Supplemental material for *Markov-switching State Space Models for Uncovering Musical Interpretation*

1 Algorithms

For completeness, we include here concise descriptions of the Kalman filter and smoother we employ as inputs to our main algorithm. The filter is given in Algorithm A1.

Algorithm A1 Kalman filter: estimate x_i conditional on $\{y_j\}_{j=1}^{i}$, for all $i = 1, \ldots, n$ and calculate the log likelihood for θ

Input: Y, x_0, P_0, d, T, c, Z, and G

$\ell(\theta) \leftarrow 0$ ▷ Initialize the log-likelihood

for $i = 1$ to n **do**

$\mathsf{X}_i \leftarrow d + T x_{i-1|i-1}, \quad P_i \leftarrow Q + T P_{i-1|i-1} T^\top$ ▷ Predict current state

$\widetilde{y}_i \leftarrow c + Z\mathsf{X}_i, \quad F_i \leftarrow G + Z P_i Z^\top$ ▷ Predict current observation

$v_i \leftarrow y_i - \widetilde{y}_i \quad K_i \leftarrow P_i Z^\top F^{-1}$ ▷ Forecast error and Kalman gain

$x_{i|i} \leftarrow \mathsf{X}_i + K_i v_i, \quad P_{i|i} \leftarrow P_i - P_i Z^\top K_i$ ▷ Update

$\ell(\theta) = \ell(\theta) - v_i^\top F^{-1} v_i - \log(|F_i|)$

end for

return $\widetilde{Y} = \{\widetilde{y}_i\}_{i=1}^n$, $\mathsf{X} = \{\mathsf{X}_i\}_{i=1}^n$, $\widetilde{X} = \{x_{i|i}\}_{i=1}^n$, $P = \{P_i\}_{i=1}^n$, $\widetilde{P} = \{P_{i|i}\}_{i=1}^n$, $\ell(\theta)$

To incorporate all future observations into these estimates, the Kalman smoother is required. There are many different smoother algorithms tailored for different applications. Algorithm A2, due to Rauch et al. (1965), is often referred to as the classical fixed-interval smoother (Anderson and Moore, 1979). It produces only the unconditional expectations of the hidden state $\widehat{x}_i = \mathbb{E}[x_i \mid y_1, \ldots, y_n]$ for the sake of computational speed. This version is more appropriate for inference in the type of switching models we discuss in the manuscript.

2 Distance matrix from raw data

In Section 3.3 of the manuscript, we present results for clustering performances using the low-dimensional vector of performance specific parameters learned for our model. An alternative approach is to simply use the raw data, in this case, 231 individual note-by-note

Algorithm A2 Kalman smoother (Rauch-Tung-Striebel): estimate $\widehat{X}$ conditional on Y

Input: χ, $\widetilde{X}$, P, $\widetilde{P}$, T, c, Z.
$t = n$,
$\widehat{x}_n \leftarrow \widetilde{x}_n$,
while $t > 1$ **do**
  $\widehat{y}_i \leftarrow c + Z\widehat{x}_i$, ▷ Predict observation vector
  $e \leftarrow \widehat{x}_i - \chi_i, \quad V \leftarrow P_i^{-1}$,
  $t \leftarrow i - 1$, ▷ Increment
  $\widehat{x}_i = \widetilde{x}_i + \widetilde{P}_i TVe$
end while
return $\widehat{Y} = \{\widehat{y}_i\}_{i=1}^n, \widehat{X} = \{\widehat{x}_i\}_{i=1}^n$

instantaneous speeds measured in beats per minute. In Figure SM-1 we show the result of this analysis. A comparison between this clustering and that given by our model is discussed in some detail in the manuscript.

3 Plotting performances

In Section 3.3 discussed 4 distinct clusters of the 46 performances as well as an "other" category of relatively unique interpretations. Figures SM-2 to SM-6 display the note-by-note tempos along with the inferred interpretive decisions for all performances by clustering. Here we include some of the discussion of these clusters from the main text to clarify the figures.

The first cluster (Figure SM-2) corresponds to performances which are reasonably staid. The emphasis state is rarely visited with the performer tending to stay in the constant tempo state with periods of slowing down at the ends of phrases. Acceleration is never used. Such state preferences are clearly inferred by the model as shown in, e.g., the top row of Figure 11. Furthermore, these performances have relatively low average tempos, and not much difference between the A and B sections.

Recordings in the second cluster (Figure SM-3) tend to transition quickly between states, especially constant tempo and slowing down accompanied by frequent transitory emphases. The probability of remaining in state 1 is the lowest for this cluster while the probability of entering state 2 from state 1 is the highest. The acceleration state is visited

only rarely.

Cluster three (Figure SM-4) is somewhat like cluster one in that performers tend to stay in state 1 for long periods of time, but they transition more quickly from state 3 back to state 1. They also use state 4 frequently whereas cluster one did not. They also tend to have very large tempo contrasts between the A and B sections.

Cluster four (Figure SM-5) has both faster average tempos and more variability from one period of constant tempo to the next. State 4 is rare, with fast constant tempo changes that persist for small amounts of time tending to reflect note emphases.

The remaining performances are relatively different from all other performances (Figure SM-6. If the distance to the third closest performances exceeded 0.35, then the performance was grouped with "other". Essentially, these recordings had at most one similar recording while the four other clusters contained at least 4.

References

Anderson, B. D., and Moore, J. B. (1979), *Optimal filtering*, Prentice-Hall, Englewood Cliffs, NJ.

Rauch, H. E., Striebel, C., and Tung, F. (1965), "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, **3**(8), 1445–1450.

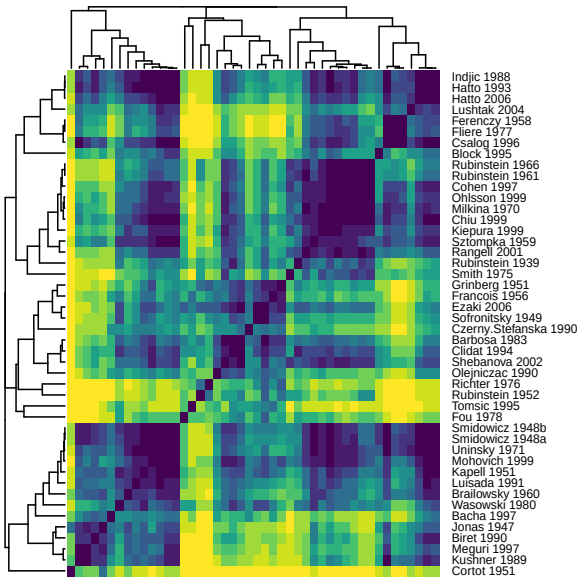

Figure SM-1: This figure presents a heatmap and hierarchical clustering based only on the note-by-note onset timings for each of the 46 recordings.

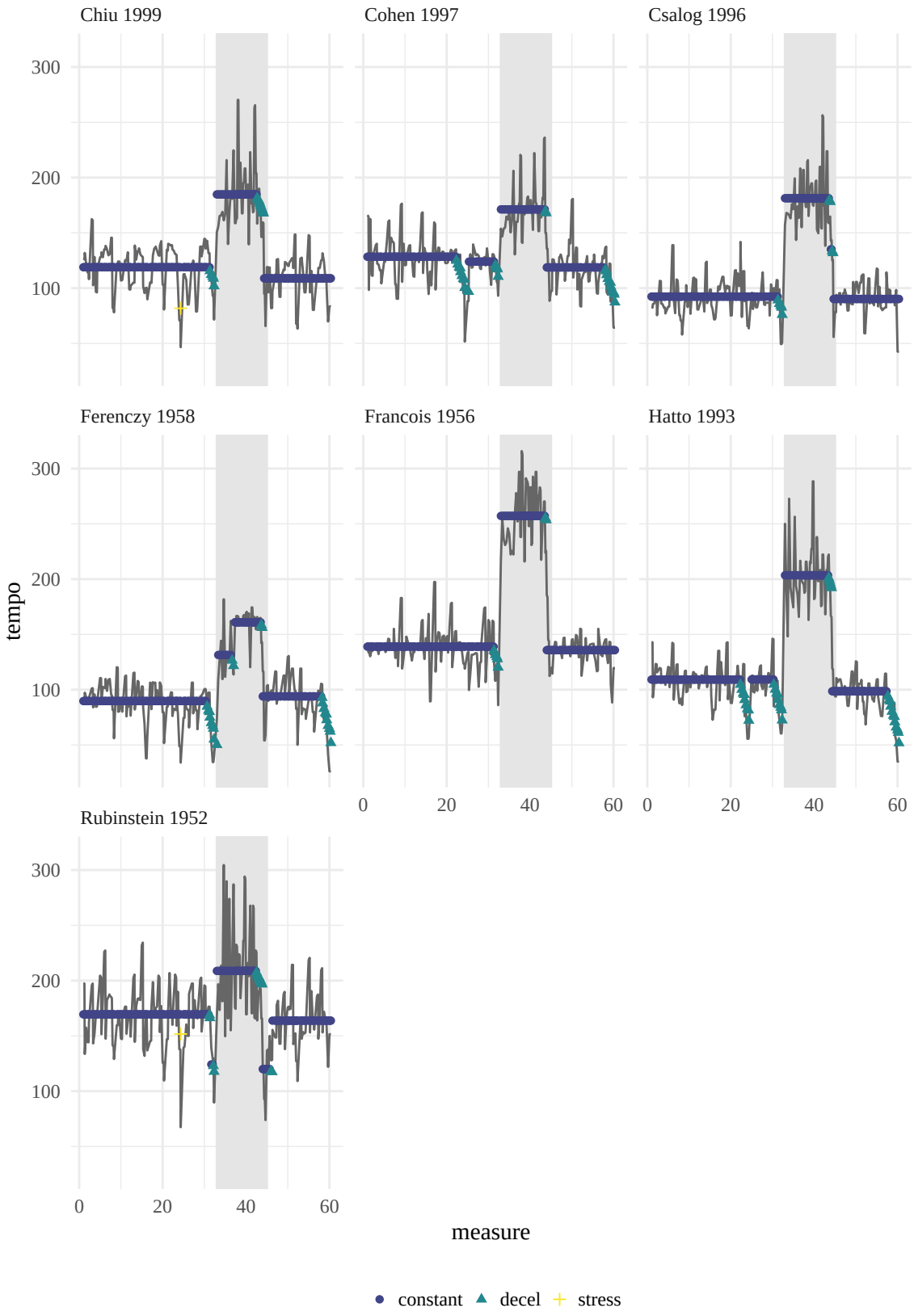

Figure SM-2: Performances in the first cluster.

Figure SM-3: Performances in the second cluster.

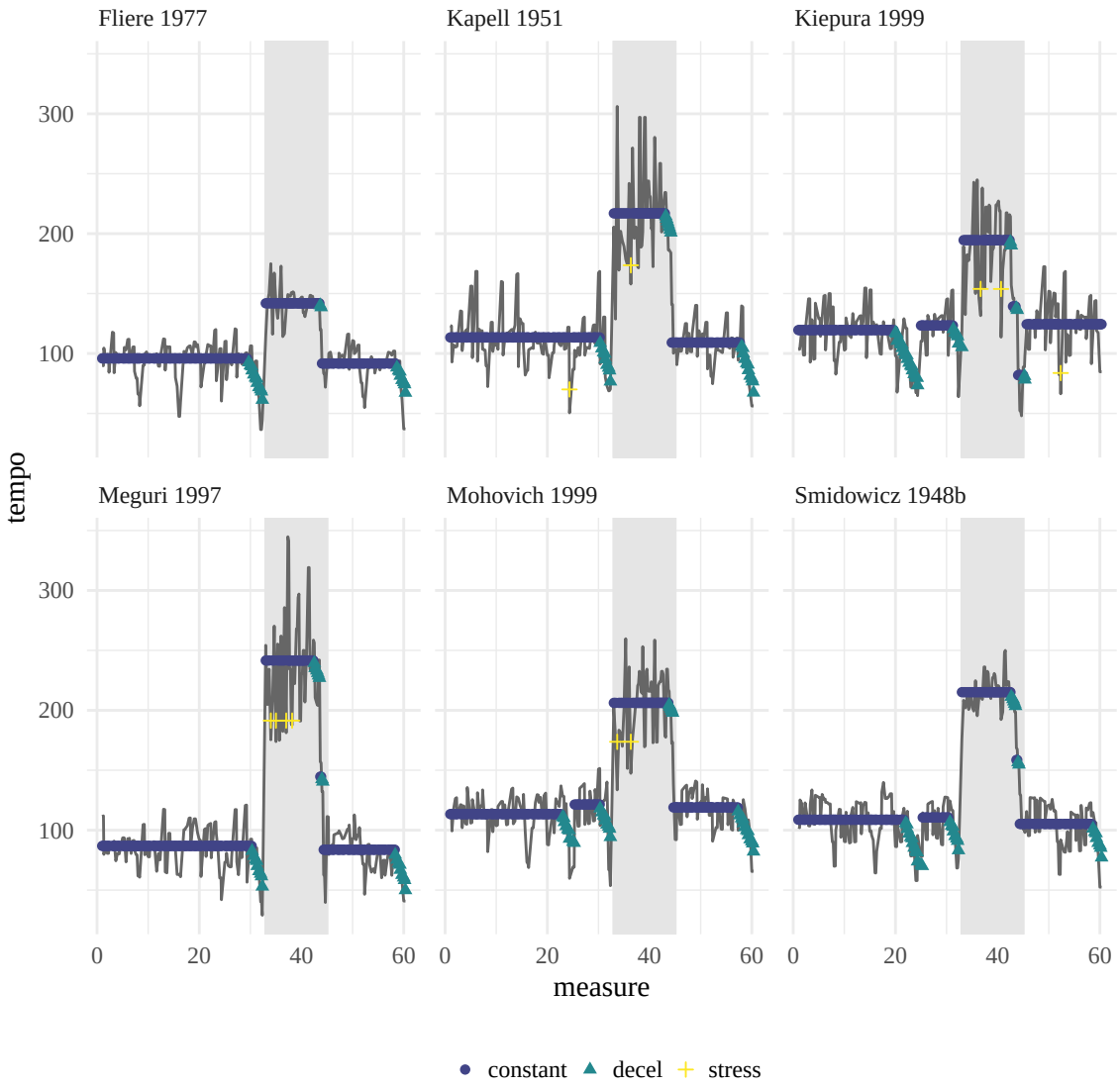

Figure SM-4: Performances in the third cluster.

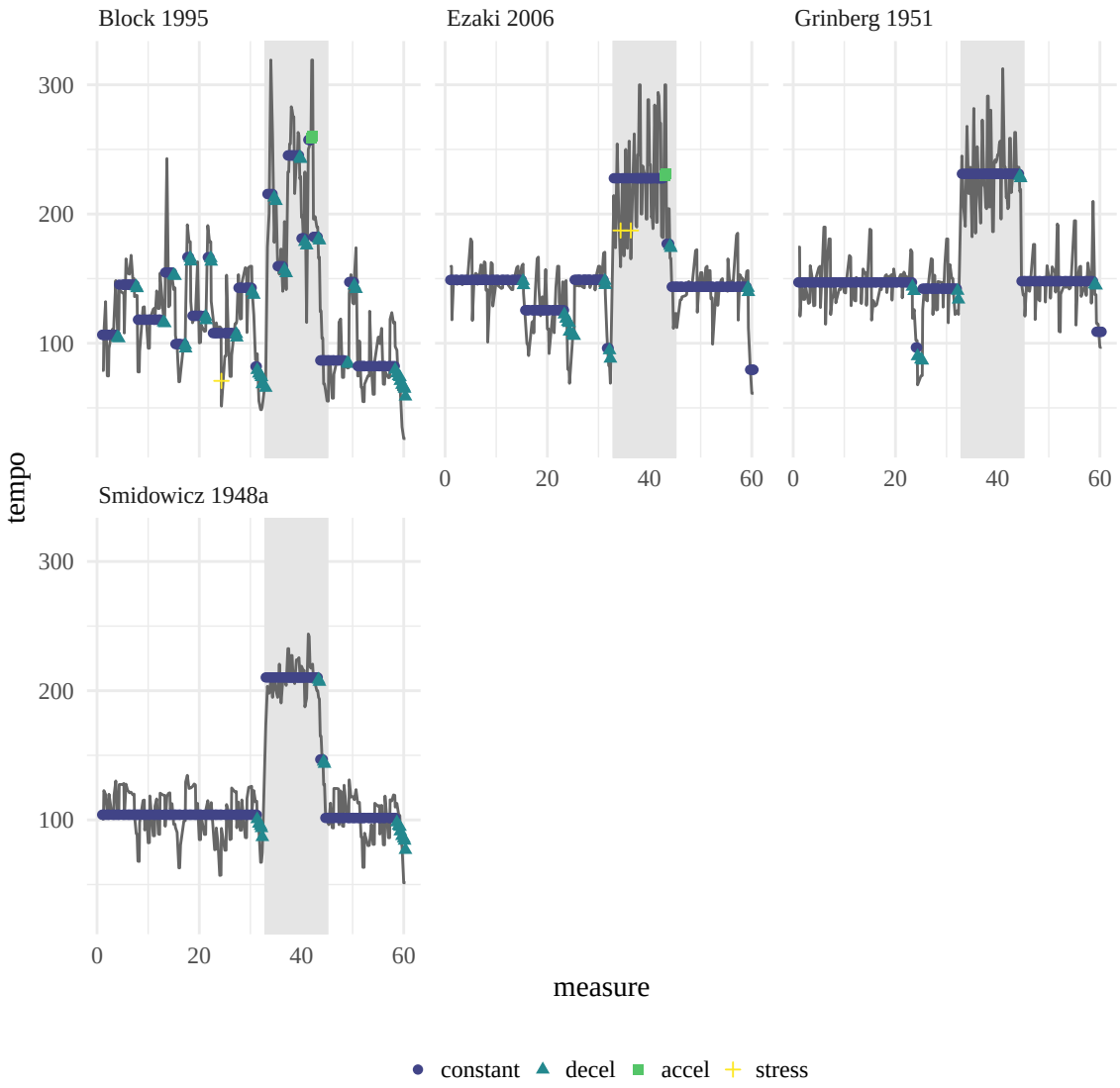

Figure SM-5: Performances in the fourth cluster.

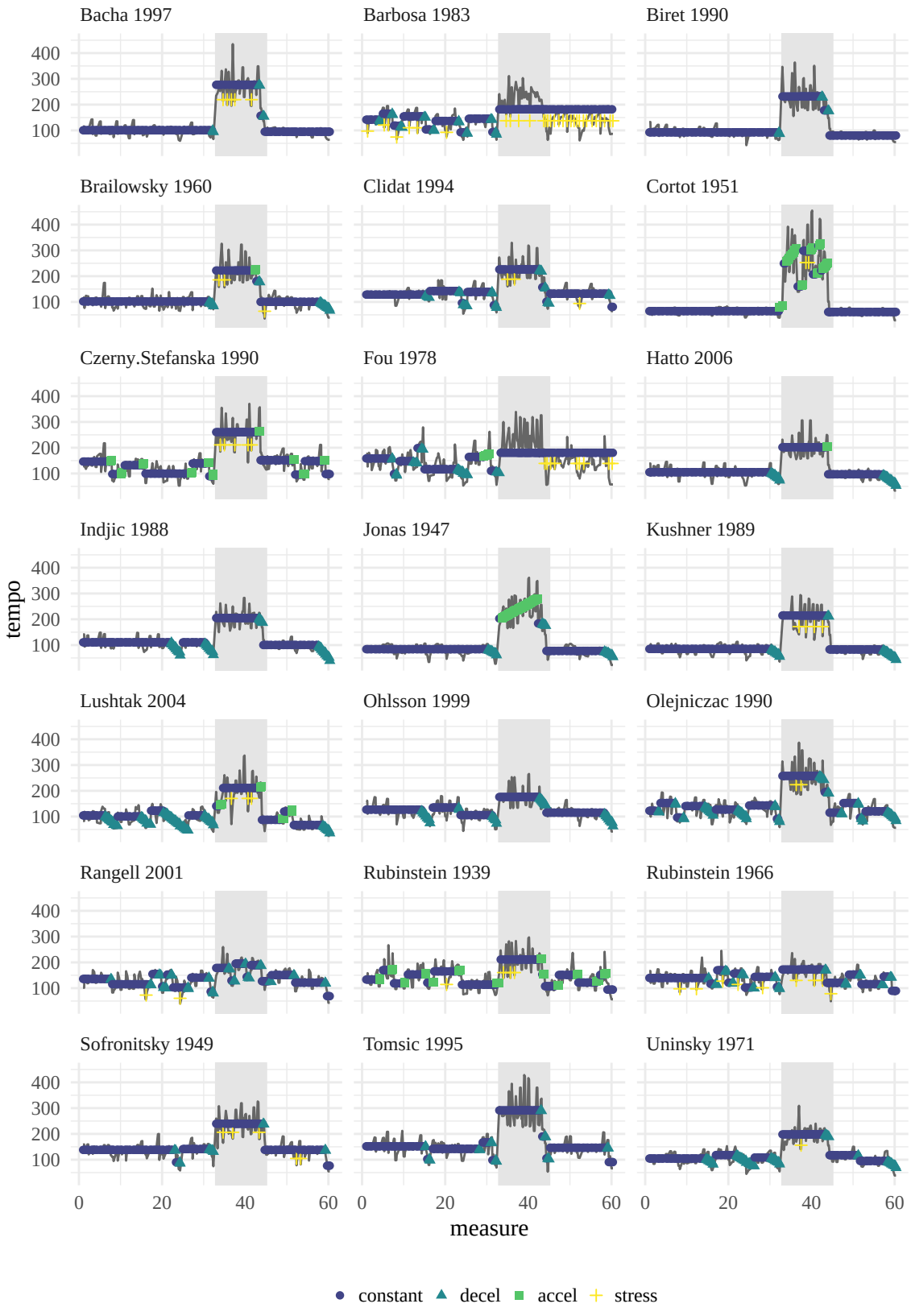

Figure SM-6: Performances in the "other" cluster.