Nikola Jajcay; Milan Paluš

Štatistické modelovanie javu El Niño - Južná oscilácia v klimatológii

*Pokroky matematiky, fyziky a astronomie*, Vol. 62 (2017), No. 1, 52–70

Persistent URL: http://dml.cz/dmlcz/146723

# Štatistické modelovanie javu El Niño — Južná oscilácia v klimatológii

*Nikola Jajcay, Milan Paluš, Praha*

*Abstrakt.* Pri modelovaní v klimatológii a meteorológii rozlišujeme dva základné druhy modelov — dynamické a štatistické. Dynamické modely majú fyzikálny základ, ktorý pozostáva z diskretizovaných diferenciálnych rovníc a súčasného stavu ako počiatočnej podmienky a následne modelujú stav systému integrovaním týchto rovníc v čase. Štatistické modely sú už v základe odlišné: ich fungovanie sa nezakladá na fyzikálnych mechanizmoch tvoriacich dynamiku modelovaného systému, ale sú odvodené z analýzy chodu počasia v minulosti. V tomto článku opíšeme príklad štatistického modelu, ktorý modeluje atmosféricko-oceánsky jav El Niño — Southern Oscillation. Zvýšenú pozornosť venujeme modelovaniu nelineárnych medziškálových interakcií. Okrem štatistických vlastností modelu sa tiež zaoberáme parametrizáciami šumu. Taktiež zvažujeme možnosť použitia štatistických modelov nízkej komplexity ako surogátnych modelov na generovanie dát za účelom štatistického testovania hypotéz.

## 1. Modelovanie v atmosférických vedách

Klimatické a atmosférické modely, ktoré používajú kvantitatívne metódy na simulovanie interakcií v klimatickom systéme, sú jedným z najdôležitejších nástrojov pre predikciu a porovnanie klímy v budúcnosti, alebo štúdium klímy v minulosti. Takmer každý sa denne stretáva s použitím dynamických modelov v predpovedi počasia. Vo všeobecnosti sa používajú dva hlavné typy modelov: dynamické a štatistické. Princíp dynamického modelovania spočíva v použití diferenciálnych rovníc (udávajúcich vzťah medzi časovým vývojom rôznych veličín), ktoré sú numericky integrované v čase z tzv. počiatočného stavu. Počiatočný stav reprezentuje východzie podmienky pri integrácii. Asi najznámejším príkladom použitia dynamických modelov sú globálne klimatické modely (GCM — *general circulation model*). Sú to v princípe matematické modely cirkulácie atmosféry a oceánov v planetárnom merítku, ktoré na modelovanie používajú Navier-Stokesove rovnice na rotujúcej sfére (reprezentujúce zachovanie hybnosti v kvapaline), rovnicu kontinuity (zachovanie hmoty v kvapaline) a iné základné rovnice dynamiky a statiky tekutín spolu s termodynamickými členmi na vyjadrenie energetických zdrojov a prepadov. Takto zostrojené modely sa používajú v numerickej

Mgr. Nikola Jajcay, Oddělení nelineární dynamiky a složitých systémů, Ústav informatiky AV ČR, v. v. i., Pod Vodárenskou věží 2, 182 07 Praha 8, Katedra fyziky atmosféry, Matematicko-fyzikální fakulta, Univerzita Karlova, V Holešovičkách 2, 180 00 Praha 8, e-mail: jajcay@cs.cas.cz; RNDr. Milan Paluš, DrSc., Oddělení nelineární dynamiky a složitých systémů, Ústav informatiky AV ČR, v. v. i., Pod Vodárenskou věží 2, 182 07 Praha 8, e-mail: mp@cs.cas.cz

predpovedi počasia, na vytváranie dát (datasetov) minulého počasia, tzv. reanalýzy a taktiež na projekcie klímy do budúcnosti vo veľkých porovnávacích projektoch ako je CMIP3 [15] alebo CMIP5 [22].

Denne sa však stretávame s chybami pri modelovaní (koľkokrát už nevyšla predpoveď počasia?). Pri klimatickom modelovaní za pomoci dynamických modelov sa stretávame s dvoma druhmi chýb: prvý typ súvisí s počiatočnou chybou (chyby v inicializácii počiatočného stavu klimatického systému), druhý je priamo naviazaný na chyby modelu (rozlíšenie — či už časové alebo priestorové, chyby spojené s diskretizáciou diferenciálnych rovníc apod.) [2]. Problémy s chybami v počiatočnom stave sa väčšinou riešia rozšírením jednej predpovede (jednej integrácie) na súbor predpovedí s mierne sa líšiacimi počiatočnými podmienkami, kde sa na konci vyhodnocuje štatistika (napríklad priemer a rozptyl) daného súboru. S chybami spojenými priamo s modelom to už je zložitejšie — sú „vstavané“ do modelu prostredníctvom exponenciálneho rastu chyby (kde rast chyby znamená citlivosť na počiatočné podmienky — dve integrácie, ktoré sú v čase $t_0$ veľmi blízko pri sebe, sa v čase od seba môžu exponenciálne vzďaľovať). Táto vlastnosť súvisí s chaotickým správaním, ktoré je spojené s nelinearitami v diskretizovaných diferenciálnych rovniciach [13] a so systematickými chybami, kde vo všeobecnosti dochádza k posunu a deformáciam štatistických rozdelení simulovaných veličín. Práve toto je hlavný dôvod obmedzenia prediktability numerických predpovedných modelov na cca 6 až 10 dní (napr. [26]).

Druhým hlavným typom modelov, výrazne rozdielnym od dynamických, sú štatistické modely. Nie sú priamo založené na fyzikálnych mechanizmoch, ktoré ovladajú dynamiku modelovaného systému, ale sú odvodené z analýz chodu počasia v minulosti (v jazyku stále populárnejšieho strojového učenia by sme povedali, že sú „naučené“ na minulých dátach). Pravdepodobne najpoužívanejším konceptom v štatistickom modelovaní je inverzný stochastický model [18], pri ktorom sa najskôr model navrhne, potom natrénuje pomocou dát z minulosti a nakoniec sa stochasticky (v procese figuruje náhodná premenná — šum) integruje do budúcnosti. Tento typ modelu má samozrejme tiež svoje slabiny — musíme správne vybrať premenné tak, aby čo najvernejšie zachytili dynamiku systému, ktorý sa snažíme modelovať. Ďalším typickým problémom štatistických modelov je nestacionarita modelovaného systému. Keďže štatistický model sa nezakladá na fyzikálnych mechanizmoch zodpovedných za vývoj dynamiky, model, ktorý je natrénovaný na podmnožine dát, nemusí správne reprodukovať všetky možné stavy systému (matematicky môžeme povedať, že nemusí navštíviť všetky oblasti fázového priestoru). Stále však platí, že štatistické modely majú svoje veľké opodstatnenie, hlavne v prípade, kedy je dynamika systému neznáma a zostrojenie dynamického modelu takmer nemožné. Ďalšou výhodou štatistických modelov je skutočnosť, že užívateľ si môže prispôsobiť komplexitu modelu podľa potreby. S tým súvisí aj motivácia použiť štatistický model na štúdium systému s neznámou dynamikou — zostrojíme štatistický model a pokiaľ nám vie správne modelovať rôzne aspekty systému, tak je jednoduchšie študovať samotný model, v ktorom presne vieme, ktoré procesy — a ako konkrétne — sa podieľajú na dynamike študovaného systému. V nasledujúcej časti textu vybudujeme štatistický model javu El Niño — Southern Oscillation (ENSO) a špeciálne budeme klásť dôraz na parametrizáciu šumu.

## 2. Štatistický model ENSO

V tejto sekcii vybudujeme model na predikciu javu El Niño — Southern Oscillation. Spomedzi atmosférických a oceánskych javov s viac ako ročnou periódou vykazuje najsilnejší signál a má veľký socioekonomický dopad. Jeho najvýraznejším prejavom je nepravidelná oscilácia teploty povrchu Tichého oceánu v rovníkovej oblasti a s tým súvisiace zmeny atmosférickej a aj oceánskej cirkulácie.

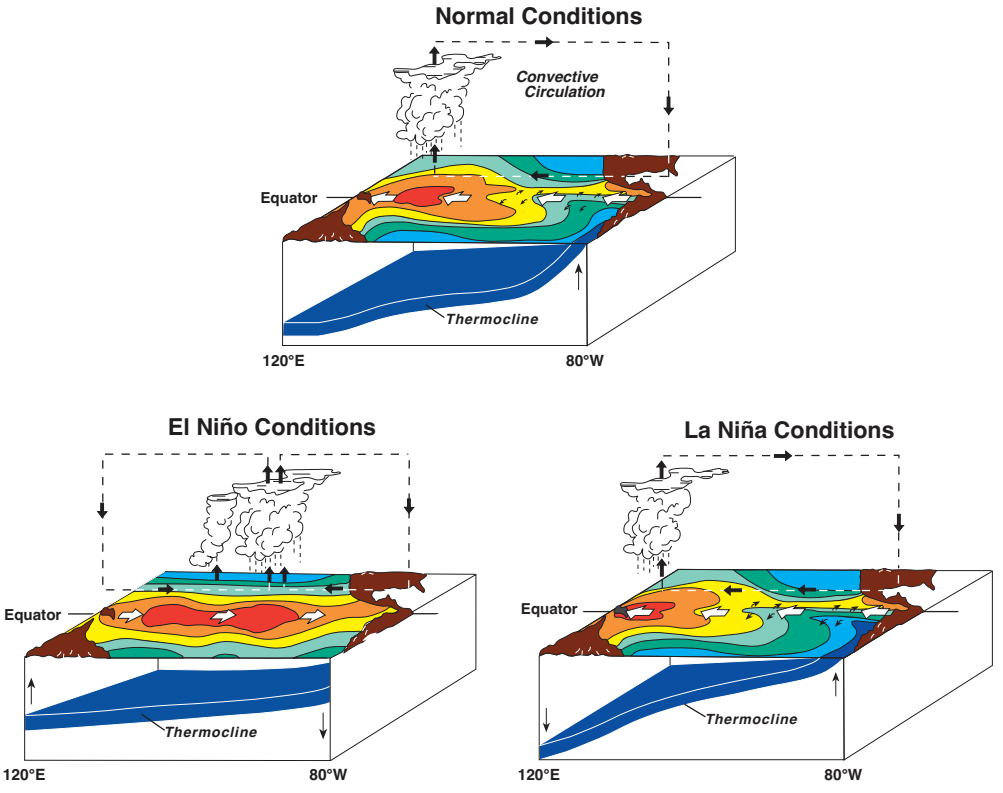
### 2.1. Fenomén ENSO

Fenomén ENSO a jeho tri fázy — neutrálna (normálne podmienky), teplá (El Niño) a studená (La Niña) — sú zobrazené na obr. 1. V neutrálnej fáze (obr. 1 hore), ktorá reprezentuje normálne podmienky v Pacifiku, tvorí základ teplý oceán v západnom Pacifiku (pri pobreží Austrálie) a studený oceán vo východnom Pacifiku (pri pobreží Peru). Vieme, že nad teplým oceánom (západný Pacifik) existuje silný výpar a keďže teplý vzduch je ľahší a stúpa nahor, v oblasti západného Pacifiku vzniká hlboká konvekcia a pásmo permanentných zrážok. S tým je spojené aj pásmo nižšieho tlaku nad západným Pacifikom. Tento teplý vzduch stúpa nahor, často až k hranici troposféry a následne vrchnou troposférou smeruje na východ, kde klesá k povrchu a tvorí pásmo vysokého tlaku (nad východným Pacifikom). Potom, v súlade s pasátmi, fúka po povrchu späť na západ a tým dokončuje slučku známu ako Walkerova cirkulácia. Otázkou ostáva, prečo je v západnom Pacifiku teplejšia voda ako vo východnom. Vysvetlenie súvisí s povrchovými vetrami, ktoré vejú smerom na západ, čiže na západe je „viac“ vody a termoklína (prechodová vrstva, v ktorej rýchlo klesá teplota s hĺbkou a pod ňou začína studený, hlboký oceán) sa musí tomuto efektu prispôsobiť naklonením. Vďaka pasátom vzniká tiež povrchový prúd v oceáne, ktorý smeruje k pólom na oboch hemisférach a tento úbytok vody sa vyrovnáva tzv. *upwellingom*, alebo pumpovaním vody na povrch oceánu, ktorý sa v rovníkovej oblasti deje z hĺbky cca 50 metrov. Termoklína je položená hlbšie na západe rovníkového Pacifiku, čiže voda vypumpovaná na povrch je stále teplá. Avšak na východe je termoklína položená v nižšej hĺbke ako je hranica upwellingu, čiže na povrch sa dostáva studená voda z tzv. hlbokého oceánu.

Pri teplej fáze ENSO (obr. 1, dole vľavo) sa teplejšia voda presúva smerom na východ k pobrežiu Peru, termoklína sa vyrovnáva a absencia studeného upwellingu ešte zvýrazňuje teplú anomáliu. S tým súvisí aj oslabenie Walkerovej cirkulácie a presun oblasti permanentných zrážok na východ, kde teraz pokrýva takmer celý rovníkový Pacifik.

Naopak, pri studenej fáze (obr. 1 dole vpravo) sa teplejšia voda posúva ešte viac na západ, až k pobrežiu Austrálie, termoklína je strmšia ako bežne a konvektívna cirkulácia nad Pacifikom je taktiež zosilnená. Za pôvodom ENSO síce stojí spojená dynamika atmosféry a oceánu v ekvatoriálnom (rovníkovom) Pacifiku [20], jeho efekty na všeobecnú cirkuláciu a interakciu medzi oceánom a pevninou sú však dobre badateľné aj mimo tropického pásu (oblasť medzi obratníkmi, čiže cca 23.5° severnej šírky až 23.5° južnej šírky), pretože sa prenášajú poväčšine za pomoci telekonekcií [1].

Obe extrémne fázy (teplá aj studená) môžu viesť k extrémnym klimatickým podmienkam v princípe na celej Zemi. Ako príklad nám poslúži priemerná globálna teplota, ktorá je dočasne vyššia počas teplej fázy ENSO a naopak dočasne nižšia počas studenej fázy ENSO [25]. Okrem globálnej teploty ovplyvňuje tento atmosféricko-oceánsky




Obr. 1. ENSO — normálne podmienky (hore), teplá fáza (dole vľavo) a studená fáza (dole vpravo). Obrázky prevzaté z [3].

fenomén samozrejme aj lokálnu teplotu (teplá fáza ENSO napríklad prináša zvýšenie teploty na stredozápade a východe USA, na juhu Brazílie a v celej juhovýchodnej Ázii a studenšie podmienky na Floride, kým studená fáza funguje presne opačne) a taktiež ovplyvňuje zrážky (suché podmienky v juhovýchodnej Ázii, Austrálii a v Indii počas teplej fázy ENSO a naopak zvýšené zrážky na tých istých miestach počas studenej fázy). Okrem už spomenutých javov súvisí teplá fáza ENSO tiež so slabšími vetrami okolo rovníka, so zvýšenou konvekciou pozdĺž ekvatoriálneho Pacifiku a zníženým rizikom hurikánov v Karibskej oblasti. Obrátene, počas studenej fázy ENSO môžeme počítať so silnejšími vetrami okolo rovníka, zníženou konvekciou a taktiež s vyššou pravdepodobnosťou výskytu hurikánov v Karibskej oblasti [6].

Dôležitým aspektom ENSO je skutočnosť, že teplá fáza — El Niño — sa vyznačuje väčšou magnitúdou (veľkosťou) ako jej náprotivok — La Niña. Táto štatistická nerovnosť čiastočne napovedá, že na dynamike ENSO sa podieľajú aj nelineárne procesy [5]. Napriek tomu aj najdetailnejšie numerické dynamické modely podceňujú tieto nelineárne efekty [8], čiže kvalita predpovede je stále nedostatočná. Ešte pred rokom 2000 bola väčšina modelov ENSO lineárna a nelineárne modely začali nadobúdať popularitu iba pomerne nedávno (viz napr. [24]).

### 2.2. Inverzný stochastický model

Koncept inverzného stochastického modelu použijeme ako štartovací bod pri stavaní modelu ENSO. Táto časť je prevažne prevzatá z [11]. Nech $\mathbf{x}(t)$ je stavový vektor anomálií nejakej veličiny, čiže $\mathbf{x}(t) = \mathbf{X}(t) - \overline{\mathbf{X}}(t)$, kde $\mathbf{X}(t)$ je klimatický stavový vektor danej veličiny a $\overline{\mathbf{X}}(t)$ je jeho klimatológia — časový priemer.

V našom prípade môže byť $\mathbf{X}(t)$ vektor teplôt povrchu oceánu vo vybraných bodoch v čase $t$. S ohľadom na veľké množstvo dát je však výhodnejšie časové rady teplôt najskôr transformovať metódou analýzy hlavných komponent (tzv. PCA rozklad). Táto slúži k dekorelácii dát a následným zanedbaním málo významných komponent môžeme znížiť dimenziu problému. Je to vlastne rozklad kovariančnej matice dát pomocou singulárnej dekompozície na priestorové komponenty nemenné v čase a k nim priradené časové rady.

Časová evolúcia anomálií $\mathbf{x}$ sa dá vyjadriť ako

$$\dot{\mathbf{x}} = \mathbf{L}\mathbf{x} + \mathbf{N}(\mathbf{x}), \tag{1}$$

kde $\mathbf{L}$ je lineárny operátor, $\mathbf{N}$ reprezentuje nelineárne zložky a bodkou značíme časovú deriváciu (explicitnú závislosť stavového vektora $\mathbf{x}(t)$ na čase už neuvádzame). Táto rovnica reprezentuje všeobecný inverzný model, o ktorom predpokladáme, že má lineárnu zložku reprezentovanú operátorom $\mathbf{L}$ a nelineárnu zložku reprezentovanú operátorom $\mathbf{N}$, ktorý závisí od stavového vektora $\mathbf{x}$. Tieto operátory vo všeobecnosti nezávisia od času.

Najjednoduchším typom inverzných modelov sú lineárne inverzné modely (*linear inverse models* — LIM [18]). Za pomoci predpokladu $\mathbf{N}(\mathbf{x})\mathrm{d}t \approx \mathbf{T}\mathbf{x}\,\mathrm{d}t + \mathrm{d}\mathbf{r}^{(0)}$, kde $\mathbf{T}$ je matica, ktorá reprezentuje lineárnu spätnú väzbu nerozlíšených (skrytých) procesov v $\mathbf{x}$, a $\mathrm{d}\mathbf{r}^{(0)}$ je šumový proces, zjednodušíme rovnicu na lineárnu. Šumový proces je pri odhadovaní parametrov rezíduum — zvyšok po odhadnutí, ktorý sa nedá vysvetliť lineárnym odhadom, a ako uvidíme neskôr, pri následnom integrovaní modelu sa z neho stane biely šum. Vďaka nášmu predpokladu môžeme rovnicu (1) prepísať ako

$$\mathrm{d}\mathbf{x} = \mathbf{B}^{(0)}\mathbf{x}\mathrm{d}t + \mathrm{d}\mathbf{r}^{(0)}, \quad \mathbf{B}^{(0)} = \mathbf{L} + \mathbf{T}. \tag{2}$$

Maticu $\mathbf{B}^{(0)}$, reprezentujúcu lineárne spätné väzby, a kovariančnú maticu šumu $\mathbf{Q} \equiv \langle \mathbf{r}^{(0)}\mathbf{r}^{(0)T} \rangle$, ktorú budeme potrebovať pri integrácii modelu, môžeme priamo odhadnúť z pozorovania veličiny $\mathbf{x}$ (z dát) pomocou viacnásobnej lineárnej regresie [28]. Samozrejme predpokladáme, že reziduály $\mathbf{r}^{(0)}$ majú nulovú strednú hodnotu. Stavový vektor $\mathbf{x}$ pozostáva z časových radov hlavných komponent a tzv. vektor odoziev $\dot{\mathbf{x}}$ pozostáva z ich tendencií.

### 2.3. Nelineárny viac-stupňový model

Predpoklady o lineárnej, stabilnej dynamike a o aditívnom bielom šume, ktoré sme použili pri predstavovaní lineárnych inverzných modelov, sú, bohužiaľ, platné iba pri príliš veľkej aproximácii. Konkrétne, rezíduum po lineárnom odhade $\mathrm{d}\mathbf{r}^{(0)}$ obsahuje v sebe autokorelácie, čo je v rozpore s definíciou šumového procesu — rezíduum má byť tzv. biely šum, čiže náhodný signál s rovnomerne rozdeleným spektrálnym výkonom. Okrem toho, matice $\mathbf{B}^{(0)}$ a $\mathbf{Q}$, ktoré sme získali odhadom z dát, vykazujú silnú závislosť

na samotnom procese $\mathbf{x}$ a z definície by mali byť nezávislé [19]. V nasledujúcom texte teda predstavíme dve modifikácie nášho modelu, ktoré adresujú nelinearitu a taktiež spôsob, ako sa zbaviť korelácií v v rezíduách (prevzaté z [11]).

Prvú modifikáciu dosiahneme uvažovaním polynomiálneho tvaru $\mathbf{N}(\mathbf{x})$ v rovnici (1), presnejšie kvadratického polynómu. $i$-tu komponentu $N_i(\mathbf{x})$ môžeme písať ako

$$N_i(\mathbf{x})\mathrm{d}x \approx \left(\mathbf{x}^T\mathbf{A}_i\mathbf{x} + \mathbf{t}_i\mathbf{x} + c_i^{(0)}\right)\mathrm{d}t + \mathrm{d}r_i^{(0)}, \tag{3}$$

kde matice $\mathbf{A}_i$ reprezentujú bloky tenzoru tretieho rádu, kým vektory $\mathbf{b}_i^{(0)} = \mathbf{l}_i + \mathbf{t}_i$ sú riadkami matice $\mathbf{B}^{(0)} = \mathbf{L} + \mathbf{T}$ (presne ako v rovnici (2)). Tieto matice, podobne ako zložky vektoru $\mathbf{c}^{(0)}$, môžeme odhadnúť pomocou viacnásobnej polynomiálnej regresie [14].

Druhá modifikácia sa vysporadúva s problémom korelácií v reziduálnom šume. Ak pri odhadovaní modelu zistíme, že rezíduá po odstránení kvadratickej a lineárnej zložky nie sú biely šum, tak tieto rezíduá odhadneme, teraz už len lineárne, pridaním ďalšieho stupňa do modelu. Konkrétne, $i$-tu zložku prvého, tzv. hlavného stupňa nášho inverzného stochastického modelu, môžeme písať ako

$$\mathrm{d}x_i = \left(\mathbf{x}^T\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{(0)}\mathbf{x} + c_i^{(0)}\right)\mathrm{d}t + \mathrm{d}r_i^{(0)}, \tag{4}$$

kde $\mathbf{x} = \{x_i\}$ je stavový vektor a matice $\mathbf{A}_i$, vektory $\mathbf{b}_i^{(0)}$, zložky $c_i^{(0)}$ vektoru $\mathbf{c}^{(0)}$, rovnako ako zložky $r_i^{(0)}$ vektoru reziduálneho šumu $\mathbf{r}^{(0)}$, zistíme pomocou metódy najmenších štvorcov. Ďalší stupeň modelu je pridaný na modelovanie známych časových inovácií (tendencií) $\mathrm{d}\mathbf{r}^{(0)}$ ako lineárnej funkcie tzv. rozšíreného stavového vektoru, $[\mathbf{x}, \mathbf{r}^{(0)}] \equiv \left(\mathbf{x}^T, \mathbf{r}^{(0)T}\right)^T$. Rezíduá tohto stupňa opäť zistíme pomocou metódy najmenších štvorcov. Rovnakým spôsobom pridáme ďalšie a ďalšie stupne modelu, až pokiaľ rezíduá $L$-tého stupňa, $\mathbf{r}^{(L+1)}$, nie sú biele v čase (vzájomne nezávislé a ich autokorelácia je nulová). Matematicky môžeme dodatočné stupne modelu písať ako

$$\begin{aligned}
\mathrm{d}r_i^{(0)} &= \mathbf{b}_i^{(1)}[\mathbf{x}, \mathbf{r}^{(0)}]\mathrm{d}t + r_i^{(1)}\mathrm{d}t, \\
\mathrm{d}r_i^{(1)} &= \mathbf{b}_i^{(2)}[\mathbf{x}, \mathbf{r}^{(0)}, \mathbf{r}^{(1)}]\mathrm{d}t + r_i^{(2)}\mathrm{d}t, \\
&\cdots \\
\mathrm{d}r_i^{(L)} &= \mathbf{b}_i^{(L+1)}[\mathbf{x}, \mathbf{r}^{(0)}, \ldots, \mathbf{r}^{(L)}]\mathrm{d}t + r_i^{(L+1)}\mathrm{d}t.
\end{aligned} \tag{5}$$

Rovnice (4) a (5) reprezentujú širokú paletu procesov, kde potrebujeme explicitne modelovať spätnú väzbu modelovaného procesu $\mathbf{x}$ na jeho šume. Lineárny viacstupňový model dostaneme predpokladom $\mathbf{A}_i \equiv 0$ a $\mathbf{c}^{(0)} \equiv 0$ v rovnici (4). Detaily metodológie a taktiež diskusia rôznych aspektov tohto modelu je uvedená v pôvodnom článku [12].

V konkrétnom prípade modelovania ENSO (následujúci princíp sa dá zovšeobecniť a použiť aj na iné účely) využijeme tiež fakt, že teplá i studená fáza zvyknú mať svoje maximum v novembri a decembri. Je viacero ciest, ako podobnú synchronizáciu zahrnúť do modelu. My, v súlade s [12], zahŕňame sezónnu závislosť do dynamickej časti hlavného stupňa modelu. Rigorózne povedané, predpokladáme, že maticová

funkcia $\mathbf{B}^{(0)}(t)$ a vektorová funkcia $\mathbf{c}^{(0)}(t)$ sú časovo závislé a periodické s periódou $T = 12$ mesiacov:

$$\begin{aligned} \mathbf{B}^{(0)}(t) &= \mathbf{B}_0 + \mathbf{B}_s \sin(2\pi t/T) + \mathbf{B}_c \cos(2\pi t/T), \\ \mathbf{c}^{(0)}(t) &= \mathbf{c}_0 + \mathbf{c}_s \sin(2\pi t/T) + \mathbf{c}_c \cos(2\pi t/T), \end{aligned} \tag{6}$$

čiže v našom prípade použijeme celú dostupnú dĺžku dát na odhad štyroch sezónne závislých koeficientov. Model (čiže jeho parametre — tenzor $\mathbf{A}$, matice $\mathbf{B}^{(i)}$ a vektory $\mathbf{c}^{(i)}$) odhadneme v priestore komponent z PCA rozkladu (pripomíname, že PCA je rozklad časovo-priestorových dát pomocou singulárnej dekompozície kovariančnej matice na priestorové módy, ktoré vysvetľujú najviac variancie; z rozkladu dostaneme časovo-nemenné, priestorové komponenty a k nim priradené časové rady) [7] teplôt povrchu oceánu tropického Pacifiku. Optimálna (v zmysle podobnosti modelu a skutočných dát) dimenzionalita modelu (počet zložiek vektoru $\mathbf{x}$), na ktorých model natrénujeme, sa určuje heuristicky, pomocou kros-validácie, kedy skúšame rôzne možnosti a vyberieme možnosť s najvyššou podobnosťou voči reálnym dátam.

## 3. Výsledky: porovnanie modelu s dátami

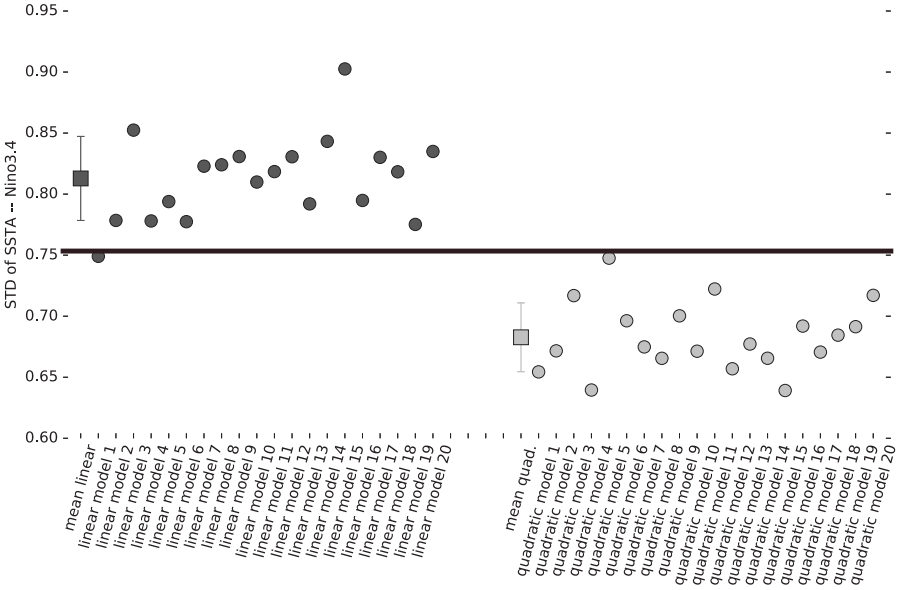
V porovnávaní výsledkov nášho modelu sa zameriame na jeden z indexov opisujúcich fenomén ENSO, a to konkrétne tzv. NINO3.4 index. Tento je definovaný ako časový rad priestorového priemeru povrchovej teploty oceánu v boxe ohraničenom 5° južnej až 5° severnej šírky a 120° až 170° západnej dĺžky. Daný index sa zakladá na HadISST1 datasete [21], ktorý predstavuje súbor dát povrchových teplôt oceánu a koncentrácie ľadovcov s mesačným rozlíšením od roku 1871, vzorkovaných pri rozlíšení $5° \times 5°$. Ak použijeme model a značenie z predchádzajúcej sekcie, potom $\mathbf{X}(t)$ je vektor teplôt povrchu oceánu v bodoch mriežky v čase $t$, ktorý meriame v mesiacoch, $\overline{\mathbf{X}}(t)$ je vektor priemerných teplôt v danom mesiaci a vektor $\mathbf{x}(t) = \mathbf{X}(t) - \overline{\mathbf{X}}(t)$ popisuje anomálie teplôt. Pripomíname, že s ohľadom na zníženie dimenzie problému nepracujeme priamo s vektormi teplôt, ale s ich PCA transformáciami.

V prvej fáze porovnávania sa zameriame na základné charakteristiky indexu NINO3.4 — jeho veľkosť, sezónnu závislosť a spektrum.

### 3.1. Základné charakteristiky

Ako prvú porovnáme amplitúdu (veľkosť) NINO3.4 indexu. Definujeme ju ako štandardnú odchylku (*STD*) anomálií (odchýliek od priemeru za celé obdobie, pre ktoré máme dáta k dispozícii) teploty povrchu oceánu (*SSTA — sea surface temperature anomalies*) v boxe NINO3.4. Na obr. 2 môžeme vidieť amplitúdu indexu NINO3.4 z dát ako tučnú čiernu čiaru a po 20 realizácií (20 nezávislých integrácií modelu z rôznych náhodných počiatočných podmienok a za použitia náhodného bieleho šumu pri integrácii) z lineárneho (tmavo-šedé guličky) a kvadratického modelu (bledo-šedé guličky). Oba modely boli natrénované na podmnožine prvých 20 priestorových komponent z PCA rozkladu a integrované v čase, aby sme dosiahli to, že časové rady budú mať rovnakú dĺžku ako reálne dáta.

Z obrázku vyplýva, že lineárny model ľahko preceňuje amplitúdu ENSO a kvadratický model naopak amplitúdu mierne podceňuje. Niektoré realizácie oboch modelov



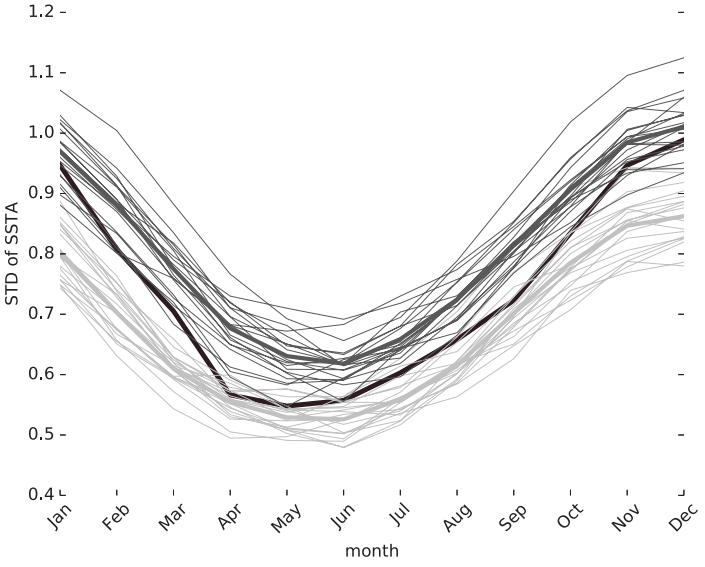

Obr. 2. Amplitúda ENSO, ako štandardná odchýlka NINO3.4 anomálií v pozorovaných dátach (čierna čiara) a v 20 realizáciach lineárneho (tmavo-šedé guličky) a kvadratického (bledošedé guličky) modelu. V rovnakých farbách je vykreslený aj súborový priemer ako štvorec pre tieto modely.

sa reálnej amplitúde ENSO veľmi približujú, na základe čoho môžeme tvrdiť, že táto závisí na počiatočných podmienkach a na šume dodanom počas integrácie. Súborové priemery lineárneho ako aj kvadratického modelu sa celkovo s dostatočnou presnosťou približujú reálnej amplitúde.

Ďalšou základnou a dôležitou charakteristikou ENSO je jeho sezónnosť. Ako bolo spomenuté vyššie, ENSO sa vyznačuje tendenciou vrcholiť v zime na severnej pologuli, čo sa preukazuje zvýšenou varianciou počas zimných mesiacov a zníženou varianciou počas jari a leta na severnej pologuli. Ako charakteristiku teda zvolíme štandardnú odchýlku SSTA za jednotlivé mesiace a tieto charakteristiky pre reálne dáta a takisto aj pre naše modely môžeme vidieť na obr. 3.

Vidíme, že oba modely dobre reprezentujú sezónnosť ENSO v zmysle zvýšenej variancie v zimných mesiacoch s minimom na jar na severnej pologuli. Je potrebné podotknúť, že v reálnych dátach je rozdiel v mesačných varianciách väčší ako v oboch modeloch. Opäť platí, že súborové priemery oboch modelov dostatočne dobre reprezentujú sezónnosť, akú môžeme nájsť v dátach.

Poslednou, ale nemenej dôležitou charakteristikou je frekvenčné spektrum teplotných anomálií, z ktorého sa dozvieme typické periodicity (frekvencie) signálu. Fenomén ENSO nemá veľmi konkrétnu periodicitu, väčšinou sa v literatúre stretávame s pojmom „ENSO band“, ktorý reprezentuje fakt, že periodicita ENSO nie je pevne definovaná (jeho spektrum nemá jasné maximum) a uvádza sa v rozmedzí 3–7 rokov. Tento spektrálny pás môžeme vidieť aj na obr. 4 (spektrálna hustota bola počítaná Welchovou metódou [27]) s mierne výrazným maximom na 9 rokoch.

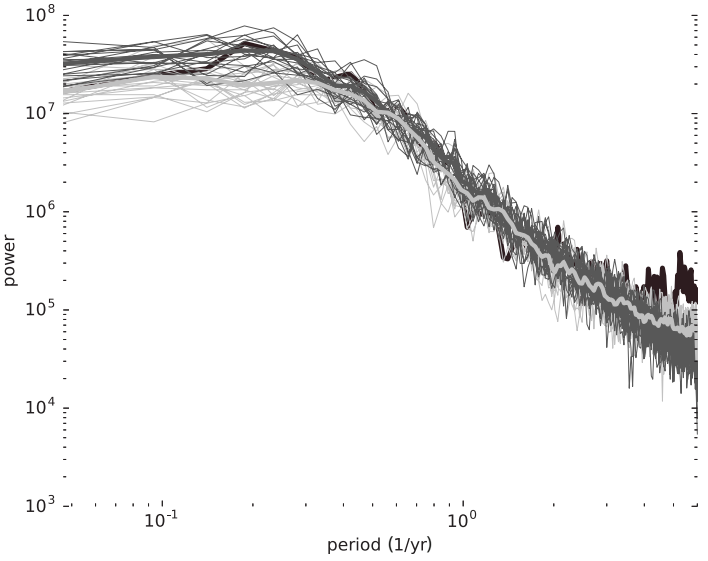



Obr. 3. Sezónnosť ENSO, ako štandardná odchýlka NINO3.4 anomálií za mesiac v pozorovaných dátach (čierna čiara) a v 20 realizáciach lineárneho (tmavo-šedé čiary) a kvadratického (bledo-šedé čiary) modelu. V rovnakých farbách je vykreslený súborový priemer ako tučné čiary pre tieto modely.

Oba modely relatívne dobre kopírujú spektrum NINO3.4 indexu — či už jednotlivé realizácie, alebo súborové priemery. Vo výsledku môžeme tvrdiť, že štatistický model integrovaný na rovnaké obdobie ako dáta verne modeluje fenomén ENSO a dokáže kopírovať jeho základné (lineárne) štatistiky.

## 4. Vylepšenia modelu: parametrizácia šumu

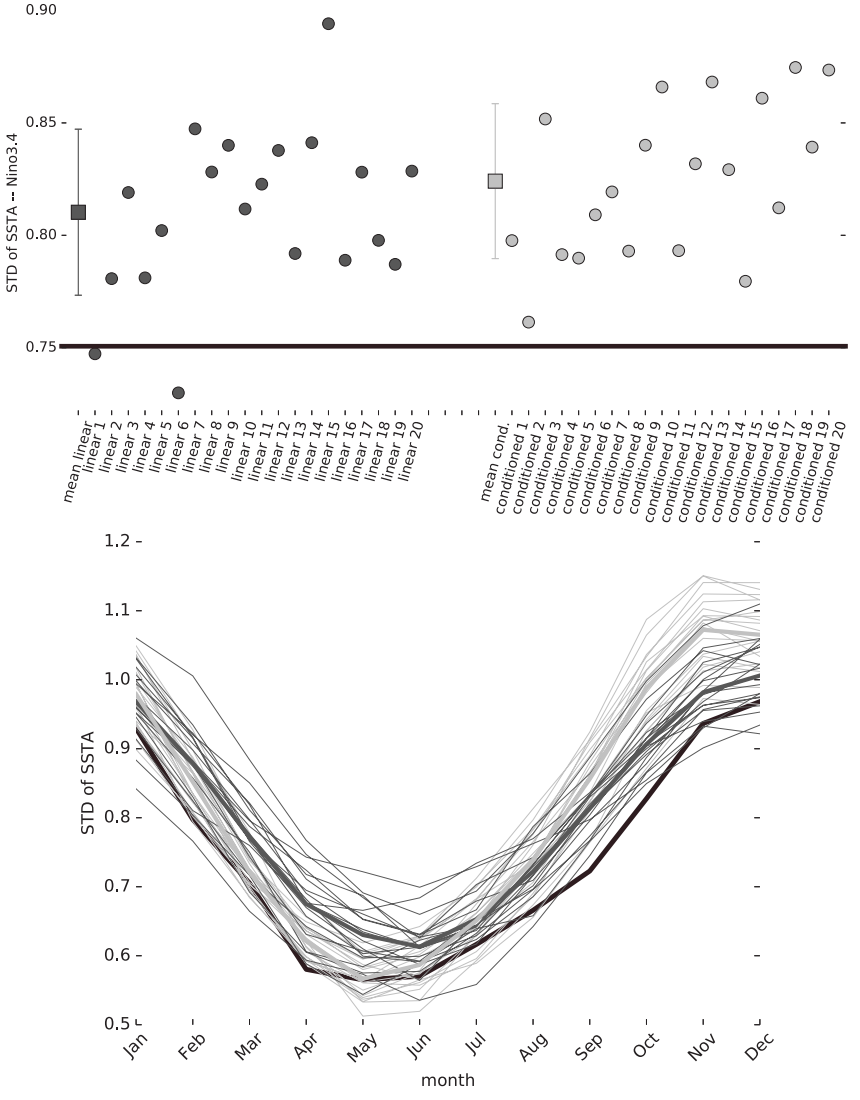
Po navrhnutí dizajnu a natrénovaní štatistického modelu nasleduje samotná integrácia, počas ktorej pridávame do modelu šum. Tento šum môže mať priestorovú štruktúru v zmysle korelácií medzi jednotlivými zložkami, ktorá by mala ostať rovnaká ako v dátach. Ako bolo spomenuté vyššie, model obsahuje toľko stupňov, aby rezíduá posledného stupňa boli biele v čase. To však neznamená, že tento šum nemôže byť v priestore nejako korelovaný. V najzákladnejšom a najintuitívnejšom nastavení si jednoducho spočítame kovariančnú maticu rezíduí z posledného stupňa a následne jej Choleského dekompozíciu na spodnú trojuholníkovú maticu a k nej konjugovanú transponovanú maticu $\mathbf{R}$ (tento krok je potrebný kvôli stabilite integrácie). Pri integrácii si v každom časovom kroku vyrobíme náhodnú realizáciu bieleho šumu o časovej dĺžke 1 a tento náhodný vektor prenásobíme maticou $\mathbf{R}$, z čoho dostaneme realizáciu priestorovo korelovaného bieleho šumu, ktorý dosadíme do modelu. Tomuto procesu hovoríme parametrizácia šumu. V praxi sa však ukazuje, že aj pri vysokostupňových modeloch je stále možné, že rezíduá budú mať nejakú zložitú štruktúru (autokorelácie pri vysokých oneskoreniach, sezónnosť apod.) a parametrizácia za pomoci priestorovej korelácie by nebola dostatočná. V tom prípade môžeme siahnuť po jednej zo zložitejších parametrizácií.




Obr. 4. Spektrum ENSO, odhad spektrálnej hustoty pomocou Welchovej metódy NINO3.4 anomálií v pozorovaných dátach (čierna čiara) a v 20 realizáciach lineárneho (tmavo-šedé čiary) a kvadratického (bledo-šedé čiary) modelu. V rovnakých farbách je vykreslený súborový priemer ako tučné čiary pre tieto modely.

Jedno z vylepšení parametrizácie šumu vychádza z konceptu modelovania klimatických procesov, ktoré vykazujú variabilitu na nízkych frekvenciách (LFV — *low frequency variability*). Princíp tejto metódy spočíva v hľadaní vzoriek (časových úsekov) v rezíduách (tzv. *noise snippets*), ktoré sa objavovali v systéme už v minulosti, vo fáze LFV, tesne pred práve pozorovaným stavom. Tieto (alebo informácie, ktoré z nich získame) následne použijeme na integráciu nášho systému do budúcnosti. Táto metóda je popísaná v [10].

Šumové vzorky, ktoré nájdeme v minulých stavoch systému, môžeme použiť dvoma rôznymi spôsobmi: prvým je priame použitie týchto vzoriek z minulého stavu systému na integráciu modelu do budúcnosti (ako to je použité v [10]) ako súboru viacerých integrácií. Napríklad, ak nájdeme 4 rôzne časové intervaly, ktoré pripomínaju súčasný stav LFV systému, tak použijeme priamo všetky 4 vzorky ako šum pri integrácii a následne integračný výsledok spriemerujeme. Druhý spôsob (použitý v našej štúdii) je nájdenie daného počtu vzoriek (napríklad 100) z minulosti systému, ktoré sú najbližšie súčasnému stavu LFV systému a vytvorenie kovariančnej matice z týchto 100 vzoriek. Následne vytvoríme maticu $\mathbf{R}$ pomocou Choleského dekompozície a náhodný šumový vektor sa touto maticou vynásobí. Bez ohľadu na to, ktorý z vyššie uvedených spôsobov použijeme, existuje viacero možností, ako odhadnúť súčasný stav systému. Jednou z najpoužívanejších je odhad pomocou korelácie SSA časových radov (SSA — *singular spectrum analysis* [4] je rozklad kovariančnej matice, ktorá je navyše rozšírená o rôzne časové oneskorenia, v princípe podobný PCA), ktoré vytvoríme z časových radov hlavných komponent vstupov do modelu. Druhý spôsob spočíva v použití euklidovskej vzdialenosti v podpriestore modelu prvých pár hlavných komponent.




Obr. 5. Amplitúda ENSO (hore) a sezónnosť ENSO (dole), dáta ako čierna čiara, lineárny model v tmavo-šedej farbe, lineárny model so šumom závislým na stave systému v bledo-šedej farbe.

Na obr. 5 môžeme vidieť porovnanie klasického lineárneho modelu (s bielym šumom s klasickou kovariančnou maticou tvorenou z celých dát) a lineárneho modelu so šumom s kovariančnou maticou závislou na súčasnom stave systému. V našom prípade bol súčasný stav systému odhadnutý pomocou metódy SSA. Ako je zrejmé z obrázku, iná parametrizácia šumu amplitúde nepomohla, avšak pri sezónnosti môžeme vidieť lepší prechod medzi zimou na severnej pologuli s vysokou varianciou, a naopak, jarou a letom na severnej pologuli s nízkou varianciou práve v modeli so šumom závislým na stave systému.

## 5. Nelineárne interakcie v dátach a v modeli

Fenomén ENSO je modelovaný počas viacerých dekád a rôzne modely majú svoje výhody a nevýhody, ale v princípe všetky dosahujú veľmi dobré výsledky v predpovedi na 3 mesiace vopred. My sme náš model chceli zostrojiť tak, aby verne kopíroval nelineárne interakcie v dynamike ENSO. Tým, že model je jednoduchší ako realita, je tento krok dôležitý v hľadaní mechanizmov, ktoré stoja za nelineárnymi interakciami. Nás zaujímajú nelineárne interakcie medzi rôznymi časovými škálami v dynamike ENSO, konkrétne sa zameriame na fázovú synchronizáciu a kauzalitu.

### 5.1. Metódy

Najskôr musíme NINO3.4 index, ktorý je v princípe časovou radou teplotných anomálií v oceáne, rozložiť na jednotlivé zložky, kde každá reprezentuje časť signálu na konkrétnej časovej škále. Z dát si pomocou pásmového filtra vieme získať napríklad ročnú zložku, ktorej typická periodicita je jeden rok a iné frekvencie sa v nej — až na malé odchýlky — nevyskytujú. Z pôvodného časového radu $s$ tak získame škálovú zložku $s_f$ o frekvencii $f$. Najjednoduchším spôsobom, ako odhadnúť jej fázu a amplitúdu, je vypočítať jej imaginárnu zložku $\hat{s}_f$ pomocou Hilbertovej transformácie (lineárna operácia, pomocou ktorej sa počíta analytická reprezentácia signálu). Potom môžeme písať

$$s_f(t) + \mathrm{i}\hat{s}_f(t) = A_f(t)\exp(\mathrm{i}\phi_f(t)), \tag{7}$$

kde

$$\phi_f(t) = \operatorname{arctg}\frac{\hat{s}_f(t)}{s_f(t)} \tag{8}$$

je fáza oscilačného signálu s frekvenciou $f$ a

$$A_f(t) = \sqrt{s_f^2(t) + \hat{s}_f^2(t)} \tag{9}$$

je jeho amplitúda.

Vyššie popísané dva kroky — filtrovanie a následnú Hilbertovu transformáciu — môžeme zjednodušiť na jeden použitím waveletovej transformácie. Waveletová transformácia je vlastne konvolúcia vopred predpísanej vlny o určitej perióde s dátami, čím vznikajú oscilačné časové rady na frekvencii predpísanej vlny. Výstupom komplexnej waveletovej transformácie sú reálna a imaginárna zložka časového radu, z ktorých môžeme priamo vypočítať fázu a amplitúdu škálovej zložky o danej frekvencii (perióde); pre detaily metodológie odporúčame [16].

Našim hlavným cieľom bolo odhadnúť kauzálne vzťahy medzi oscilačnými zložkami na rôznych periódach. Existuje viacero metód na rozpoznávanie synchronizácie a kauzality, my používame nástroj z teórie informácie, konkrétne vzájomnú informáciu a jej podmienenú verziu. Vzájomná informácia sa v prípade diskrétnych vstupných veličín počíta ako

$$I(X;Y) = \sum_x \sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{10}$$

kde $p(x,y)$ je združená pravdepodobnosť náhodných veličín $X$, $Y$ a $p(x)$, $p(y)$ sú marginálne pravdepodobnosti veličín $X$ a $Y$. Vzájomná informácia v podstate vyjadruje (ako už jej názov napovedá), koľko informácie zdieľajú dve rôzne veličiny, $X$ a $Y$.

Ak nezdieľajú žiadnu spoločnú informáciu (sú nezávislé), ich združená distribučná funkcia sa rovná súčinu jednotlivých marginálnych distribučných funkcií a v sume je logaritmus 1, čo sa rovná 0. Vzájomná informácia nemá vrchnú hranicu, je zhora neobmedzená. Na rozdiel, napríklad od korelácie, odhalí aj nelineárne vzťahy medzi náhodnými premennými a preto je niekedy prezývaná aj „nelineárna korelácia".

O vzájomnej informácii môžeme hovoriť nie iba v prípade náhodných veličín, ale tiež v prípade dvojice časových radov $x(t)$ a $y(t)$, $t \in T$. Značíme ju opäť $I(x(t); y(t))$ a počítame podľa vzorca (10), kde príslušné distribučné funkcie odhadneme pomocou histogramov; pre prehľad rôznych odhadov vzájomnej informácie odporúčame [9].

Ešte zaujímavejším konceptom je podmienená vzájomná informácia, ktorá udáva, koľko spoločnej informácie máme v dvoch náhodných veličinách, ak si odmyslíme pôsobenie tretej veličiny. Pomocou podmienenej vzájomnej informácie môžeme odhadovať mimo iného aj kauzálne vzťahy medzi dvoma veličinami a to tak, že za tretiu veličinu, na ktorú podmieňujeme, dosadíme minulosť jednej z prvých dvoch veličín, číže efektívne počítame vzájomnú informáciou medzi dvoma veličinami, podmieňujúc na minulosť jednej z nich.

V tejto štúdii pozorujeme dve miery nelineárných vzťahov: fázovú synchronizáciu a fázovo-amplitúdovú kauzalitu. Fázová synchronizácia je jednoducho vzájomná informácia medzi fázami dvoch oscilačných zložiek na rôznych frekvenciách. Môže nás napríklad zaujímať fázová synchronizácia medzi oscilačnou zložkou s ročnou periódou a zložkou pomalšieho, 5-ročného cyklu. Matematicky píšeme $I(\phi_{f_1}(t); \phi_{f_2}(t))$.

Fázovo-amplitúdová kauzalita je už trocha zložitejšia a v jej počítaní používame koncept podmienenej vzájomnej informácie. V našom prípade počítame kauzalitu smerom od fáz k amplitúde. V princípe počítame vzájomnú informáciu fázy $\phi_1$ v čase $t$ na amplitúdu $A_2$ v budúcnosti (v čase $t + \tau$, čiže počítame ako nám znalosť momentálnej fázy pomôže predpovedať amplitúdu v budúcnosti). Zároveň však podmieňujeme na prítomnú a minulú amplitúdu $A_2$ v časoch $t$, $t - \eta$, ... (tu závisí na dimenzii podmienky, my používame 3-dimenzionálnu podmienku), čím sa zbavujeme jej pôsobenia. Celkovo píšeme, že odhad kauzality od fáz k amplitúde počítame ako $I(\phi_{f_1}(t); A_{f_2}(t+\tau) \mid A_{f_2}(t), A_{f_2}(t-\eta), A_{f_2}(t-2\eta))$.

### 5.2. Surogátne dáta

Predtým, ako prejdeme priamo k výsledkom našich analýz, vysvetlíme metódu surogátnych dát. Surogátnymi dátami (*surrogate data* alebo *analogous data*) nazývame metódu na generovanie umelých časových radov, ktoré zachovávajú niektoré vybrané štatistické vlastnosti pôvodných dát, kým ostatné znáhodnia konzistentne s nulovou hypotézou. Ich najčastejšie využitie je v štatistickom testovaní významnosti. Najskôr si zvolíme nulovú hypotézu, ktorá opisuje nejaký proces, a následne vygenerujeme súbor surogátnych dát zodpovedajúcich nulovej hypotéze použitím Monte Carlo metód (opakované náhodné vzorkovanie). Jednou z najpoužívanejších techník na generovanie surogátnych dát sú tzv. fourierovské surogáty (*Fourier transform surrogates* [23]), ktoré zachovávajú lineárne korelácie v dátach (spektrum alebo periodogram a taktiež autokoreláciu), ale znehodnotia akékoľvek interakcie medzi nimi. Pri tvorbe týchto surogátov dáta pretransformujeme do frekvenčného priestoru pomocou Fourierovej transformácie, ich fázy znáhodníme a následne späť pretransformujeme inverznou Fourierovou transformáciou do priestoru reálneho času.

Metódu a jej použitie ilustrujeme na jednoduchom príklade — predstavme si dva prepojené dynamické systémy (napríklad Lorenzov systém), kde jeden z nich ovláda druhý v kauzálnom zmysle (to znamená, že v rovniciach pre druhý systém sa vyskytuje premenná z prvého systému). Tento systém naintegrujeme a máme k dispozícii merania — časové rady z oboch systémov. Chceme zistiť, či a ako sú prepojené a ak sú, chceme vedieť smer kauzality. Na tento účel použijeme napríklad podmienenú vzájomnú informáciu. Z analýzy vzájomnej informácie dostaneme výsledok — jedno číslo. Toto nám však na zistenie výsledku nestačí — hodnotu sme mohli dostať aj „náhodou“. Za účelom štatistického testovania si teda vytvoríme súbor surogátnych dát (napríklad už zmienených Fourierových surogátov) a analýzu podmienenej vzájomnej informácie zopakujeme pre každý člen súboru zvlášť. Z hodnôt vzájomných informácií pre každý člen zostrojíme histogram a pozrieme sa, kam padne hodnota z dát. Ak prekročí vopred zvolený percentil (často to býva 95 %), povieme, že kauzálny vzťah je signifikantný.

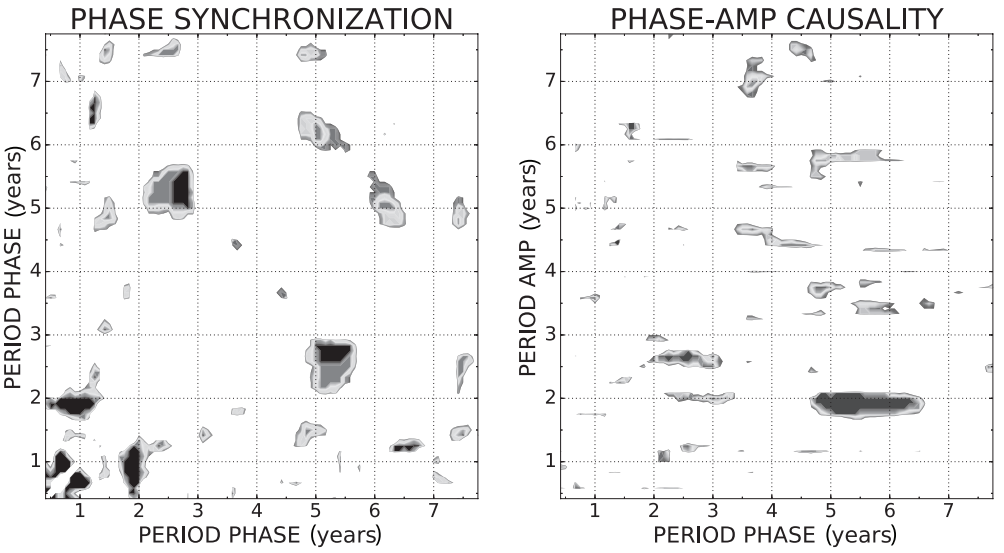
### 5.3. Interakcie v dátach

Pri študovaní nelineárnych interakcií v dátach sa najskôr zameriame na fázovú synchronizáciu. Fázovou synchronizáciou rozumieme vo všeobecnosti proces, v ktorom dva alebo viac cyklických signálov oscilujú s opakujúcou sa sekvenciou fázových rozdielov. V našom klimatickom príklade to znamená, že dve rôzne oscilačné zložky systému sú spolu zosynchronizované — ak dôjde k zmene jednej, simultánne by sa mala zmeniť aj druhá. Fázovú synchronizáciu počítame pomocou vyššie zavedenej vzájomnej informácie ako $I(\phi_{f_1}(t); \phi_{f_2}(t))$, kde $\phi_i(t)$ je časový rad fázy $i$-tej oscilačnej zložky. Výsledky môžeme vidieť na obr. 6 vľavo. Pre každú dvojicu období odhadneme vzájomnú informáciu ich fáz a porovnávame ju so súborom synteticky vyrobených dát zodpovedajúcich nulovej hypotéze. Nulová hypotéza v našom prípade zodpovedá lineárnemu stochastickému procesu s rovnakým spektrom ako dáta, v ktorom ale žiadne medziškálové interakcie neexistujú. V prípade, že hodnota vzájomnej informácie je v dátach vyššia ako daný percentil (často 95 %) z distribúcie surogátnych dát, nulová hypotéza je zamietnutá a tvrdíme, že vzájomná informácia je štatisticky významná. Podobne si zobrazíme aj výsledky z kauzálnej interakcie medzi fázou a amplitúdou ako na obr. 6 vpravo.

Ako je zrejmé z obr. 6, v observačných dátach môžeme vidieť synchronizáciu ročného cyklu s periódami tesne pod jedným rokom, tzv. kombinačnými tónmi a taktiež fázovú synchronizáciu 2:1, čiže tzv. bienálneho (dvojročného) cyklu s ročným. Čo sa týka kauzalít, tvrdíme, že fáza pomalého cyklu s periódou 5–6 rokov kauzálne vplýva na amplitúdu bienálneho cyklu.

### 5.4. Interakcie v modeli

V nasledujúcich riadkoch sa pozrieme na podobné interakcie, ale tentoraz v modeloch. Či dokážeme nájsť rovnaké interakcie v modeloch a v observačných dátach je dôležité z hľadiska ozrejmenia mechanizmov stojacich za týmito interakciami. Komplexita modelov je na rozdiel od prírody značne znížená a model môžeme navyše rozložiť na jeho jednotlivé zložky a sledovať, z akých interakcií v modeli náš výsledok vychádza. Na test synchronizácie a kauzality sme opäť integrovali lineárny model v súbore, pre každú realizáciu zo súboru sme spravili štatistický test oproti surogátnym časovým



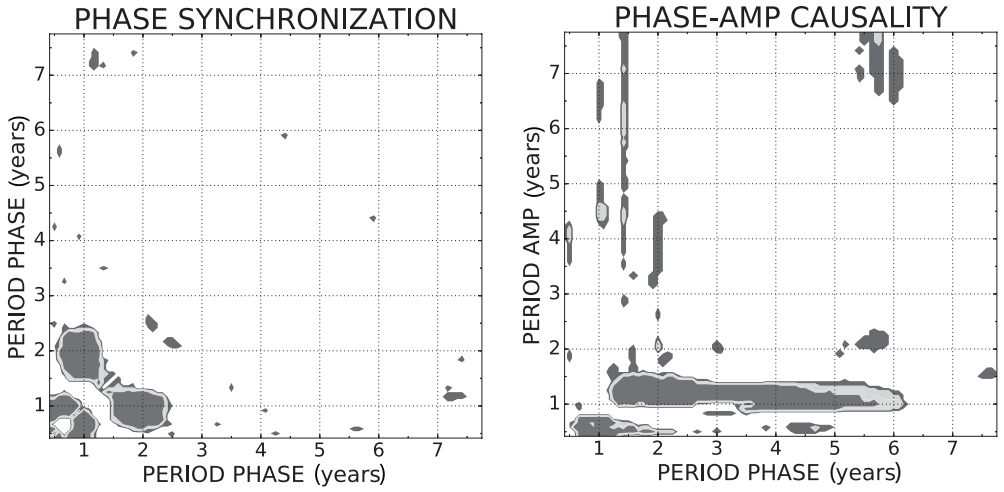

Obr. 6. Fázová synchronizácia (vľavo) a kauzalita fáza → amplitúda (vpravo) v časovej rade NINO3.4 indexu. Zobrazené sú stupne významnosti ako stupne šedi (od 95% percentilu) z odhadu vzájomnej informácie. Štatistický test významnosti oproti 500 syntetickým, surogátnym časovým radom.

radom, takéto grafy sme binarizovali (ak bol daný bod na grafe významný, tak sme mu pridelili hodnotu 1, v opačnom prípade 0) a nakoniec sčítali. Výsledky ukazujeme na obr. 7.

Ako vidno z ľavej časti obr. 7, fázová synchronizácia je správne modelovaná a obsahuje obe synchronizačné pásma, menovite synchronizáciu ročného cyklu s kombinačnými tónami a synchronizáciu ročného a bienálneho cyklu. S kauzalitou fáza-amplitúda to už je horšie – túto interakciu model nedokázal zachytiť. Dôvodom môže byť nízka komplexita modelu, alebo absencia nejakých nelineárnych interakcií v modeli, ktoré observačné dáta (a tým pádom dynamika ENSO) obsahujú.

## 6. Modelovanie surogátnych dát pomocou štatistického modelu

Už sme si vysvetlili, čo sú surogátne dáta a na čo sa používajú. V poslednej časti tohoto článku ukážeme, ako generovať surogátne dáta pomocou nášho štatistického modelu. Pri štúdiu nelineárnych interakcií použitím metód ako vzájomná informácia, je použitie surogátnych dát na testovanie štatistickej významnosti nutné. Samozrejme, môžeme použiť už spomenuté fourierovské surogátne dáta, čím sa nulovou hypotézou stane lineárny proces s rovnakým spektrom. Môžeme však vytvoriť sofistikovanejšiu nulovú hypotézu, využívajúc možnosti štatistických modelov: ak vezmeme lineárny model, zanedbáme dynamickú sezónnosť členov $\mathbf{B}^{(0)}$ a $\mathbf{c}^{(0)}$ v rovnici (6) a použijeme základnú parametrizáciu šumu (vezmeme do úvahy iba priestorové korelácie), tento model bude kopírovať základné (lineárne) štatistické vlastnosti modelovaných časových radov, ale nebude umožňovať napríklad nelineárne interakcie. Týmto spôsobom ho môžeme využiť ako nulovú hypotézu.



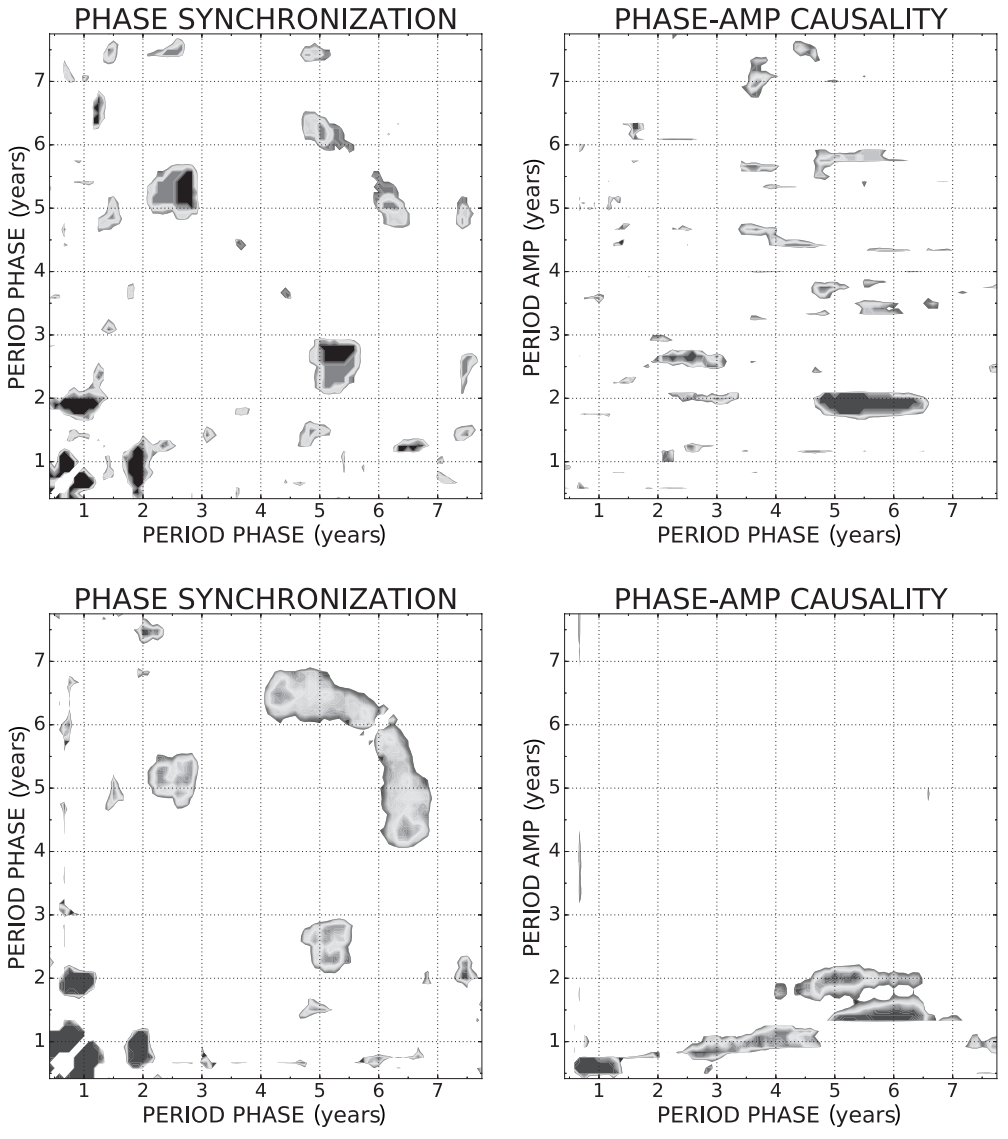

Obr. 7. Suma 5 členov súboru pre fázovú synchronizáciu (vľavo) a kauzality fáza → amplitúda (vpravo) v v časovom rade modelovaného NINO3.4 indexu zo súboru lineárneho modelu. Ukázané sú hladiny významnosti (od 95% percentilu) z odhadu vzájomnej informácie. Štatistický test významnosti oproti 500 syntetickým, surogátnym časovým radom.

Použitie štatistického modelu na vytváranie nulových hypotéz v praxi môžeme vidieť na obr. 8. Keď porovnáme riadky (odlišné typy surogátov, čiže odlišné nulové hypotézy), v oboch máme rovnaké významné interakcie. Konkrétne sa jedná o synchronizáciu fáz ročného cyklu s kombinačnými tónmi a ročného cyklu s bienálnym. Taktiež pozorujeme kauzalitu fáza — amplitúda, kde fáza pomalého cyklu (5–6 rokov) kauzálne ovplyvňuje amplitúdu bienálneho cyklu. Pri použití štatistického modelu nízkej komplexity ako nulovej hypotézy vidíme — hlavne pri kauzálnych interakciách — že model tohto typu má menej štatistických „fluktuácií“, resp. falošných pozitív, ako fourierovské surogáty. Toto je očakávateľné, pretože model verne kopíruje základnú (lineárnu) štruktúru dát a vynecháva iba sezónnu závislosť na ročnom cykle a nelineárne interakcie. Použitím štatistického modelu na testovanie hypotéz dostávame lepšiu predstavu o významných interakciách, zároveň je takýto test prísnejší (nulová hypotéza je konkrétnejšia), čo má za následok zníženie počtu falošne pozitívnych výsledkov.

## 7. Zhrnutie

Štatistickým modelom v klimatológii sa dostáva čoraz väčšej pozornosti, aj z dôvodu, že ich použitie nie je limitované iba na predpovedné účely (napríklad spomenutého fenoménu ENSO), ale aj na identifikáciu interakcií v zložitých systémoch. Keďže sú väčšinou zostrojené v redukovanom fázovom priestore s nízkou dimenzionalitou, identifikácia zdrojov týchto interakcií sa stane jednoduchšou.

Ukázali sme, že štatistické modely s vhodným nastavením (ktoré súvisí so základným pochopením modelovaného javu) vedia generovať syntetické časové rady modelovaného systému (u nás to bolo ENSO), ktoré kopírujú vlastnosti observačných dát — lineárne aj nelineárne. Ďalej sme sa venovali rôznym parametrizáciam šumu.




Obr. 8. Porovnanie nulových hypotéz: fázová synchronizácia (vľavo) a kauzalita fáza → amplitúda (vpravo) v časovom rade NINO3.4 indexu. Zobrazené sú hladiny významnosti (od 95% percentilu) z odhadu vzájomnej informácie. Štatistický test významnosti oproti 500 (hore) fourierovským surogátnym časovým radom a (dole) časovým radom z nízkokomplexného štatistického modelu.

Keďže stochasticita je dôležitým aspektom štatistických modelov, správne zvolený šum môže mať veľký vplyv na výsledné modelovanie. V našom konkrétnom prípade nám aj zložitejšia parametrizácia šumu pomohla. Záverom sme predstavili spôsob, ako zo štatistických modelov nízkej komplexity vytvárať modely na štatistické testovanie hy-

potéz, ktoré v našom prípade ukázalo lepšie výsledky ako použitie známych metód (napr. fourierovských surogátov).

Budúcnosť štatistických modelov sa môže vyvíjať viacerými smermi. Jedným z nich môže byť vývoj štatistických modelov ako takých — experimentovanie s rôznymi premennými v modeli, kombinovanie meteorologických premenných na vstupe, rôzne metódy a varianty preprocessingu a pod. Ďalším môže byť prepojenie štatistických modelov s dynamickými. Ako konkrétny príklad uveďme možnosť použitia štatistických modelov na parametrizáciu javov s menším rozlíšením ako je mriežka (*sub-grid phenomena*) pri dynamických modeloch, napríklad mikrofyziku oblakov, lokálne konvekčné procesy a pod.

**Poďakovanie.** Autori ďakujú za cenné pripomienky anonymným recenzentom a Lucii Hraškovej za gramatické a štylistické korektúry. Tento článok vznikol za podpory Ministerstva školstva, mládeže a telovýchovy v rámci projektu KONTAKT II, číslo LH14001.

L i t e r a t ú r a

[1] Alexander, M. A., Bladé, I., Newman, M., Lanzante, J. R., Lau, N.-C., Scott, J. D.: *The atmospheric bridge: The influence of ENSO teleconnections on air–sea interaction over the global oceans.* J. Climate *15* (2002), 2205–2231.

[2] Bjerknes, V.: *Dynamic meteorology and hydrology, Part II.* Kinematics. Gibson, Carnegie Institute, New York, 1911.

[3] El Niño–Southern Oscillation diagrams [cit. 13. 12. 2016]. Dostupné z: `http://www.pmel.noaa.gov/tao/proj_over/diagrams/index.html`

[4] Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., et al.: *Advanced spectral methods for climatic time series.* Reviews Geophys. *40*, (2002), 1–41.

[5] Ghil, M., Robertson, A. W.: *Solving problems with GCMs: General circulation models and their role in the climate modeling hierarchy.* Academic Press, 2000, 285–325.

[6] Global patterns — El Nino-Southern Oscillation (ENSO) [cit. 29. 11. 2016]. Dostupné z: `http://climate.ncsu.edu/climate/patterns/ENSO.html`

[7] Hannachi, A., Jolliffe, I. T., Stephenson, D. B.: *Empirical orthogonal functions and related techniques in atmospheric science: A review.* Int. J. Climatol. *27* (2007), 1119–1152.

[8] Hannachi, A., Stephenson, D. B., Sperber, K. R.: *Probability-based methods for quantifying nonlinearity in the ENSO.* Clim. Dyn. *20* (2003), 241–256.

[9] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J.: *Causality detection based on information-theoretic approaches in time series analysis.* Phys. Rep. *441 (1)* (2007), 1–46.

[10] Chekroun, M. D., Kondrashov, D., Ghil, M.: *Stochastic systems by noise sampling, and application to the El Nino-Southern Oscillation.* Proc. Natl. Acad. Sci. USA *108* (2011), 11766–11771.

[11] Kondrashov, D., Kravtsov, S., Robertson, A. W., Ghil, M.: *A hierarchy of data-based ENSO models.* J. Climate *18* (2005), 4425–4444.

[12] Kravtsov, S., Kondrashov, D., Ghil, M.: *Multilevel regression modeling of nonlinear processes: Derivation and applications to climate variability.* J. Climate *18* (2005), 4404—4424.

[13] Lorenz, E. N.: *Deterministic nonperiodic flow.* J. Atmos. Sci. *20* (1963), 130–141.

[14] McCullagh, P., Nelder, J. A.: *Generalized linear models.* Chapman and Hall, 1989.

[15] Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., Taylor, K. E.: *The WCRP CMIP3 multi-model dataset: A new era in climate change research.* A Bull. Amer. Meteor. Soc. *88* (2007), 1383–1394.

[16] Paluš, M.: *Multiscale atmospheric dynamics: Cross-frequency phase-amplitude coupling in the air temperature.* Phys. Rev. Lett. *112* (2014), 1–5.

[17] Paluš, M.: *From nonlinearity to causality: statistical testing and inference of physical mechanisms underlying complex dynamics.* Contemp. Phys. *48* (2007), 307–348.

[18] Penland, C.: *Random forcing and forecasting using principal oscillation pattern analysis.* Mon. Weath. Rev. *117* (1989), 2165–2185.

[19] Penland, C., Ghil, M.: *Forecasting northern hemisphere 700-mb geopotential height anomalies using empirical normal modes.* Mon. Weath. Rev. *121* (1993), 2355–2372.

[20] Philander, S. G. H.: *El Nino, La Nina, and the southern oscillation.* Academic Press, 1990.

[21] Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., Kaplan, A.: *Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century.* J. Geophys. Res. *108* (2003), 4407.

[22] Taylor, K. E., Stouffer, R. J., Meehl, G. A.: *An overview of CMIP5 and the experiment design.* A Bull. Amer. Meteor. Soc. *93* (2012) 485–498.

[23] Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J. D.: *Testing for nonlinearity in time series: The method of surrogate data.* Phys. D *58* (1992), 77–94.

[24] Timmermann, A., Voss, H. U., Pasmanter, R.: *Empirical dynamical system modeling of ENSO using nonlinear inverse techniques.* J. Phys. Oceanogr. *31* (2001), 1579–1598.

[25] Trenberth, K. E., Caron, J. M., Stepaniak, D. P., Worley, S.: *Evolution of El Nino–Southern Oscillation and global atmospheric surface temperatures.* J. Geophys. Res.: Atmospheres *107* (2002), 4065.

[26] Van den Dool, H. M.: *Long-range weather forecasts through numerical and empirical methods.* Dyn. Atmos. Oceans *20* (1994), 247–270.

[27] Welch, P. D.: *The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms.* IEEE Trans. Audio *AU-15* (1967), 70–73.

[28] Wetherill, G. B.: *Regression analysis with applications.* Chapman and Hall, 1986.