NH winter forecast skill of AO and NAO indices: results and sampling issues

Tim Stockdale, ECMWF

Reusing material from earlier talks given together with Laura Ferranti and Franco Molteni
Outline

• Intro: ECMWF System 4

• Predicting NH winter circulation modes

• Challenges of sampling

• Discussion
System 4 configuration

- **IFS**: Tₗ255L91 Cy36r4

- **Real time forecasts:**
  - 51 member ensemble forecast to 7 months
  - SST and atmos. perturbations added to each member

- **Back integrations from 1981-2010 (30 years)**
  - 15 member ensemble every month
  - 15 members extended to 13 months once per quarter
  - **51 members** for Feb/May/Aug/Nov starts
ENSO forecasts are good .....
So are probabilistic scores ....

15 members

JJA Europe T2m>upper tercile
Re-forecasts from 1 May, 1981-2010
Reliability score: 0.987
ROC skill score: 0.38

51 members

JJA Europe T2m>upper tercile
Re-forecasts from 1 May, 1981-2010
Reliability score: 0.996
ROC skill score: 0.43

(Figures from Susanna Corti)
Ensemble size important for low-signal areas

15 members
DJF Europe T2m>upper tercile
Re-forecasts from 1 Nov, 1981-2010
Reliability score: 0.902
ROC skill score: 0.06

51 members
DJF Europe T2m>upper tercile
Re-forecasts from 1 Nov, 1981-2010
Reliability score: 0.981
ROC skill score: 0.22

(Figures from Susanna Corti)
Arctic Oscillation

Calculated as first EOF of monthly mean MSLP anomalies, poleward of 20N.

Use same method as CPC, but using ERA interim analysis, 1981-2010.

Model and analysis time-series both obtained by projection onto observed EOF.

Correlation of our observed time-series with CPC is 0.996.
AO re-forecast skill

Correlation (30y) = 0.608

26 years (no volcanoes)
Correlation = 0.73

Surprising because model AO is very noisy ....
Statistical analysis

Unbiased variance estimates: Obs/Tot/Int/Ext: 1.0000 0.8390 0.8316 0.0074

Model/obs stddev ratio: 0.9159
Model/obs stddev ratio interval: 0.693 1.129 \( \leftarrow \) model variability consistent with obs
Bootstrap over nens, pval for ratio=1: 0.7960

SNR actual : 0.0941
SNR jackknife over nens : 0.0202 0.1029 0.1857

ACC actual : 0.6085
ACC basic bootstrap over nens : 0.5568 0.7121 0.8144 \( \leftarrow \) 95% interval due to ensemble size
ACC basic bootstrap over nyears: 0.2052 0.6069 0.8326 \( \leftarrow \) bigger uncertainty range here

ACP from internal sampling: -0.2947 0.0583 0.4010
Mean ACC for nens-1: 0.6049
p val of measured acc if model perfect: 0.9996 \( \leftarrow \) only a 0.0004 chance we could get this correlation

- Model skill for these years is relatively high
- Model predictability limit must be wrong (because we exceed it so much)
## Other teleconnection patterns

<table>
<thead>
<tr>
<th></th>
<th>ACC</th>
<th>S/N</th>
<th>ACP</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNA (EOF)</td>
<td>0.696</td>
<td>0.64</td>
<td>0.54</td>
<td>0.065</td>
</tr>
<tr>
<td>NAO (EOF)</td>
<td>0.465</td>
<td>0.13</td>
<td>0.10</td>
<td>0.017</td>
</tr>
</tbody>
</table>

**PNA** has high skill and high predictability  
**NAO** has moderate skill, and low predictability  

NAO skill is, like AO, higher than expected
Does resolution help?

Project Minerva has run the ECMWF coupled model at different atmospheric resolutions. We have 30 years of winter forecasts, with 51 member ensembles:

<table>
<thead>
<tr>
<th></th>
<th>T319</th>
<th>T639</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNA (EOF)</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>NAO (EOF)</td>
<td>0.36</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>S/N</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNA (EOF)</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>NAO (EOF)</td>
<td>0.17</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*S/N* does not seem to be affected by resolution.

**NAO** structure and skill is significantly (at 5% level) improved by higher atmosphere resolution.
Challenge: sampling errors are large!

Correlation scores, ECMWF S4
95% C.I.s over NENS and NYEARS

Box = 95% interval, bootstrapping on ensemble size
Whiskers = 95% interval, bootstrapping on years included

© ECMWF
NH winter forecasts

Even with 101 members, ensemble mean signal not always well defined
Conclusions

- S4 has substantial skill in predicting AO phase over a 30 year period
  - How typical this is of expected future performance is unknown
  - Amplitude of model signal is too weak
  - Models are noisy

- Scores are unstable
  - Sensitive to choice of years, especially for shorter periods
  - Relative skill of AO and NAO indices can vary between model versions

- Higher resolution (to T639)
  - DOES help NAO in particular (quite big improvement)
  - Does NOT help S/N ratios
  - … according to a single experiment
Conclusions - Sampling

- Sampling over NYEARS
  - Is an obvious problem for systems without high S/N ratios
  - Skill estimates need as many years as possible, but there are limits
  - We need to understand sources of skill to know how far back we can go (to 1979? to 1960? Even earlier??)

- Ensemble size is often too small
  - Given how noisy our models are, we should probably be doing our experiments with ensembles $O(100)$ to get clean results

- Costs
  - So all we need are very high resolution models, large ensembles, lots of start dates … and lots of different experiments to improve our models.