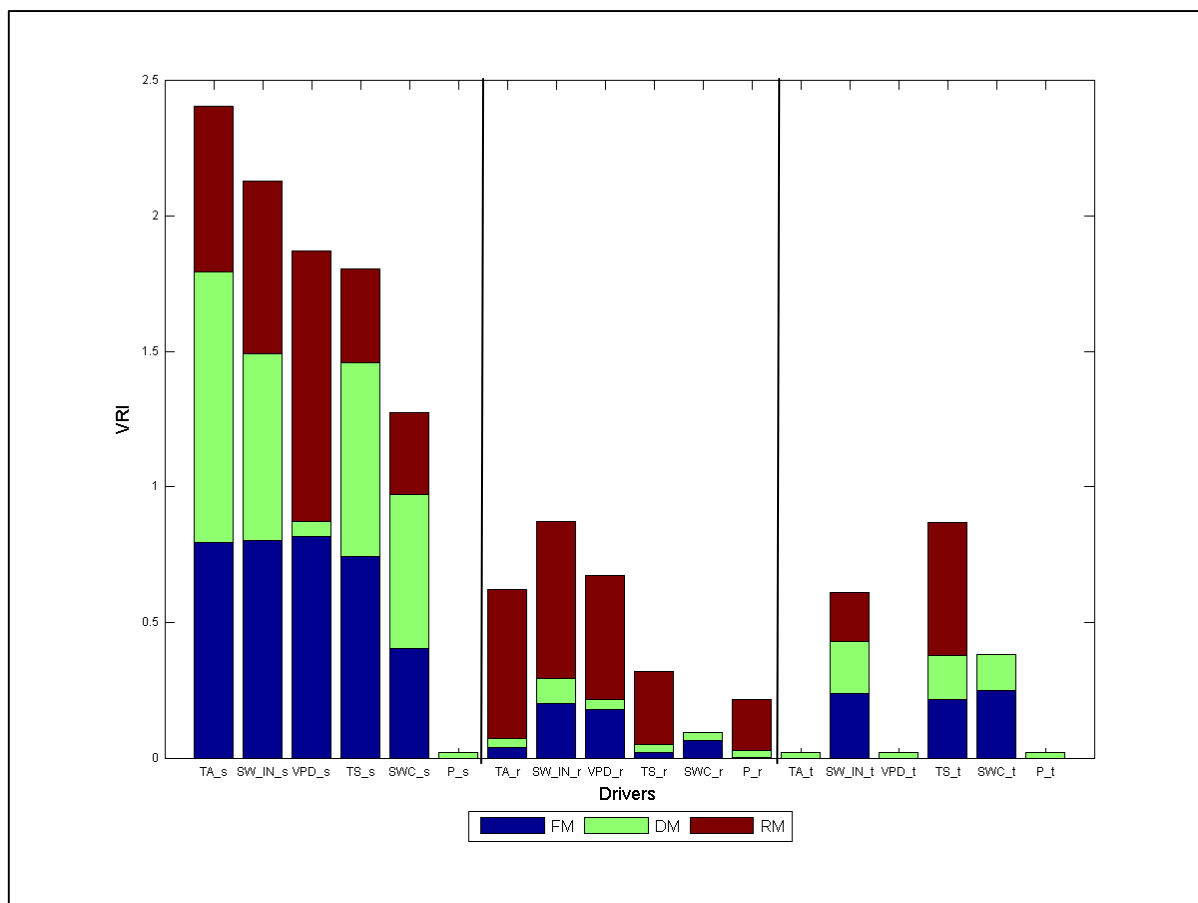


Figura 14: VRI riportato dai 3 metodi (FM – Forward stepwise method; DM- Destruction order method e RM – Random Forest method) per i segnali s (stagionalità), r (residui) e t (trend) dei driver con la GPP non decomposta (sito DE-Tha).

Si è proceduto quindi all'applicazione dei metodi alle componenti di lungo periodo (rappresentate dai trend - T) e di breve periodo (rappresentate dai residui - R), per individuare quale variabile influenzasse maggiormente il trend della GPP (Figura 15) e quale fosse il driver che più spiegava le alte frequenze della GPP, ovvero nel breve periodo (Figura 16).

Sono state prese in considerazione solo le prime 3 variabili individuate come driver principali della GPP da ogni metodo.

I grafici riportanti i VRI dei 3 metodi, applicati alle variabili prima della decomposizione, e dopo la decomposizione (con tutti i segnali, con il trend delle variabili e con i residui), sono esposti in appendice B.

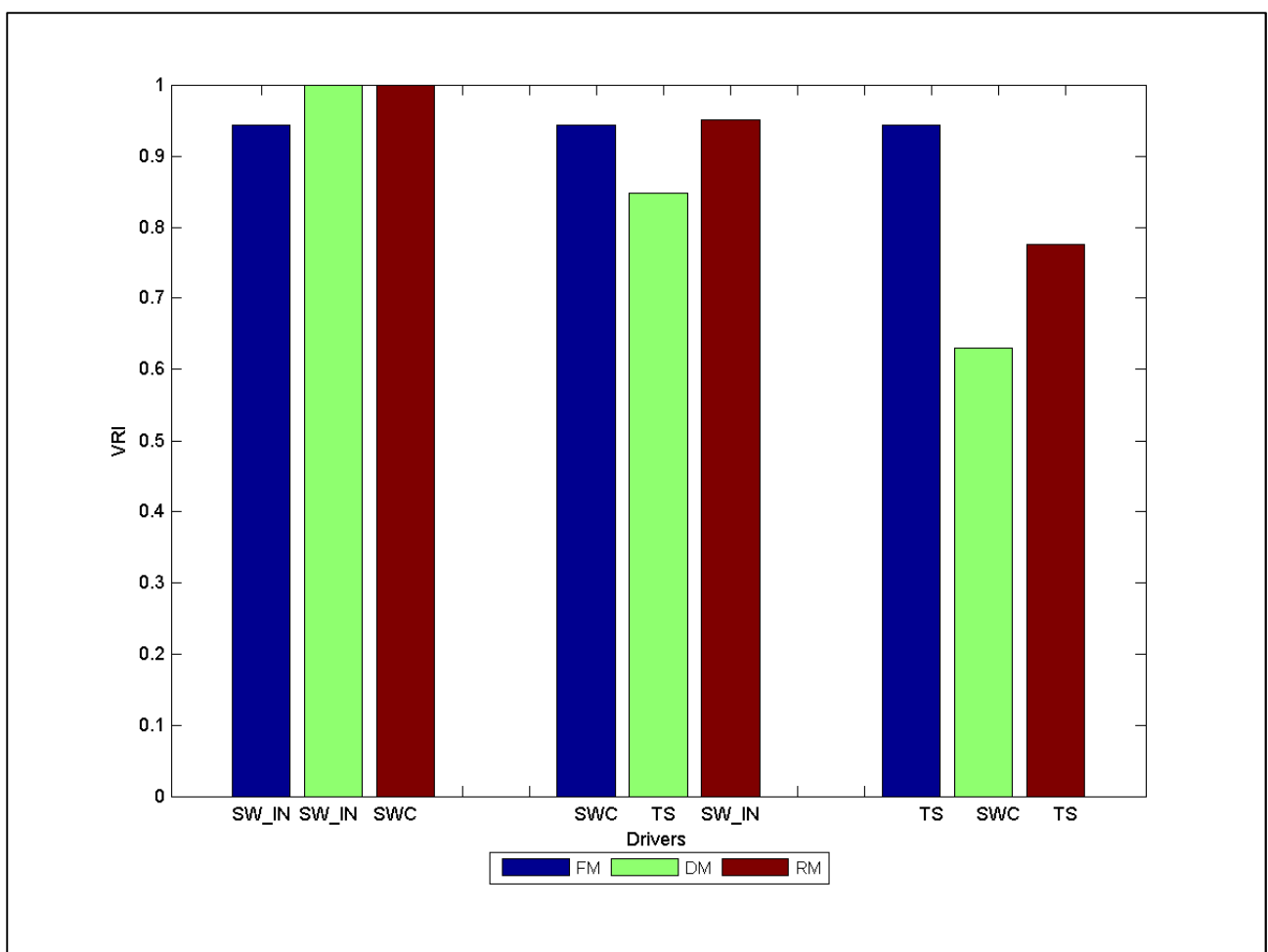


Figura 15: VRI riportato dai 3 metodi (FM – Forward stepwise method; DM- Destruction order method e RM – Random Forest method) per i trend dei driver con la GPP di lungo periodo (sito DE-Tha).

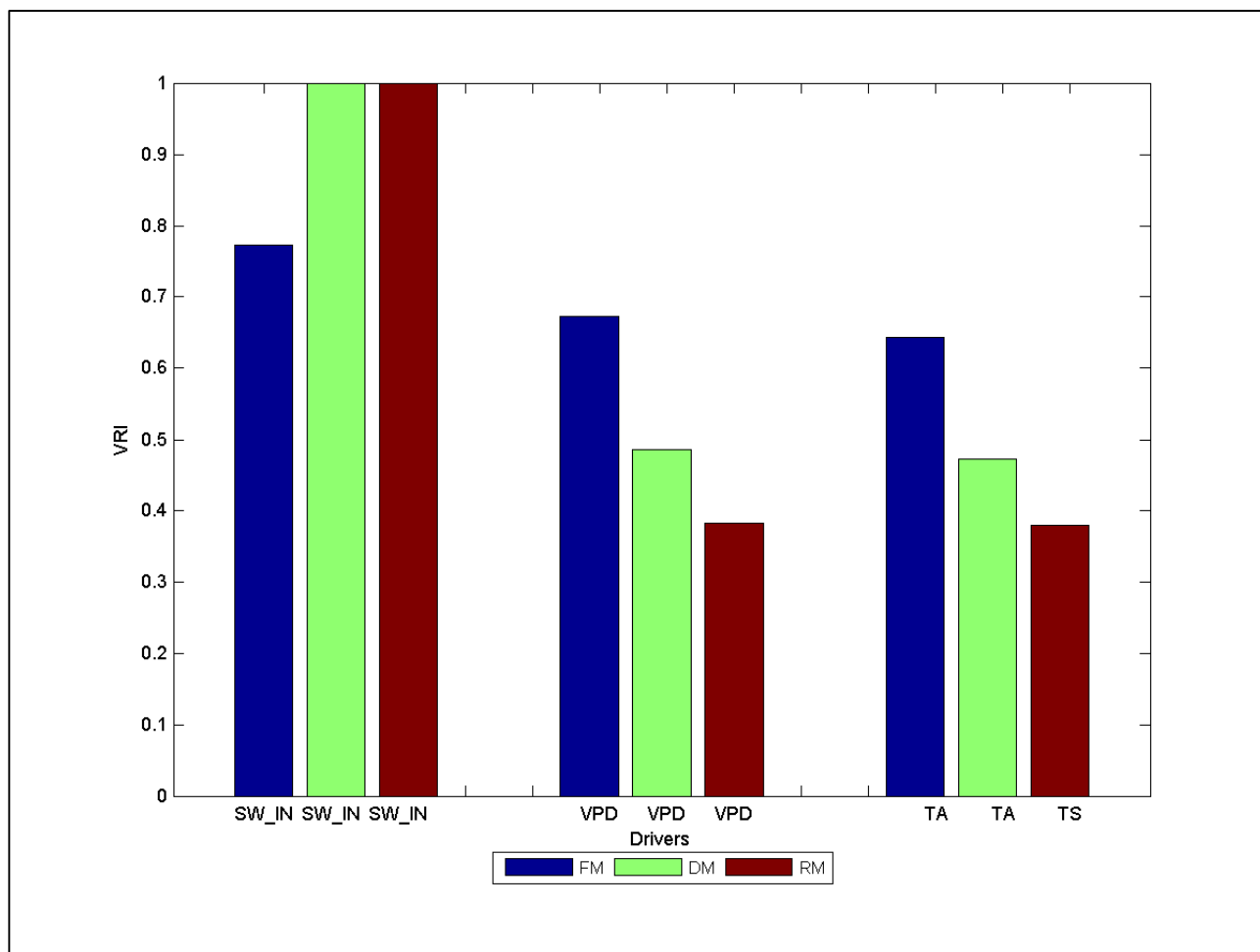


Figura 16: VRI riportato dai 3 metodi (FM – Forward stepwise method; DM- Destruction order method e RM – Random Forest method) per i residui dei driver con i residui GPP (sito DE-Tha).

3.5.1 Confronto dei risultati tra siti

Al fine di confrontare i risultati ottenuti per i diversi siti relativi all'importanza delle variabili nello spiegare le variabilità a bassa ed alta frequenza, i VRI ottenuti dai tre metodi impiegati sono stati mediati, assegnando alle variabili non selezionate valore $VRI = 0$. Il VAI è stato calcolato per ciascuna delle prime tre variabili (più importanti) per ogni sito moltiplicando il VRI medio per la meff del modello (tabella 10).

SITE_ID	IGBP	TREND						RESIDUI					
		SW-IN	TA	TS	VPD	SWC	P	SW-IN	TA	TS	VPD	SWC	P
AT-Neu	GRA	0,00	0,62	0,00	0,92	0,63	0,60	0,22	0,00	0,11	0,05	0,00	0,16
AU-DaP	GRA	0,00	0,00	0,68	0,67	0,65	0,00	0,42	0,00	0,34	0,00	0,30	0,00
AU-DaS	SAV	0,00	0,00	0,82	0,62	0,49	0,61	0,29	0,08	0,20	0,00	0,03	0,14
AU-Tum	EBF	0,51	0,56	0,00	0,59	0,98	0,00	0,49	0,00	0,00	0,08	0,35	0,21
BE-Lon	CRO	0,00	0,86	0,00	0,60	0,26	0,95	0,20	0,08	0,00	0,00	0,17	0,08
BE-Vie	MF	0,00	0,00	0,17	0,61	0,96	0,39	0,49	0,00	0,09	0,19	0,28	0,00
CA-Qfo	ENF	0,00	0,00	0,23	0,00	0,17	0,00	0,37	0,15	0,23	0,00	0,00	0,06
CA-SF3	OSH	0,00	0,77	0,00	0,89	0,00	0,69	0,31	0,05	0,06	0,19	0,00	0,06
CH-Oe1	GRA	0,00	0,00	0,00	0,00	0,00	0,00	0,29	0,11	0,18	0,06	0,00	0,00
DE-Hai	DBF	0,00	0,00	0,00	0,00	0,00	0,00	0,41	0,16	0,00	0,30	0,08	0,00
DE-Kli	CRO	0,00	0,58	0,33	0,55	0,31	0,97	0,27	0,10	0,21	0,00	0,09	0,00
DE-Obe	ENF	0,00	0,00	0,00	0,00	0,00	0,00	0,61	0,26	0,00	0,36	0,09	0,00
DE-Tha	ENF	0,93	0,00	0,83	0,00	0,82	0,00	0,53	0,21	0,07	0,29	0,00	0,00
DK-Sor	DBF	1,00	0,00	0,00	0,43	0,57	0,48	0,59	0,15	0,00	0,08	0,22	0,14
DK-ZaH	GRA	0,60	0,00	0,67	0,99	0,12	0,00	0,00	0,00	0,22	0,00	0,31	0,25
FI-Hyy	ENF	0,13	0,33	0,33	0,00	0,69	0,98	0,46	0,07	0,00	0,27	0,00	0,20
FR-Gri	CRO	0,00	0,00	0,00	0,00	0,00	0,00	0,23	0,12	0,26	0,00	0,16	0,00
GF-Guy	EBF	0,53	0,00	0,27	0,99	0,33	0,53	0,46	0,00	0,22	0,07	0,00	0,30
IT-Lav	ENF	0,00	0,00	0,59	0,64	0,57	0,90	0,49	0,17	0,00	0,27	0,08	0,00
IT-Noe	CSH	0,00	0,00	0,33	0,00	0,16	0,27	0,27	0,08	0,00	0,20	0,11	0,00
IT-Ren	ENF	0,99	0,33	0,00	0,33	0,47	0,55	0,39	0,00	0,00	0,23	0,29	0,00
IT-Ro1	DBF	0,00	0,60	0,81	0,00	0,82	0,20	0,31	0,00	0,00	0,12	0,26	0,11
IT-SRo	ENF	0,62	0,00	0,33	0,54	0,49	0,52	0,41	0,00	0,08	0,07	0,24	0,07
IT-Tor	GRA	0,00	0,00	0,00	0,08	0,00	0,00	0,46	0,00	0,21	0,00	0,32	0,07
NL-Hor	GRA	0,00	0,26	0,19	0,00	0,00	0,00	0,39	0,08	0,25	0,00	0,20	0,00
NL-Loo	ENF	0,31	0,33	0,00	0,00	0,00	0,30	0,53	0,14	0,12	0,20	0,07	0,00
RU-Fyo	ENF	0,49	0,00	0,33	0,00	0,66	0,85	0,33	0,00	0,00	0,07	0,24	0,16
US-ARM	CRO	0,64	0,00	0,58	0,12	0,95	0,00	0,20	0,00	0,00	0,08	0,19	0,11
US-Blo	ENF	0,54	0,00	0,85	0,33	0,33	0,60	0,36	0,12	0,03	0,12	0,22	0,14
US-IB2	GRA	0,00	0,54	0,33	0,64	0,33	0,86	0,30	0,00	0,00	0,15	0,22	0,06
US-MMS	DBF	0,00	0,00	0,32	0,78	0,96	0,62	0,26	0,00	0,00	0,19	0,16	0,00
US-SRM	WSA	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,06	0,00	0,00	0,22	0,35
US-Ton	WSA	0,00	0,59	0,63	0,84	0,60	0,00	0,13	0,08	0,09	0,13	0,22	0,00
US-UMd	DBF	0,00	0,86	0,90	0,31	0,53	0,00	0,28	0,00	0,18	0,13	0,07	0,00
US-Var	GRA	0,91	0,53	0,56	0,00	0,00	0,62	0,14	0,10	0,08	0,10	0,13	0,10
US-Whs	OSH	0,00	0,00	0,00	0,16	0,18	0,00	0,00	0,20	0,00	0,10	0,14	0,28
US-Wkg	GRA	0,00	0,00	0,00	0,00	0,18	0,26	0,08	0,00	0,00	0,06	0,17	0,19
ZA-Kru	SAV	0,55	0,94	0,53	0,21	0,18	0,00	0,15	0,00	0,16	0,10	0,02	0,11

Tabella 10: VAI relativo ad ogni sito per i TREND e i RESIDUI di tutte le variabili driver.

Sono state plottate anche le relazioni tra i VAI medi, da cui stati eliminati i valori 0 per rendere più chiara l'interpretazione, sia per il trend dei driver sia per i residui (figura 17 e figura 18).

Per il trend non vi sono relazioni particolarmente evidenti, l'unico andamento simile del VAI si riscontra per TA e TS. Per quanto riguarda i residui, gli unici due driver che sembrano avere un VAI con andamento simile, sono l'SW-IN e la TA.

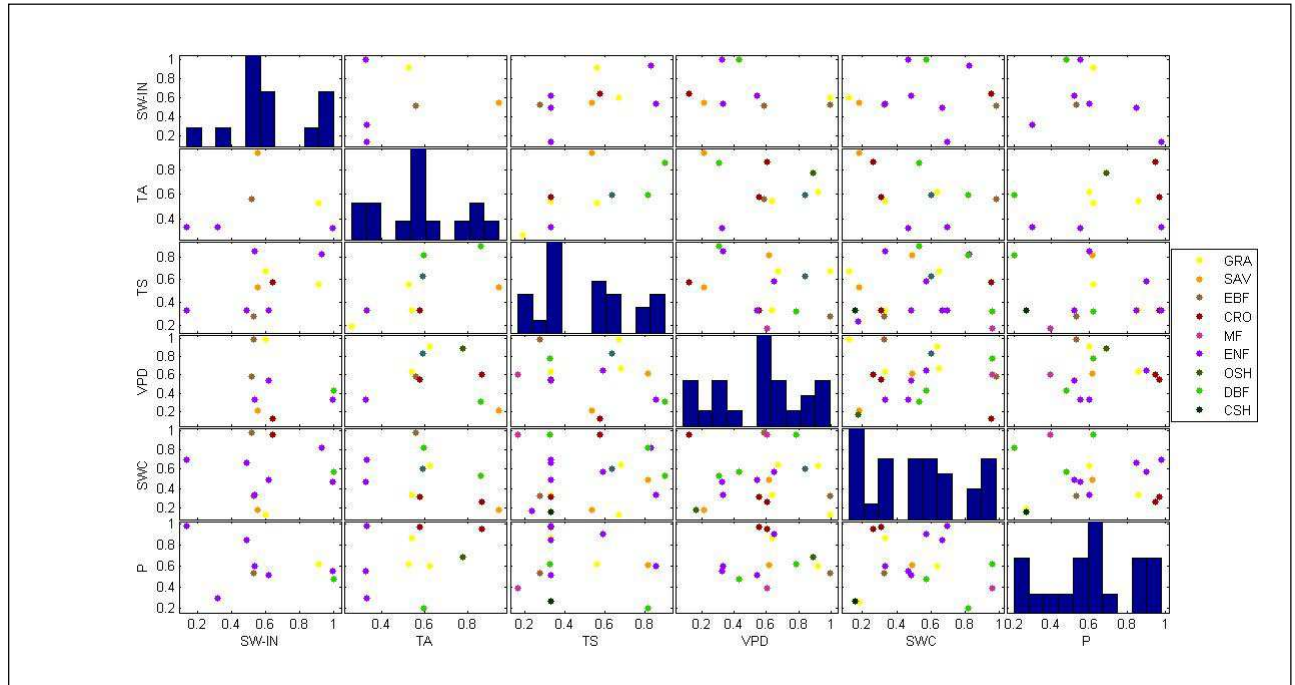


Figura 17: Scatterogramma per visualizzare le relazioni tra i VAI dei diversi driver per il trend.

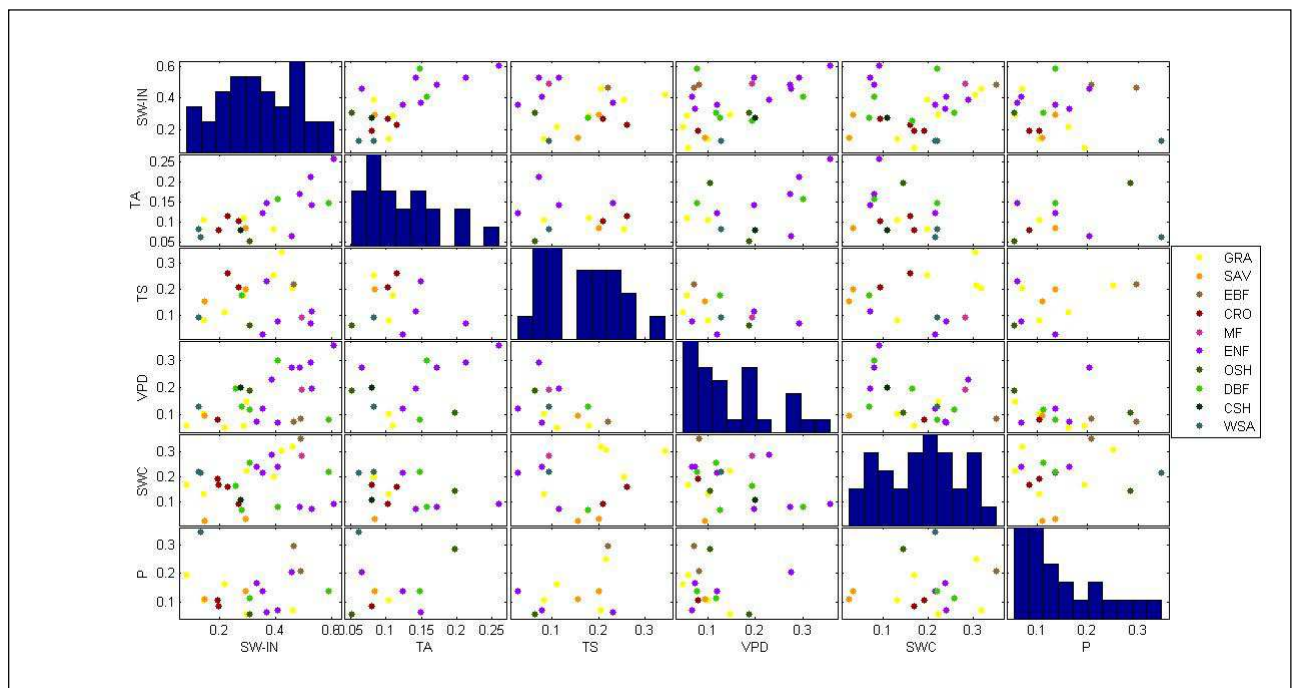


Figura 18: Scatterogramma per visualizzare le relazioni tra i VAI dei diversi driver per i residui.

3.5.2 Relazione tra importanza variabili e caratteristiche dei siti

I VAI sono stati confrontati con le caratteristiche di ogni sito: IGBP, precipitazioni medie annue (MAP), temperatura media annua (MAT), leaf area index (LAI), altitudine (Alt) e latitudine (Lat). Al fine di individuare se ci fossero delle relazioni dirette tra le caratteristiche dei siti che potessero confondere l'analisi, queste sono state confrontate attraverso scatterogrammi (figura 19). Solo latitudine e temperatura media annua sono risultate correlate, presentando una relazione di carattere inverso come atteso.

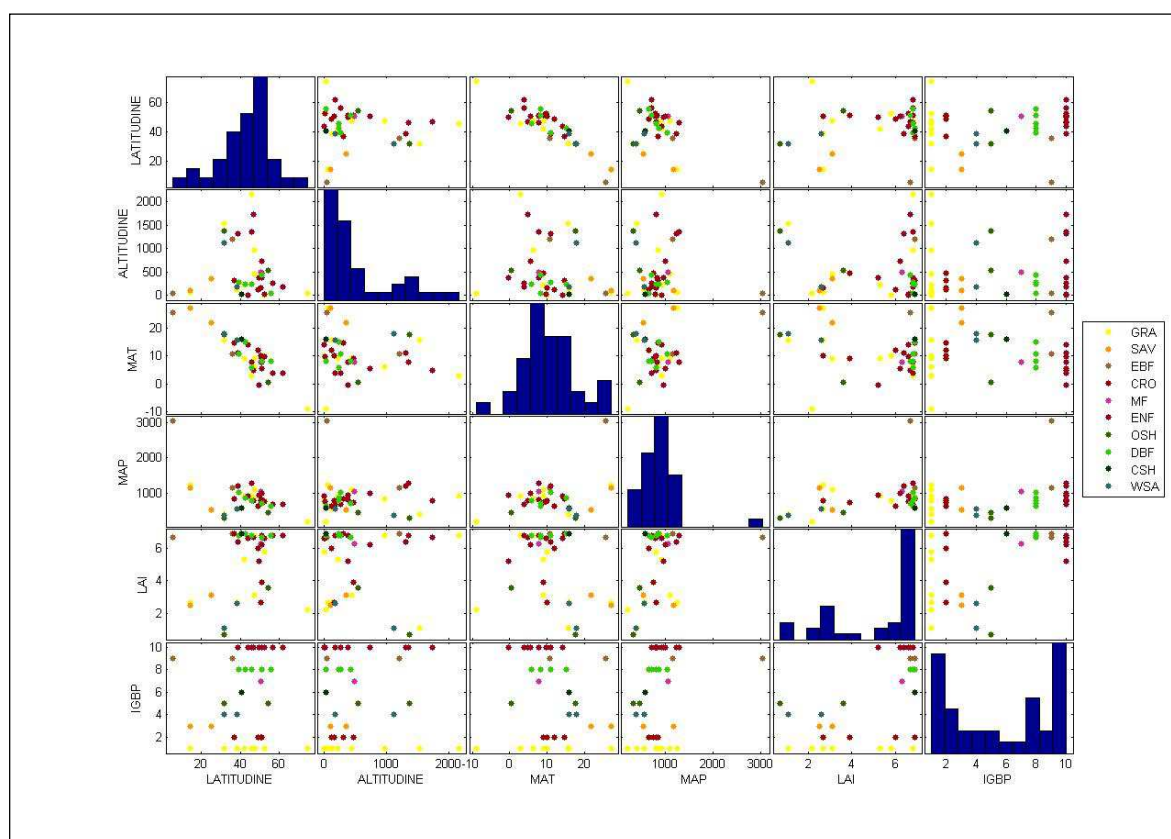


Figura 19: Scatterogramma per visualizzare le relazioni tra le diverse caratteristiche dei siti prese in considerazione (LATITUDINE, ALTITUDINE, MAT, MAP, LAI e IGBP).

Nelle figure 20 e 21 sono riportati i VAI per ogni driver (asse x) e per ogni sito (asse y) associando una scala di colori alla grandezza del VAI, quindi all'importanza attribuita ai driver, rispettivamente per i trend e per i residui. In verde sono riportati i siti forestali, mentre in giallo quelli non forestali.

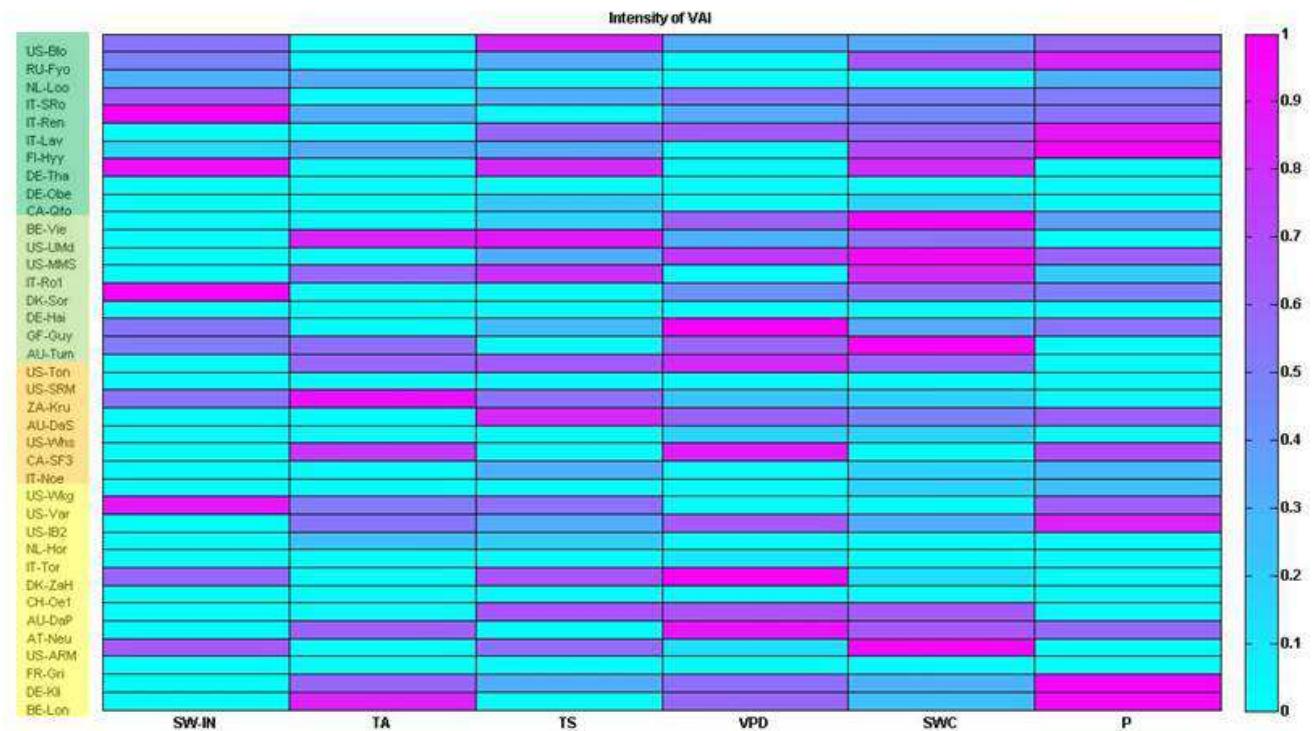


Figura 20: Intensità del VAI per i trend di tutti i siti. Più il valore tende al rosso maggiore sarà l'importanza attribuita alla variabile corrispondente. In verde sono riportati i siti forestali (verde scuro gli ENF, in verde chiaro MF, DBF e EBF), mentre i siti con altri *plant functional type* sono riportati in giallo (GRA e CRO) e in arancione (SAV, WSA, OSH, CSH).

I VAI delle variabili che spiegano il trend della GPP appaiano molto diversificati. In generale si nota un VAI più elevato per i drivers legati alla disponibilità d'acqua (VPD, SWC e P) soprattutto per le ENF nel guidare la GPP nel lungo periodo.

L'SW-IN è la variabile a cui i metodi attribuiscono maggiore importanza per la GPP nel breve periodo, per tutti i siti ad eccezione di DK-ZaH e US-Whs.

In tutti i siti forestali il VAI dell'SW-IN risulta più alto rispetto al VAI attribuito all'SW-IN per i siti non forestali, mostrando una maggiore importanza nella spiegazione della GPP di breve periodo nelle foreste.

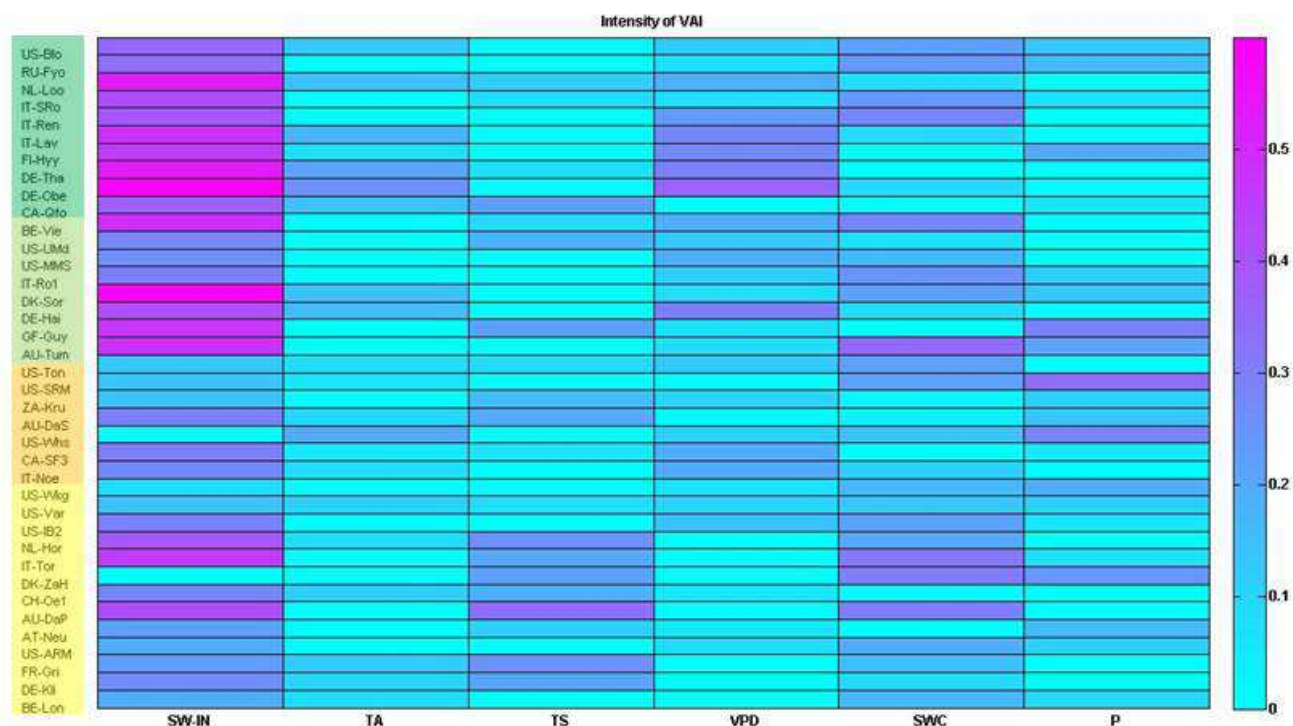


Figura 21: Intensità del VAI per i residui di tutti i siti. Più il valore tende al rosa maggiore sarà l'importanza attribuita alla variabile corrispondente. In verde sono riportati i siti forestali (verde scuro gli ENF, in verde chiaro MF, DBF e EBF), mentre i siti con altri *plant functional type* sono riportati in giallo (GRA e CRO) e in arancione (SAV, WSA, OSH, CSH).

I siti, suddivisi in forestali (EBF, DBF, MF, ENF) e non forestali (CRO, GRA, CSH, OSH, SAV, WSA), sono stati messi in relazione con l'indice di aridità di De Martonne (1923), e ordinati dal sito più arido a quello più umido (figure 22 e 23).

Non sono presenti siti forestali aridi, (definiti come indice inferiore a 15), mentre i siti non forestali US-Whs, US-SRM, US-Wkg e ZA-Kru sono siti aridi e semiaridi.

Il sito più arido (US-Whs) e quello più umido (DK-ZaH) sono gli unici due siti per cui la SW-IN non viene individuata come driver della GPP residua.

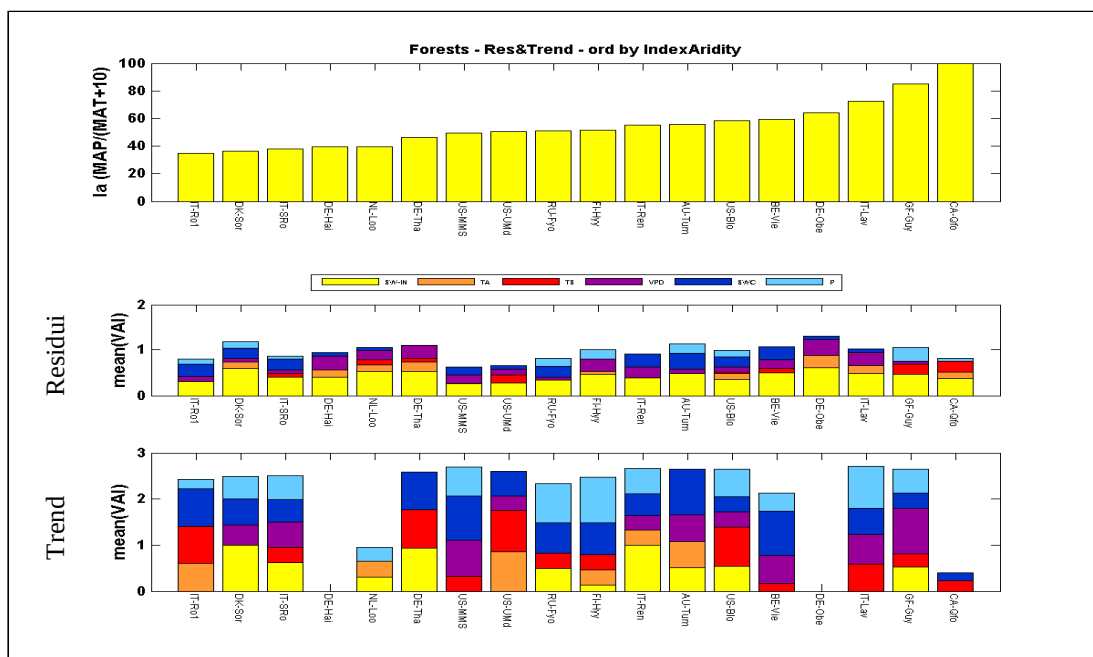


Figura 22: VAI dei siti forestali in ordine di aridità secondo l'indice di De Martonne, dal più arido al più umido. Nel primo subplot sono riportati gli indici di aridità attribuiti ai siti. Nel secondo subplot sono riportati i VAI attribuiti dai metodi ai residui dei driver nello spiegare la GPP di breve periodo. Nel terzo subplot sono riportati i VAI attribuiti dai metodi ai trend dei driver nello spiegare la GPP di lungo periodo. I colori differenti per il secondo e terzo subplot sono relativi ai singoli driver come mostrato dalla legenda.

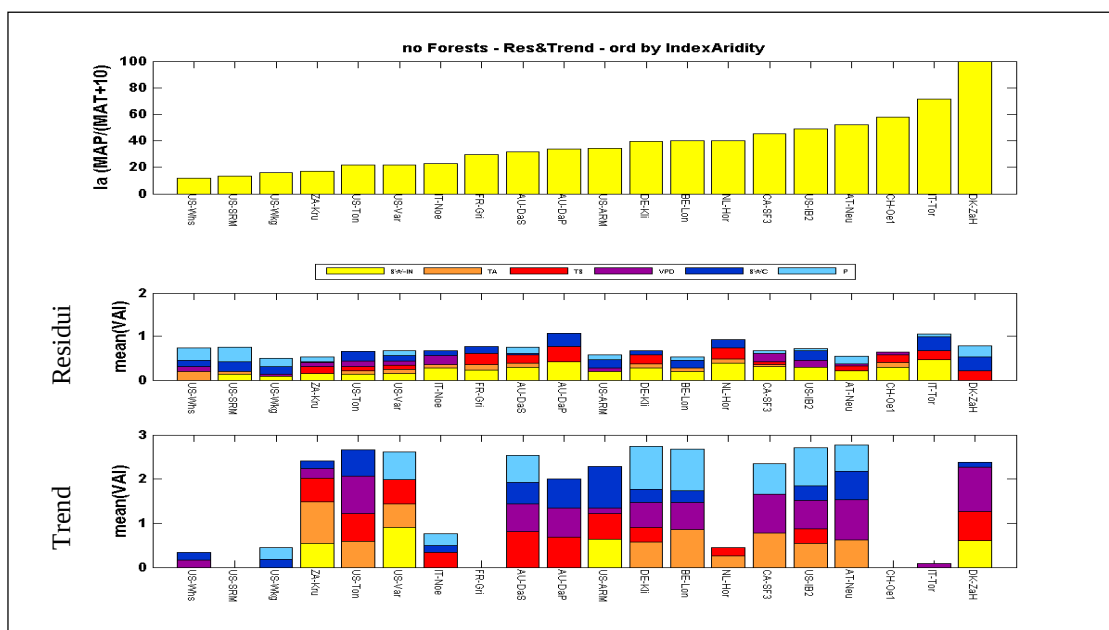


Figura 23: VAI dei siti non forestali in ordine di aridità secondo l'indice di De Martonne, dal più arido al più umido. Nel primo subplot sono riportati gli indici di aridità attribuiti ai siti. Nel secondo subplot sono riportati i VAI attribuiti dai metodi ai residui dei driver nello spiegare la GPP di breve periodo. Nel terzo subplot sono riportati i VAI attribuiti dai metodi ai trend dei driver nello spiegare la GPP di lungo periodo. I colori differenti per il secondo e terzo subplot sono relativi ai singoli driver come mostrato dalla legenda.

Per un'analisi dei risultati da un punto di vista ecologico, ai siti è stata attribuita la classificazione FAO *global ecological zones*, che riassume le caratteristiche climatiche a livello mondiale.

Nelle tabelle 11 e 12, sono riportati rispettivamente i siti non forestali (CRO, GRA, CSH, OSH, SAV e WSA) e quelli forestali (EBF, DBF, MF, ENF) con le relative caratteristiche fitoclimatiche (IGBP, la classificazione FAO e l'indice di aridità *Ia*) e la variabile individuata dai metodi come driver primario della GPP nel breve (R) e nel lungo periodo (T). Tra parentesi sono riportate le seconde variabili driver più importanti.

Tra i siti non forestali, la GPP nel breve periodo è guidata dall'SW-IN tranne che nel sito DK-ZaH (classificato come *Polar* dalla classificazione FAO) dove la GPP residua sembra essere influenzata dal SWC, e nei tre siti classificati come *Subtropical desert*, che hanno come driver primario le precipitazioni (P).

Per quanto riguarda il trend della GPP per i siti non forestali, il VPD è il driver che la condiziona maggiormente nei due siti più freddi (DK-ZaH e CA-SF3), seguito dalle temperature. Anche per i siti ricadenti in ecosistemi montani (US-Ton e AT-Neu), ad eccezione di US-Var (dove il trend della GPP è guidato da SW-IN), il driver che influenza di più la GPP nel lungo periodo è il VPD. Nei siti tropicali la GPP è guidata dai trend delle temperature; TA è il driver primario per ZA-Kru (sito semiarido), mentre per AU-DaP e AU-DaS (siti umidi) la variabile più importante è la TS. La TS viene individuata come driver di lungo periodo anche per IT-Noe. Nei siti subtropicali desertici la disponibilità d'acqua condiziona la GPP, infatti SWC per US-Whs e P per US-Wkg sono i drivers selezionati dai metodi sul lungo periodo. Nelle due cropland che appartengono alla zona FAO *Temperate oceanic forest* (BE-Lon e DE-Kli), il trend della GPP è spiegato dalle precipitazioni, e in entrambi i siti come secondo driver viene individuata la TA. Le precipitazioni guidano anche il trend della GPP per la grassland US-IB2, mentre la TA spiega il trend della GPP nel sito NL-Hor.

SITE_ID	IGBP	FAO	Ia	Driver_GPP_R	Driver_GPP_T
DK-ZaH	GRA	Polar	PERIUMIDO	SWC (P)	VPD (TS)
CA-SF3	OSH	Boreal coniferous forest	UMIDO	SW-IN (VPD)	VPD (TA)
BE-Lon	CRO	Temperate oceanic forest	UMIDO	SW-IN (SWC)	P (TA)
CH-Oe1	GRA	Temperate oceanic forest	UMIDO	SW-IN (TS)	\
DE-Kli	CRO	Temperate oceanic forest	UMIDO	SW-IN (TS)	P (TA)
FR-Gri	CRO	Temperate oceanic forest	UMIDO	SW-IN (TS)	\
NL-Hor	GRA	Temperate oceanic forest	UMIDO	SW-IN (TS)	TA (TS)
US-IB2	GRA	Temperate continental forest	UMIDO	SW-IN (SWC)	P (VPD)
AT-Neu	GRA	Temperate mountain system	UMIDO	SW-IN (P)	VPD (SWC&TS)
IT-Tor	GRA	Temperate mountain system	PERIUMIDO	SW-IN (SWC)	\
US-ARM	CRO	Temperate steppe	UMIDO	SW-IN (SWC)	SWC (SW-IN)
IT-Noe	CSH	Subtropical dry forest	SUBUMIDO	SW-IN (VPD)	TS (P)
US-Ton	WSA	Subtropical mountain system	SUBUMIDO	SWC (SW-IN)	VPD (TS)
US-Var	GRA	Subtropical mountain system	SUBUMIDO	SW-IN (SWC)	SW-IN (P)
US-SRM	WSA	Subtropical desert	ARIDO	P (SWC)	\
US-Whs	OSH	Subtropical desert	ARIDO	P (TA)	SWC (VPD)
US-Wkg	GRA	Subtropical desert	SEMIARIDO	P (SWC)	P (SWC)
AU-DaP	GRA	Tropical dry forest	UMIDO	SW-IN (TS)	TS (VPD&SWC)
AU-DaS	SAV	Tropical dry forest	UMIDO	SW-IN (TS)	TS (VPD&P)
ZA-Kru	SAV	Tropical dry forest	SEMIARIDO	SW-IN (TS)	TA (SW-IN)

Tabella 11: Siti non Forestali - Caratteristiche fitoclimatiche a confronto con i primi due drivers selezionati dai metodi

I risultati riportati dalla tabella 12, mostrano come per tutti i siti forestali, la prima variabile guida della GPP nel breve periodo è l'SW-IN. Ad eccezione del sito CA-Qfo, US-UMd e GF-Guy, le seconde variabili individuate sono sempre SWC e VPD. Per il sito tropicale (GF-Guy) il secondo driver sono le precipitazioni, mentre per CA-Qfo e US-UMd la temperatura del suolo.

La GPP nel lungo periodo è guidata da differenti driver in base ai siti. Nel sito più freddo (CA-Qfo) è influenzata maggiormente dalla TS, mentre nel sito tropicale (GF-Guy) dal VPD. TS è anche il driver principale nei siti US-UMd e US-Blo. SWC influenza il trend della GPP in tre siti (BE-Vie, US-UMMS e IT-Ro1). P guida la GPP per i siti montani di conifere (FI-Hyy, RU-Fyo e IT-Lav). SW-IN condiziona la GPP nei siti DE-Tha e DK-Sor (seguito dal SWC), IT-Ren (seguito dalle precipitazioni) e IT-SRo (che ha come secondo driver il VPD).

SITE_ID	IGBP	FAO	Ia	Driver_GPP_R	Driver_GPP_T
CA-Qfo	ENF	Boreal coniferous forest	PERIUMIDO	SW-IN (TS)	TS (SWC)
FI-Hyy	ENF	Boreal coniferous forest	UMIDO	SW-IN (VPD)	P (SWC)
BE-Vie	MF	Temperate oceanic forest	UMIDO	SW-IN (SWC)	SWC (VPD)
DE-Hai	DBF	Temperate oceanic forest	UMIDO	SW-IN (VPD)	\
DE-Tha	ENF	Temperate oceanic forest	UMIDO	SW-IN (VPD)	SW-IN (TS&SWC)
DK-Sor	DBF	Temperate oceanic forest	UMIDO	SW-IN (SWC)	SW-IN (SWC)
NL-Loo	ENF	Temperate oceanic forest	UMIDO	SW-IN (VPD)	TA (SW-IN&P)
RU-Fyo	ENF	Temperate continental forest	UMIDO	SW-IN (SWC)	P (SWC)
US-MMS	DBF	Temperate continental forest	UMIDO	SW-IN (VPD)	SWC (VPD)
US-UMd	DBF	Temperate continental forest	UMIDO	SW-IN (TS)	TS (TA)
AU-Tum	EBF	Temperate mountain system	UMIDO	SW-IN (SWC)	SWC (VPD)
DE-Obe	ENF	Temperate mountain system	PERIUMIDO	SW-IN (VPD)	\
IT-Lav	ENF	Temperate mountain system	PERIUMIDO	SW-IN (VPD)	P (VPD)
IT-Ren	ENF	Temperate mountain system	UMIDO	SW-IN (SWC)	SW-IN (P)
IT-Ro1	DBF	Subtropical dry forest	UMIDO	SW-IN (SWC)	SWC (TS)
IT-SRo	ENF	Subtropical dry forest	UMIDO	SW-IN (SWC)	SW-IN (VPD)
US-Blo	ENF	Subtropical mountain system	UMIDO	SW-IN (SWC)	TS (P)
GF-Guy	EBF	Tropical rainforest	PERIUMIDO	SW-IN (P)	VPD (SW-IN&P)

Tabella 12: Siti Forestali - Caratteristiche fitoclimatiche a confronto con I primi due drivers selezionati dai metodi

4. Discussioni

4.1. Performance di ANN e RF

Dall'analisi delle performance delle reti neurali create nella fase 1, fase relativa alla scelta dell'architettura più idonea, si sono riscontrate buone model efficiency in tutte le tipologie di architettura testate. Le diverse strutture non influenzano quindi in modo significativo gli output della rete.

Le ANN e le RF create con i segnali originali delle variabili, mostrano performance relativamente alte (> 0.6). Per la maggior parte dei siti i driver selezionati sono risultati quindi sufficienti a spiegare il flusso della GPP senza andare incontro ad overfitting. Questi modelli mostrano quindi una buona flessibilità nel riprodurre tale flusso e, per quanto riguarda le ANN, l'architettura della rete non influenza la capacità del modello nel riuscire a spiegare il fenomeno ecologico. In futuro quindi si potrebbe considerare l'utilizzo di un'unica struttura della rete, senza andare incontro a perdite di performance.

I modelli addestrati per alcuni siti mostrano performance più basse, questo può essere dovuto a driver di base mancanti. Infatti su sei siti con meff inferiori a 0.6, sono presenti tre cropland. I campi coltivati sono sottoposti a gestioni diverse che ne influenzano flussi e cicli vegetativi che non possono essere spiegati da variabili climatiche. Input derivanti da telerilevamento o informazioni circa le operazioni di gestione potrebbero aiutare a migliorare le performances dei modelli.

L'unico sito forestale a presentare meff dei modelli inferiore alla media è il sito GF-Guy, caratterizzato da una foresta pluviale naturale. Probabilmente le variabili prese in considerazione per i driver, non sono sufficienti a spiegare pienamente la GPP essendo un ecosistema caratterizzato da una grande biodiversità e da processi ecologici molto complessi rispetto alle altre tipologie di foreste.

Entrambi i modelli hanno trovato difficoltà nel modellizzare la GPP di DK-ZaH, il sito più freddo tra quelli analizzati nello studio, caratterizzato da oltre 200 giorni di copertura nevosa (tundra artica); la presenza della neve potrebbe essere un driver mancante ma è anche la causa di serie di misura limitate alla sola stagione vegetativa per questo sito, in particolare nei primi anni, che potrebbero aver influenzato l'addestramento.

Le performances dei modelli sono anche relativamente basse per i siti aridi US-Whs e ZA-Kru (il primo è un cespuglieto dell'Arizona, il secondo è un savana in Sudafrica). In questi ecosistemi le precipitazioni hanno di solito un ruolo importante, non solo come quantità ma anche la loro

distribuzione sia nell'anno che nell'arco della giornata. Questa informazione non è stata utilizzata come driver (la scala era giornaliera) e potrebbe migliorare la capacità predittiva dei due modelli.

4.2. Selezione dei metodi più robusti per la selezione delle variabili guida

Dall'analisi dei risultati ottenuti con l'applicazione delle metodologie per restituire l'importanza dei driver, si è osservato che tutti e sei i metodi testati sono in grado di discriminare i driver che guidano un certo fenomeno dalle variabili che invece non hanno alcun peso nella determinazione dell'output.

Nell'attribuire il rank di importanza all'interno delle variabili che hanno un ruolo nella formazione dell'output i metodi WM (Weight method) e PM (Perturb method) mostrano delle incertezze. Già Gevrey et al. (2003), confrontando questi metodi con altre metodologie, aveva ottenuto dei risultati meno stabili da quelli ottenuti con altri metodi. Le restanti quattro metodologie testate in questo studio invece sono riuscite ad assegnare il rank corretto alle variabili drivers.

Per simulare il rumore presente nei dati eddy covariance e nei drivers, i quattro metodi più robusti sono anche stati testati aggiungendo del rumore alle variabili in input e alla variabile in output. Le metodologie anche in questo caso non hanno incontrato difficoltà nel riconoscere quali fossero le variabili più importanti.

La robustezza dei metodi è stata verificata anche tramite l'aggiunta di una variabile ripetuta all'interno del dataset. In questo modo si è potuta valutare la risposta delle metodologie nel caso di variabili altamente correlate. Da questo test è emerso che il metodo BM (Backward stepwise method), non riesce ad attribuire il rank corretto alle variabili. Esso si basa infatti sulla rimozione di una variabile alla volta; con la presenza di una doppia variabile, non riesce a distinguere l'importanza del driver rimosso perché mantiene all'interno del dataset la sua variabile gemella. Gli altri tre metodi (FM, DM e RM) risultano efficaci anche in questo caso. Anche Gevrey et al. (2003) aveva trovato una differenza nei risultati tra i due metodi *stepwise* (BM e FM), senza individuare però quale dei due fosse più affidabile.

La performance dei modelli utilizzati nelle tre metodologie robuste, è stata testata tramite l'analisi della correlazione di Spearman. Sono state confrontate le risposte dei metodi nei differenti test effettuati. Per quanto concerne i due metodi basati sulle ANN (FM e DM), mostrano un indice di correlazione ρ elevato per tutti i test e, a conferma di quanto esposto in precedenza, non presentano differenti risultati in base alle diverse architetture della rete utilizzate. Tuttavia, l'indice ρ risulta diminuire nel caso del dataset corto (100 valori) evidenziando dei limiti in caso di serie

di dati relativamente corte. Questo aspetto è da tenere in considerazione nel caso di pochi anni di dati o qualora si volesse applicare la metodologia descritta a dati con risoluzione mensile o annuale.

Il metodo delle random forest (RM) mostra invece rho elevati in tutti i test, mostrando una ottima affidabilità anche nel caso di dataset con meno di 100 punti, in linea con quanto riportato anche da Genuer et al., (2010), rendendo il metodo adatto anche ad analisi di dati eddy covariance a risoluzione mensile (che su una finestra temporale di 10-12 anni hanno massimo 150 valori).

4.3 SSA – segnali ottenuti dalla scomposizione delle variabili

La stagionalità influenza notevolmente gli andamenti di molte variabili ecologiche. La GPP è caratterizzata in genere da un forte segnale stagionale così come la temperatura e la radiazione, e tali segnali sono altamente correlati tra loro (Mahecha et al., 2007). Si è reso necessario quindi individuare ed eliminare il segnale stagionale all'interno delle variabili utilizzate nello studio. L'applicazione della SSA ha permesso di individuare le stagionalità della GPP e dei driver e quindi sottrarlo alla serie.

Tutti i siti, a diverse latitudini ed ecosistemi, presentano una marcata stagionalità della GPP. Alcuni driver non mostrano sempre segnali stagionali, in particolare le precipitazioni (P) e il contenuto d'acqua nel suolo (SWC), almeno in parte dei siti. Queste variabili infatti non sempre seguono chiari pattern stagionali e comunque non sempre correlati con la stagionalità della GPP.

Dal segnale destagionalizzato è stato poi estratto, se presente, il trend. La GPP non presenta un trend (o segnale di bassa frequenza) in sette siti. Anche in questo caso i siti sono molto diversi l'uno dall'altro. Probabilmente per la cropland FR-Gri e la grassland CH-Oe1 la gestione annuale del campo non permette la facile individuazione di un trend definito. Per gli altri siti che non presentano trend per la GPP, si può ipotizzare che realmente tale andamento di lungo periodo non sia presente. Infatti in ecosistemi in cui è stato raggiunto un equilibrio (come ad esempio foreste mature come DE-Hai, DE-Obe e NL-Loo) la GPP non presenta un trend.

Il driver per cui, nel maggior numero di siti, non è stata individuata una componente di lungo periodo è la radiazione. Questo era atteso in quanto in media la quantità di radiazione incidente al suolo non è soggetta a cambiamenti, tranne nel caso di effetti indiretti da parte dell'atmosfera. SW-IN mostra un trend solo in 15 siti, la maggior parte sono siti forestali. Tra questi sono presenti i siti tropicali (ZA-Kru, GF-Guy, AU-Tum, AU-DaS) dove l'effetto potrebbe essere dovuto ad un cambio nella dinamica della copertura nuvolosa. Anche per gli altri siti possibili cambiamenti nell'atmosfera (aerosol ad esempio, come individuato da Niyogi et al., 2004) potrebbero esserne la

causa. Tuttavia non è da escludere che parte del trend sia dovuta semplicemente a problematiche strumentali, essendo in generale i radiometri usati in campo soggetti a lento deperimento se non ricalibrati con frequenza.

Tutti gli altri driver mostrano andamenti di lungo periodo per la maggior parte dei siti. Il SWC è la variabile che ha un trend definito in un più alto numero di siti (34 su 38) e questo è in linea con quanto evidenziato da altri studi a scala globale (Jung et al., 2010). I siti che non presentano il trend nell'SWC sono due GRA (IT-Tor e NL-Hor), una CRO (FR-Gri) e una OSH (CA-SF3). Per la cropland la spiegazione è da ricercare nel tipo di gestione (irrigazione), mentre per gli altri siti le basse temperature (IT-Tor, sulle Alpi e CA-SF3) o la disponibilità continua di acqua (NL-Hor) possono spiegare l'assenza di trend.

I residui sono stati analizzati in base alla deviazione standard per verificare se vi fossero scostamenti evidenti dal valore medio tra i siti. In generale i residui dei siti si presentano omogenei con deviazioni standard prossime al valore medio. Solo per RU-Fyo i residui del SWC hanno uno scostamento significativamente elevato, probabilmente dovuto al cambio nei sensori durante le misure.

In generale la SSA si è dimostrato uno strumento utile, in grado di estrarre le componenti di un segnale in modo efficace per la maggior parte dei casi come affermato già da Mahecha et al. (2007). Questo ha permesso l'analisi del ruolo delle variabili sulla GPP al netto del forte segnale stagionale. Probabilmente serie temporali più lunghe, una maggiore qualità delle misure ed un parametrizzazione della SSA sito specifica potrebbe portare a risultati ancora più puliti e chiari.

4.3. Performance ANN e RF addestrate con le componenti estratte dalle variabili

Una volta scomposte le variabili nei loro sottosegnali (stagionalità, trend e residui), questi sono stati usati nella costruzione dei modelli ANN e RF su cui si basano i metodi dell'importanza dei driver. Sui modelli creati le componenti delle variabili, sono stati calcolati i relativi indici di determinazione (R^2). Sia per le ANN che per le RF addestrate con tutte le componenti dei driver, mantenendo come output la GPP non decomposta, le performance del fitting risultano molto alte, dimostrando la capacità dei modelli di fittare correttamente i dati.

Anche le ANN e le RF addestrate con le componenti di lungo periodo dei driver hanno performance molto alte, riuscendo a fittare ottimamente il segnale di lungo periodo della GPP.

Queste buone performances erano comunque attese in quanto i trend sono funzioni abbastanza semplici da fittare, quasi lineare, sia per i drivers che per l'output.

Per quanto riguarda le reti e le random forest addestrate utilizzando i residui, presentano R^2 sul fitting differenti. La ANN mostrano coefficienti di determinazioni inferiore alle RF, ma comunque in media superiori a 0.6, riuscendo a fittare più del 60% dei dati. Mentre le RF hanno R^2 più alti, fittando il 90% dei dati relativi a residui della GPP. I residui sono segnali a frequenza molto elevata ed includono le componenti di rumore (tipiche delle misure) e le ANN mostrano più difficoltà nel fittare i dati, mentre le RF risultano più flessibili.

4.4.Importanza dei driver

I tre metodi selezionati tramite il dataset artificiale sono stati applicati sia alle variabili non decomposte, sia all'insieme delle componenti, per confrontare quale componente dei driver guidasse il segnale non decomposto della GPP. Le variabili con una forte stagionalità (come SW-IN e TA) sono state individuate come le più importanti. Analizzando però i risultati dei metodi applicati a tutte le componenti insieme, è evidente come i segnali stagionali siano significativamente più importanti rispetto ai residui e ai trend. Le stagionalità infatti mascherano le interazioni che i driver hanno con la GPP nel breve (residui) e nel lungo periodo (trend) (Mahecha et al., 2007).

Si è quindi proceduto ad applicare i metodi alle componenti residuali per analizzare quale influenzasse di più la GPP nel breve periodo, e ai trend per verificare quale driver condizionasse maggiormente la GPP nel lungo periodo.

Le importanze dei driver restituite dai tre metodi (VRI) sono state mediate e pesate per le performances generali del modello (VAI), per aver un unico valore da confrontare anche con le caratteristiche dei siti (climatiche e vegetazionali). Infatti per comprendere meglio i risultati ottenuti dall'applicazione dei metodi alle componenti dei drivers, sono state prese in considerazione diverse caratteristiche dei siti: la temperatura media annua, le precipitazioni, la latitudine e l'altitudine.

Le caratteristiche dei siti risultano indipendenti l'una dall'altra, ad eccezione della temperatura media annua e della latitudine. Anche le componenti dei driver sembrano non avere alcuna relazione. Solo TA e SW-IN hanno un andamento simile nei residui. Ciò è importante da tenere in considerazione, in quanto sia le reti neurali che le random forest, di base, sono modelli che devono essere addestrati con variabili indipendenti l'una dall'altra, per non dare informazioni ridondanti che potrebbero causare overfitting.

In generale la variabile che guida le variazioni ad alta frequenza (residui) della GPP è risultata la SW-IN. La SW-IN è infatti una variabile che ha un effetto immediato sulla fotosintesi e quindi la GPP ed è normale che sia quella maggiormente correlata alle fluttuazioni giornaliere (in accordo con diversi studi quali Moffat et al., 2010, van Gorsel et al., 2013 e Runkle et al., 2014). Le uniche eccezioni sono state quattro siti: DK-ZaH, US-SRM, US-Whs e US-Wkg. Il sito DK-ZaH è in Groenlandia ed ha come driver primario il SWC, mentre la SW-IN non sembra avere un ruolo nel guidare la GPP. Questo può essere spiegato dal clima molto rigido, che durante l'anno porta a ghiacciare completamente il terreno, impedendo alle piante presenti di accedere alle risorse idriche che diventano quindi il fattore limitante nel compiere la fotosintesi anche nel breve periodo.

I tre siti americani sono invece caratterizzati da un clima arido, con temperature elevate e scarse precipitazioni. Anche per questi siti il fattore limitante sembra essere la disponibilità d'acqua, infatti mostrano come driver primario della fotosintesi le precipitazioni. A conferma di questa osservazione, diversi studi riportano lo stesso risultato per i siti aridi (Garbulsky et al., 2010, Medlyn et al., 2005).

Per i siti forestali il VAI relativo alla SW-IN ha valori in media più alti, probabilmente perché nelle colture agrarie e nelle grassland entrano in gioco anche altri fattori dovuti alla gestione che condizionano parte della variabilità dei residui della GPP. Inoltre si è notata una leggera correlazione positiva tra i siti con alte precipitazioni annuali e l'importanza della SW-IN, forse dovuta a una maggiore nuvolosità, che influisce sulla radiazione, aumentando la componente relativa alla radiazione diffusa e migliorando l'efficienza fotosintetica delle piante, come mostrato anche da van Gorsel et al. (2013).

Per quanto riguarda i driver del trend della GPP, i risultati sono più diversificati, mostrando risposte differenti tra i siti forestali e quelli non forestali (come osservato anche da Garbulsky et al. (2010).

Per quanto riguarda i siti non forestali, il VPD viene selezionato come driver nei climi più freddi (DK-ZaH e CA-SF3) e in quasi tutti i siti montani. È evidente quindi un legame tra il trend del VPD e l'andamento sul lungo periodo del GPP nei siti caratterizzati da inverni rigidi. Il forte freddo prolungato probabilmente influenza, nei siti senza vegetazione arborea, la traspirazione riducendo l'apertura stomatica e condizionando la fotosintesi sul lungo periodo come già osservato da Runkle et al. (2014) per una peatland a nord della Russia. E' anche importante rimarcare che il VPD è una variabile che è influenzata anche dalla temperatura che quindi potrebbe avere un effetto indiretto nella selezione della variabile guida.

Nei siti più aridi non forestali, la disponibilità d'acqua, rappresentata dal SWC e dalle precipitazioni, domina la fotosintesi anche sul lungo periodo. Ma P è anche il driver principale della GPP nei siti temperati. In assenza quindi di particolari condizioni climatiche, le precipitazioni sembrano essere il driver che maggiormente guida il trend della GPP nei siti non forestali (Garbulsky et al., 2010; Huxman et al., 2004).

Per quanto concerne i siti forestali, i VPD viene selezionato come variabile primaria sul lungo periodo solo nella foresta pluviale (GF-Guy). Ciò potrebbe essere però l'effetto di un elevato tasso di traspirazione per questa tipologia vegetazionale, che essendo strettamente correlata alla fotosintesi la potrebbe far figurare come driver importante. Il sito forestale più freddo (CA-Qfo), ha come driver principale la TS, che potrebbe essere spiegato dalle temperature invernali molto rigide (come osservato per altri boschi caratterizzati da climi freddi anche da Garbulsky et al. (2010).

In tre siti freddi di conifere, FI-Hyy, IT-Lav e RU-Fyo, le precipitazioni guidano il trend della fotosintesi, facendo ipotizzare una dipendenza della fotosintesi dall'andamento pluviometrico di lungo periodo nelle ENF tipiche di elevate altitudini o latitudini (risultati confrontabili con quelli ottenuti da Garbulsky et al., 2010 e Huxman et al., 2004). Infatti anche il sito montano US-Blo, ha come uno dei driver principali le precipitazioni insieme alla temperatura del suolo.

I boschi di latifoglie a latitudini e altitudine medie (BE-Vie, US-MMS, AU-Tum e IT-Ro1), anche se caratterizzati da diverse tipologie climatiche e vegetazionali, mostrano come driver primario è il SWC. Le latifoglie infatti sono molto esigenti da un punto di vista idrico e la disponibilità d'acqua nel suolo gioca un ruolo fondamentale per la fotosintesi nel lungo periodo, come riscontrato anche negli studi di Verbeeck, Samson, Granier, Montpied, & Lemeur (2008), Xie et al. (2014). Per alcuni siti forestali con caratteristiche climatiche differenti (DK-Sor, DE-Tha, IT-Ren, IT-SRo, NL-Loo) viene selezionata come driver primario la radiazione. Nonostante la SW-IN, in assenza di altri fattori limitanti, giochi un ruolo importante anche sul lungo periodo (Law et al., 2002) in questo caso è più probabile che si tratti di una selezione dovuta ad un trend strumentale nella variabile o più generalmente ad una selezione casuale in sito con trend poco pronunciato. L'unico sito che ha una netta predominanza delle temperature come driver della GPP di lungo periodo è US-UMd. In questo sito insistono prevalentemente pioppi e betulle, che evidentemente risultano più sensibili delle altre specie, per la fotosintesi, alle fluttuazioni di lungo periodo delle temperature..

5. Conclusioni

Le ANN e le RF addestrate con i segnali originali delle variabili, con tutte le componenti, o con i trend mostrano buone performance confermando l'efficacia di questi modelli nel riuscire a spiegare, nella maggior parte dei casi, fenomeni ecologici complessi.

Le diverse metodologie testate in questo studio sono validi strumenti in grado di restituire l'importanza dei driver, riuscendo a discriminare i driver che guidano un certo fenomeno dalle variabili che invece non hanno alcun peso nella determinazione dell'output.

Solo tre metodi, tra quelli testati, sono però efficienti nel riportare la corretta gerarchia dei driver nello spiegare un dato fenomeno (FM, DM e RM), reagendo bene anche nel caso di variabili ridondanti o dati rumorosi.

Qualche incertezza hanno mostrato i metodi basati sulle reti neurali (FM e DM) per quanto riguarda le serie con pochi dati a disposizione. Mentre il metodo delle random forest (RM) mostra un'ottima affidabilità anche nel caso di dataset con meno di 100 punti.

La stagionalità influenza notevolmente gli andamenti di molte variabili ecologiche e risulta difficile comprendere quali siano i driver che condizionano la GPP nel breve e nel lungo periodo. Questo è un aspetto molto importante da tenere in considerazione ogni volta che si cerca di analizzare le relazioni driver-output o anche quando misure e modelli vengono confrontati: il segnale stagionale è un segnale forte che viene facilmente spiegato da differenti variabili. La SSA è uno strumento utile per discriminare le varie componenti di un segnale e grazie all'applicazione di questa metodologia è stato possibile eliminare il segnale stagionale dalle variabili. In alcuni casi sarebbe opportuno provare diversi settaggi dei parametri per riuscire a estrarre tutte le componenti desiderate in modo efficiente.

Dall'applicazione dei metodi per l'importanza dei driver alle componenti residuali, la variabile più importante risulta essere l'SW-IN. Infatti in assenza di particolari condizioni climatiche limitanti, la radiazione è indubbiamente il driver principale della fotosintesi nel breve periodo a prescindere dal tipo di vegetazione e ecosistema. Negli ecosistemi più aridi, la GPP dipende maggiormente dalla disponibilità idrica, mettendo in secondo piano la radiazione, senza differenze tra tipi vegetazionali.

Per quanto riguarda l'andamento di lungo periodo della GPP, per i siti non forestali, il VPD viene selezionato come driver principale nei climi più freddi. È evidente quindi un legame tra il trend del VPD e quello della GPP nei siti caratterizzati da inverni rigidi senza vegetazione arborea

dove l'effetto di un alto VPD riduce l'evapotraspirazione e conseguentemente condiziona la fotosintesi.

SWC e P guidano l'andamento della GPP nel lungo periodo sia nei siti non forestali che in quelli forestali di latifoglie distribuiti lungo diversi ecosistemi (siti aridi, subtropicali e temperati). mentre la temperatura del suolo influenza la GPP per i siti forestali più freddi

In generale in questo studio è emerso, sia nel breve che nel lungo periodo, una maggiore dipendenza della fotosintesi alla disponibilità d'acqua, piuttosto che al tipo di vegetazione e alle temperature. Infatti le P e il SWC sono i driver principali sia per i residui che per il trend in tutti i siti aridi e semiaridi. Mentre, in assenza di fattori limitanti, la radiazione è il driver ecologico che influisce maggiormente sulla GPP nel breve periodo e per i siti forestali temperati di conifere anche sul lungo periodo.

Questo studio conferma l'importanza di misure ecologiche continue che rappresentano un utile strumento per comprendere fenomeni ecologici complessi a diverse scale temporali. Una maggiore lunghezza dei dataset è infatti fondamentale per individuare fenomeni che si verificano su diversi anni. Ad oggi sono ancora pochi i siti eddy covariance con serie di dati che superano i 10 anni. È importante quindi continuare le misure, per riuscire ad avere più siti e più dati di lungo periodo.

Infine per migliorare lo studio e indagare in modo più approfondito i legami tra driver e fotosintesi, sarebbe necessario aggiungere altre variabili, soprattutto legati all'andamento della stagione vegetativa, alla disponibilità di nutrienti (*N deposition* per esempio) e alla gestione dei diversi siti (ad esempio tagli del bosco e rotazione delle colture). I modelli non sembrano rappresentare un aspetto limitante per comprendere i fenomeni ecologici indagati.

Bibliografia

- Aubinet, M., Chermanne, B., Vandenhaute, M., Longdoz, B., Yernaux, M., & Laitat, E. (2001). Long term carbon dioxide exchange above a mixed forest in the Belgian Ardennes. *Agricultural and Forest Meteorology*, 108(4), 293–315. [http://doi.org/10.1016/S0168-1923\(01\)00244-1](http://doi.org/10.1016/S0168-1923(01)00244-1)
- Aubinet, M., Vesala, T., & Papale, D. (2012). *Eddy Covariance. A Practical Guide to Measurement and Data Analysis*. <http://doi.org/10.1007/978-94-007-2351-1>
- Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4), 479–492. <http://doi.org/10.1046/j.1365-2486.2003.00629.x>
- Boelman, N. T., Stieglitz, M., Griffin, K. L., & Shaver, G. R. (2005). Inter-annual variability of NDVI in response to long-term warming and fertilization in wet sedge and tussock tundra. *Oecologia*, 143(4), 588–597. <http://doi.org/10.1007/s00442-005-0012-9>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <http://doi.org/10.1023/A:1010933404324>
- Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data, 200, 217–236.
- Brosse, S., Lek, S., & Dauba, F. (1999). Predicting fish distribution in a mesotrophic lake by hydroacoustic survey and artificial neural networks. *Limnology and Oceanography*, 44(5), 1293–1303. <http://doi.org/10.4319/lo.1999.44.5.1293>
- Buttlar, J. V. (2015). Package “spectral.methods,” 29.
- Clark, K. L., Gholz, H. L., & Castro, M. S. (2004). Carbon dynamics along a chronosequence of slash pine plantations in North Florida. *Ecological Applications*, 14(4), 1154–1171. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-4444259609&partnerID=tZOtx3y1>
- Desai, A. R., Richardson, A. D., Moffat, A. M., Kattge, J., Hollinger, D. Y., Barr, A., ... Stauch, V. J. (2008). Cross-site evaluation of eddy covariance GPP and RE decomposition techniques. *Agricultural and Forest Meteorology*, 148(6-7), 821–838.

<http://doi.org/10.1016/j.agrformet.2007.11.012>

- Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 3. <http://doi.org/10.1186/1471-2105-7-3>
- Dietterich, T. G. (2000). Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, 40(2), 139–157. <http://doi.org/10.1023/A:1007607513941>
- Dolman, A. J., Moors, E. J., & Elbers, J. A. (2002). The carbon uptake of a mid latitude pine forest growing on sandy soil. *Agricultural and Forest Meteorology*, 111(3), 157–170. [http://doi.org/10.1016/S0168-1923\(02\)00024-2](http://doi.org/10.1016/S0168-1923(02)00024-2)
- Falge, E., Baldocchi, D., Tenhunen, J., Aubinet, M., Bakwin, P., Berbigier, P., ... Wofsy, S. (2002). Seasonality of ecosystem respiration and gross primary production as derived from FLUXNET measurements. *Agricultural and Forest Meteorology*, 113(1-4), 53–74. [http://doi.org/10.1016/S0168-1923\(02\)00102-8](http://doi.org/10.1016/S0168-1923(02)00102-8)
- Gallo, V. (2007). Surprising synapses deep in the brain. *Nature Neuroscience*, 10(3), 267–9. <http://doi.org/10.1038/nn0307-267>
- Garbulsky, M. F., Peñuelas, J., Papale, D., Ardö, J., Goulden, M. L., Kiely, G., ... Filella, I. (2010). Patterns and controls of the variability of radiation use efficiency and primary productivity across terrestrial ecosystems. *Global Ecology and Biogeography*, 19(2), 253–267. <http://doi.org/10.1111/j.1466-8238.2009.00504.x>
- GARBULSKY, M. F., PEÑUELAS, J., PAPALE, D., & FILELLA, I. (2008). Remote estimation of carbon dioxide uptake by a Mediterranean forest. *Global Change Biology*, 14(12), 2860–2867. <http://doi.org/10.1111/j.1365-2486.2008.01684.x>
- Garson, G. D. (1991). A Comparison of Neural Network and Expert Systems Algorithms with Common Multivariate Procedures for Analysis of Social Science Data. *Social Science Computer Review*, 9(3), 399–434.
- Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236. <http://doi.org/10.1016/j.patrec.2010.03.014>
- Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the

contribution of variables in artificial neural network models. In *Ecological Modelling* (Vol. 160, pp. 249–264). [http://doi.org/10.1016/S0304-3800\(02\)00257-0](http://doi.org/10.1016/S0304-3800(02)00257-0)

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., ... Yiou, P. (2002). Advanced Spectral Methods for Climatic Time Series. *Reviews of Geophysics*, 40(1), 1003. <http://doi.org/10.1029/2000RG000092>

Golyandina, N., & Korobeynikov, A. (n.d.). Basic Singular Spectrum Analysis and Forecasting with R.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4), 308–319. <http://doi.org/10.1198/tast.2009.08199>

Huxman, T. E., Smith, M. D., Fay, P. A., Knapp, A. K., Shaw, M. R., Loik, M. E., ... Williams, D. G. (2004). Convergence across biomes to a common rain-use efficiency. *Nature*, 429(6992), 651–4. <http://doi.org/10.1038/nature02561>

IPCC. (2007). IPCC Fourth Assessment Report (AR4). *IPCC*, 1, 976. <http://doi.org/ISSN:02767783>

Janssens, I., & Pilegard, K. (2003). Large seasonal changes in Q₁₀ of soil respiration in a beech. *Global Biogeochemical Cycles*, 9(6), 911–918. <http://doi.org/10.1046/j.1365-2486.2003.00636.x>

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., ... Zhang, K. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, 467(7318), 951–954. <http://doi.org/10.1038/nature09396>

Kato, T., & Tang, Y. (2008). Spatial variability and major controlling factors of CO₂ sink strength in Asian terrestrial ecosystems: Evidence from eddy covariance data. *Global Change Biology*, 14(10), 2333–2348. <http://doi.org/10.1111/j.1365-2486.2008.01646.x>

Korobeynikov, A. A., Shlemov, A., Usevich, K., Golyandina, N., & Korobeynikov, M. A. (2015). *Package “Rssa”*.

Laë, R., Lek, S., & Moreau, J. (1999). Predicting fish yield of African lakes using neural networks. *Ecological Modelling*, 120(2-3), 325–335. [http://doi.org/10.1016/S0304-3800\(99\)00112-X](http://doi.org/10.1016/S0304-3800(99)00112-X)

- Lasslop, G., Reichstein, M., Papale, D., Richardson, A., Arneth, A., Barr, A., ... Wohlfahrt, G. (2010a). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. *Global Change Biology*, 16(1), 187–208. <http://doi.org/10.1111/j.1365-2486.2009.02041.x>
- Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., ... Wohlfahrt, G. (2010b). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, 16(1), 187–208. <http://doi.org/10.1111/j.1365-2486.2009.02041.x>
- Law, B. ., Falge, E., Gu, L., Baldocchi, D. ., Bakwin, P., Berbigier, P., ... Wofsy, S. (2002). Environmental controls over carbon dioxide and water vapor exchange of terrestrial vegetation. *Agricultural and Forest Meteorology*, 113(1-4), 97–120. [http://doi.org/10.1016/S0168-1923\(02\)00104-1](http://doi.org/10.1016/S0168-1923(02)00104-1)
- Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, 90(1), 39–52. [http://doi.org/10.1016/0304-3800\(95\)00142-5](http://doi.org/10.1016/0304-3800(95)00142-5)
- Lek, S., & Gu??gan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120(2-3), 65–73. [http://doi.org/10.1016/S0304-3800\(99\)00092-7](http://doi.org/10.1016/S0304-3800(99)00092-7)
- Lieth, H., & Whittaker, R. H. (1975). Primary Production of the Major Vegetation Units of the World. In *Primary Productivity of the Biosphere* (pp. 203–215). <http://doi.org/10.1007/978-3-642-80913-2>
- Liong, S. Y., Lim, W. H., Kojiri, T., & Hori, T. (2000). Advance flood forecasting for flood stricken Bangladesh with a fuzzy reasoning method. *Hydrological Processes*, 14(3), 431–448. [http://doi.org/10.1002/\(SICI\)1099-1085\(20000228\)14:3<431::AID-HYP947>3.0.CO;2-0](http://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<431::AID-HYP947>3.0.CO;2-0)
- Luyssaert, S., Janssens, I. a., Sulkava, M., Papale, D., Dolman, a. J., Reichstein, M., ... Moors, E. J. (2007). Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10), 2110–2127. <http://doi.org/10.1111/j.1365-2486.2007.01432.x>
- Mahecha, M. D., Reichstein, M., Lange, H., Carvalhais, N., Bernhofer, C., Grünwald, T., ...

- Seufert, G. (2007). Characterizing ecosystem-atmosphere interactions from short to interannual time scales. *Biogeosciences*, 4(5), 743–758. <http://doi.org/10.5194/bg-4-743-2007>
- Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling and Software*, 15(1), 101–124. [http://doi.org/10.1016/S1364-8152\(99\)00007-9](http://doi.org/10.1016/S1364-8152(99)00007-9)
- Mastrorillo, S., Dauba, F., Oberdorff, T., Guégan, J.-F., & Lek, S. (1998). Predicting local fish species richness in the garonne river basin. *Comptes Rendus de l'Académie Des Sciences - Series III - Sciences de La Vie*, 321(5), 423–428. [http://doi.org/10.1016/s0764-4469\(98\)80307-7](http://doi.org/10.1016/s0764-4469(98)80307-7)
- Medlyn, B. E., Berbigier, P., Clement, R., Grelle, A., Loustau, D., Linder, S., ... McMurtrie, R. E. (2005). Carbon balance of coniferous forests growing in contrasting climates: Model-based analysis. *Agricultural and Forest Meteorology*, 131(1-2), 97–124. <http://doi.org/10.1016/j.agrformet.2005.05.004>
- Moffat, A. M., Beckstein, C., Churkina, G., Mund, M., & Heimann, M. (2010). Characterization of ecosystem responses to climatic controls using artificial neural networks. *Global Change Biology*, 16(10), 2737–2749. <http://doi.org/10.1111/j.1365-2486.2010.02171.x>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [http://doi.org/10.1016/0022-1694\(70\)90255-6](http://doi.org/10.1016/0022-1694(70)90255-6)
- Niyogi, D., Chang, H. I., Saxena, V. K., Holt, T., Alapaty, K., Booker, F., ... Xue, Y. (2004). Direct observations of the effects of aerosol loading on net ecosystem CO₂ exchanges over different landscapes. *Geophysical Research Letters*, 31(20). <http://doi.org/10.1029/2004GL020915>
- Olden, J. D., & Jackson, D. A. (2002). Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2), 135–150. [http://doi.org/10.1016/S0304-3800\(02\)00064-9](http://doi.org/10.1016/S0304-3800(02)00064-9)
- Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4), 389–397. <http://doi.org/10.1016/j.ecolmodel.2004.03.013>

- Paruelo, J., & Tomasel, F. (1997). Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. *Ecological Modelling*. [http://doi.org/10.1016/S0304-3800\(96\)01913-8](http://doi.org/10.1016/S0304-3800(96)01913-8)
- Recknagel, F. (1997). ANNA - Artificial neural network model for predicting species abundance and succession of blue-green algae. In *Hydrobiologia* (Vol. 349, pp. 47–57). <http://doi.org/10.1023/A:1003041427672>
- Reichstein, M., Ciais, P., Papale, D., Valentini, R., Running, S., Viovy, N., ... Zhao, M. (2007). Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: A joint flux tower, remote sensing and modelling analysis. *Global Change Biology*, 13(3), 634–651. <http://doi.org/10.1111/j.1365-2486.2006.01224.x>
- Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., ... Valentini, R. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biology*. <http://doi.org/10.1111/j.1365-2486.2005.001002.x>
- Rosenzweig, M. L. (1968). Net primary productivity of terrestrial communities: Prediction from climatological data. *The American Naturalist*, 102(923), 67–74. <http://doi.org/10.1086/282523>
- Runkle, B. R. K., Wille, C., Gažovič, M., Wilmking, M., & Kutzbach, L. (2014). The surface energy balance and its drivers in a boreal peatland fen of northwestern Russia. *Journal of Hydrology*, 511, 359–373. <http://doi.org/10.1016/j.jhydrol.2014.01.056>
- Sims, D. A., Luo, H., Hastings, S., Oechel, W. C., Rahman, A. F., & Gamon, J. A. (2006). Parallel adjustments in vegetation greenness and ecosystem CO₂ exchange in response to drought in a Southern California chaparral ecosystem. *Remote Sensing of Environment*, 103(3), 289–303. <http://doi.org/10.1016/j.rse.2005.01.020>
- Smith, P. (2004). Carbon sequestration in croplands: The potential in Europe and the global context. *European Journal of Agronomy*. <http://doi.org/10.1016/j.eja.2003.08.002>
- Valentini, R., Matteucci, G., Dolman, a J., Schulze, E. D., Rebmann, C., Moors, E. J., ... Clement, R. (2000). Respiration as the main determinant of carbon balance in European forests. *Nature*, 404(6780), 861–865. <http://doi.org/10.1038/35009084>

- van Gorsel, E., Berni, J. a. J., Briggs, P., Cabello-Leblic, a., Chasmer, L., Cleugh, H. a., ... Zegelin, S. (2013). Primary and secondary effects of climate variability on net ecosystem carbon exchange in an evergreen Eucalyptus forest. *Agricultural and Forest Meteorology*, 182-183, 248–256. <http://doi.org/10.1016/j.agrformet.2013.04.027>
- Vautard, R., Yiou, P., & Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4), 95–126. [http://doi.org/10.1016/0167-2789\(92\)90103-T](http://doi.org/10.1016/0167-2789(92)90103-T)
- Verbeeck, H., Samson, R., Granier, A., Montpied, P., & Lemeur, R. (2008). Multi-year model analysis of GPP in a temperate beech forest in France. *Ecological Modelling*, 210(1-2), 85–103. <http://doi.org/10.1016/j.ecolmodel.2007.07.010>
- Waring, R. H., Law, B. E., Goulden, M. L., Bassow, S. L., McCreight, R. W., Wofsy, S. C., & Bazzaz, F. A. (1995). Scaling gross ecosystem production at Harvard Forest with remote sensing: A comparison of estimates from a constrained quantum-use efficiency model and eddy correlation. *Plant, Cell and Environment*, 18(10), 1201–1213. <http://doi.org/10.1111/j.1365-3040.1995.tb00629.x>
- Xie, J., Chen, J., Sun, G., Chu, H., Noormets, A., Ouyang, Z., ... Guan, W. (2014). Long-term variability and environmental control of the carbon cycle in an oak-dominated temperate forest. *Forest Ecology and Management*, 313, 319–328. <http://doi.org/10.1016/j.foreco.2013.10.032>
- Yiou, P., Sornette, D., & Ghil, M. (2000). Data-adaptive wavelets and multi-scale singular-spectrum analysis. *Physica D: Nonlinear Phenomena*, 142(3-4), 254–290. [http://doi.org/10.1016/S0167-2789\(00\)00045-2](http://doi.org/10.1016/S0167-2789(00)00045-2)